IWQW

Institut für Wirtschaftspolitik und Quantitative
Wirtschaftsforschung

Diskussionspapier
Discussion Papers

No. 01/2014

Learning cost sensitive binary classification rules accounting for uncertain and unequal misclassification costs

Lydia Rybizki
University of Erlangen-Nuremberg

ISSN 1867-6707

Learning cost sensitive binary classification rules accounting for uncertain and unequal misclassification costs

Lydia Rybizki
University of Erlangen-Nuremberg
lydia.rybizki@fau.de

Abstract: This paper proposes cost sensitive criteria for constructing classification rules by supervised learning methods. Reinterpreting established loss functions and considering those introduced by Buja, Stuetzle, et al. (2005) and Hand (2009), we identify criteria reflecting different degrees of information about misclassification costs. To adapt classification methodology to practical cost considerations, we suggest the use of these criteria for different model selection approaches in supervised learning. In addition, we investigate the effects of cost sensitive adaptations in CART and boosting and conclude that adaptations are more promising in the selection rather than in the estimation step.

Keywords: unequal misclassification costs, proper scoring rules, AUC, boosting, CART, model selection, pruning, early stopping

1 Introduction

Classification tasks are an often encountered problem in various disciplines. In general, the aim is to combine observable predictive variables into a classification rule that best predicts unknown class membership of an object. The classification rule is learned from a training sample, i.e. objects for which both predictive variables and class membership have been observed. This amounts to methods of supervised learning, which commonly include an estimation and a model selection step.

It if often demanded to match methodology to the actual application. Hence, criteria defining the "best" prediction and thereby applicable for constructing classification rules should be chosen according to the practical problem at hand, cf. Hand and Vinciotti (2003), Berger (1985), Berk (2008).

Classification is commonly concerned with misclassification rate or misclassification cost. Some problems arise with employing these criteria for constructing classification rules. First, they would lead to discontinuous objective functions in optimization problems making many standard methods inapplicable. Second, it is often difficult in practice to exactly quantify misclassification costs for each of the classes. Instead, information about the consequences of false classifications may rather be incomplete or even nonexistent. Hence, the eligibility of a criterion is determined by the available information about misclassification costs.

Thanks to a large number of existing criteria for classifier evaluation and construction, solutions to these problems do exist. However, they have not been considered from a cost related point of view. The problem of a discontinuous objective function is overcome by focusing on class probability predictions and employing approximations reflecting equal misclassification costs. Unequal misclassification costs are commonly accounted for by adjusting prior probabilities according to assumed costs, cf. Landesa-Vázquez and Alba-Castro (2012), Therneau and Atkinson (2013)). An alternative approach is due to Masnadi-Shirazi and Vasconcelos (2007).

Buja, Stuetzle, et al. (2005) and Hand (2009) independently develop a class of criteria that are able to account for misclassification costs not exactly quantified by posing distributional assumptions on the cost ratio. Hand (2009) suggests their use in classifier evaluation. Buja, Stuetzle, et al. (2005) investigate their applicability for class probability estimation, for splitting in CART and as optimization criterion in boosting and compare them to standard approaches. However, they do not consider alternative cost sensitive approaches in their comparison. More important, the criteria's use in the selection step has not been investigated although this forms a crucial part in many learning methods.

Although it has not yet been made explicit, a criterion applicable in the case of completely unknown misclassification costs is given by the Area under the Receiver Operating Characteristics curve (AUC). Pepe and Thompson (2000) introduce the AUC as an optimization criterion for estimating parametric models, which is advanced by Ma and Huang (2005). The AUC is also proposed as a model selection criterion, for example by Ma and Huang (2007) and Huang, Qin, et al. (2011).

The aim of this paper is twofold. First, it is to reveal the ability of well known criteria to capture different degrees of information about misclassification costs. Second, we investigate different approaches of accounting for no, incomplete or complete information in classification methods to match statistical methodology to practical considerations.

The next section reviews loss functions for binary classification problems focusing on proper scoring rules, which reflect unknown, uncertain or fixed misclassification costs. The third section discusses cost sensitive adaptations for classification methods appertaining to supervised learning. Due to the principles underlying supervised learning, the proper scoring rules can be used to account for uncertain and unequal misclassification costs in estimation and selection. In particular, we construct cost sensitive prediction rules with Classification and Regression Trees (CART) and gradient boosting by reviewing, reinterpreting and advancing the ideas of Buja, Stuetzle, et al. (2005). Different cost sensitive adaptations are analyzed in a simulation study in Section 4. The effects of beta-loss are compared to those of prior probability adjustment in CART and of loss functions introduced by Masnadi-Shirazi and Vasconcelos (2007) in boosting. The final section discusses main results and avenues for further research.

2 Loss functions for cost sensitive classification

Predicting class membership can be regarded as a statistical decision problem of finding an optimal (*Bayes*) prediction rule. A prediction rule is a function f mapping predictive variables X to an output variable G, which both have joint density $f_{G,X}$. It serves to predict G when only X can be observed for an object. With a wrong prediction resulting in loss $L(G, f(X))$, the Bayes rule is to minimize expected loss given by

$$E_{G,X}[L(G, f(X))] = \int_{\mathbb{R}} \int_{\mathcal{X}} L(g, f(x)) f_{G,X}(g, x) \, \mathrm{d}x \, \mathrm{d}g.$$

This is equivalent to minimizing posterior expected loss

$$E_{G|X}[L(G, f(X))|X = x] = \int_{\mathbb{R}} L(g, f(x)) \pi_g(x) \, \mathrm{d}g,$$

where $\pi_g(x)$ denotes the conditional probability density or mass function of G given X. In regression problems, the prediction rule is a function $f : \mathcal{X} \mapsto \mathbb{R}$ and the loss of a prediction usually serves to reflect a distance between g and $f(x)$; compare e.g. Berger (1985). In a binary classification problem, G is a categorical variable with support $\{0, 1\}$ denoting the membership to a class Ω_g. The prediction rule or *classification rule* is a function $d : \mathcal{X} \mapsto \{0, 1\}$. Making use of simplifications for the two-class-case, we use loss functions $L(0, d(x)) = L_0(d(x))$, $L(1, d(x)) = L_1(d(x))$; prior class probabilities $\pi_1 = \pi$, $\pi_0 = 1 - \pi$ and posterior class probabilities $\pi_1(x) = \pi(x)$, $\pi_0(x) = 1 - \pi(x)$. Hence, expected loss simplifies to

$$E_{G,X}[L(G, d(X))] = \int_{\mathcal{X}} L_0(d(x)) f_0(x)(1 - \pi) \, \mathrm{d}x + \int_{\mathcal{X}} L_1(d(x)) f_1(x) \pi \, \mathrm{d}x \tag{1}$$

with posterior expected loss

$$E_{G|X}[L(G, d(X))|X = x] = L_0(d(x))(1 - \pi(x)) + L_1(d(x))\pi(x). \tag{2}$$

In the standard approach to classification, the loss function of a classification rule is determined by the consequences of assigning an object from Ω_i to Ω_j, which are usually termed *costs* $c_{i,j}$. Assuming perfect information about misclassification costs $c_{0,1}$ and $c_{1,0}$ leads to generalized 0-1-loss

$$\begin{aligned} L_0(d(x)) &= \mathbb{I}(d(x) = 1) c_{0,1} \\ L_1(d(x)) &= \mathbb{I}(d(x) = 0) c_{1,0}. \end{aligned} \tag{3}$$

The Bayes classification rule d_c^* for generalized 0-1-loss minimizing (posterior) expected loss is given by

$$d_c^*(x) = \mathbb{I}\left(\pi(x) > c\right), \quad \text{with } c = c_{0,1}/(c_{0,1} + c_{1,0}). \tag{4}$$

Thus, the decision only depends on $\pi(x)$ and the ratio of costs. With known mis-

classification costs, finding the optimal decision rule reduces to the task of predicting posterior class probability $\pi(x)$. Probability prediction can be considered as a special type of regression problem employing a prediction rule $p : \mathcal{X} \mapsto [0, 1]$.

In standard decision theory, probability predictions are assessed by *scoring rules*. In general, scoring rules are loss functions L^S defining the loss from stating a probability distribution and a certain event occurs. In the two-class case, a probabilistic forecast is completely given by stating a probability p for one of the classes. A scoring rule is *proper* if the expected loss is defined, minimized and less than $+\infty$ when p is the true class probability; cf. Gneiting and Raftery (2007, Section 2) and Schervish (1989). The minimum expected loss of a proper scoring rule is known as *information measure*. Guaranteeing minimum expected loss for the true probability, proper scoring rules seem a natural choice of loss functions if classification is based on class probabilities.

Buja, Stuetzle, et al. (2005) have transferred scoring rules to statistical decision theory, where predictions regard posterior class probabilities given X. In the binary case with $p(x)$ denoting predicted posterior class 1 probability, scoring rules are defined by a pair of functions $(L_0^S(p(x)), L_1^S(p(x)))$. Posterior expected loss of a proper scoring rule, given by

$$E_{G|X}[L^S(G, p(X))|X = x] = L_0^S(p(x))(1 - \pi(x)) + L_1^S(p(x))\pi(x)$$

is minimized by the Bayes rule $p^*(x) = \pi(x)$. The minimum achievable posterior expected loss yields the information measure for the data dependent case, which is defined as

$$H_x(\pi) = E_{G|X}[L^S(G, \pi(X))|X = x] = L_0^S(\pi(x))(1 - \pi(x)) + L_1^S(\pi(x))\pi(x), \tag{5}$$

for all proper scoring rules L^S; cf. Buja, Stuetzle, et al. (2005).

Using the integral representation for binary proper scoring rules, which is derived by Schervish (1989), Buja, Stuetzle, et al. (2005) reveal a practical interpretation of proper scoring rules: The integral representation of proper scoring rules is

$$\begin{aligned} L_0^S(p(x)) &= \int_0^1 \mathbb{I}(p(x) > c) c \, w(c) \, \mathrm{d}c \\ L_1^S(p(x)) &= \int_0^1 \mathbb{I}(p(x) \leq c)(1 - c) \, w(c) \, \mathrm{d}c. \end{aligned} \tag{6}$$

Restating generalized 0-1-loss as function of $p(x)$ leads to

$$\begin{aligned} L_0(p(x)) &= \mathbb{I}(p(x) > c)c_{0,1} = \mathbb{I}(p(x) > c)(c \cdot (c_{0,1} + c_{1,0})) \\ L_1(p(x)) &= \mathbb{I}(p(x) \leq c)c_{1,0} = \mathbb{I}(p(x) \leq c)((1 - c) \cdot (c_{0,1} + c_{1,0})). \end{aligned}$$

As decision problems are equivalent for a loss L and a loss aL, where a is an arbitrary constant (Dawid (2007)), $(c_{0,1} + c_{1,0})$ can be omitted without changing the optimal prediction rule. Thus, proper scoring rules can be interpreted as misclassification costs averaged for a proportion c, where the Lebesgue density $w(c)$ of $w(dc)$ serves as weight

function; cf. Theorem 1' of Buja, Stuetzle, et al. (2005).

Hand (2010) derives a similar expression employing decision rules based on a score $s : \mathcal{X} \mapsto \mathbb{R}$, which is interpreted as a monotone increasing transformation of predicted class probability $p(x)$. Expected loss from generalized 0-1-loss can then be written as

$$E_{G,X}[L(G, s(X))] = (c(1-\pi)(1-F_{s_0}(t)) + (1-c)\pi F_{s_1}(t))(c_{0,1} + c_{1,0}), \tag{7}$$

with $F_{s_g}(t) = P(s(X) \leq t | G = g)$. Assuming that F_{s_g} are differentiable, the optimal threshold t^* for a given ratio c can be found by maximizing (7). Hand (2010) shows that, given the solution is unique, the optimal threshold is $t^*(c) = P_1^{-1}(c)$, where P_1^{-1} is the inversion of

$$P_1(t) = P(G = 1|t) = \frac{\pi f_{s_1}(t)}{(1-\pi) f_{s_0}(t) + \pi f_{s_1}(t)}.$$

If $s^*(x)$ is a monotone transformation of the true posterior class probability $\pi(x)$, then $\mathbb{I}(s^*(x) > t^*(c))$ is a Bayes classification rule for generalized 0-1-loss.

Assuming that $c = c_{0,1}/(c_{0,1} + c_{1,0})$ and $v = c_{0,1} + c_{1,0}$ have a joint density $w_{CV}(c, v)$, Hand (2010) derives the expected loss of the classification rule $\mathbb{I}(s(x) > t^*(c))$ as:

$$E_{G,X}[L(G, s(X))] = \int_0^1 \Big(c(1-\pi)(1-F_{s_0}(t^*(c))) + (1-c)\pi F_{s_1}(t^*(c))\Big)\, \nu(c)\, \mathrm{d}c,$$
$$\text{with} \quad \nu(c) = \int_0^\infty v\, w_{CV}(c, v)\, \mathrm{d}v.$$

With c and v independent the weight simplifies to $\nu(c) = w(c)E(v)$; cf. Hand (2010). Setting $E(v) = 1$, it can be shown that expected loss of Hand (2010) can be rewritten as expected loss in (1) with $d(x) = \mathbb{I}(s(x) > t^*(c))$ and loss functions

$$L_0^S(s(x)) = \int_0^1 \mathbb{I}(s(x) > t^*(c)) c\, w(c)\, \mathrm{d}c$$
$$L_1^S(s(x)) = \int_0^1 \mathbb{I}(s(x) \leq t^*(c))(1-c)\, w(c)\, \mathrm{d}c,$$

so that loss functions are equivalent to proper scoring rules of Buja, Stuetzle, et al. (2005).

The weight $w(c)$ enables the construction of arbitrary proper scoring rules, on the one hand, and is interpretable as cost weight, on the other. These features can be used to derive loss functions for $p(x)$ that are tailored to the actual problem of application. That is, weights can be set to reflect assumptions about likely values of c and thereby about the ratio of the two types of misclassification costs. Buja, Stuetzle, et al. (2005) and Hand (2010) propose to use weight functions stemming from the probability density of the Beta distribution

$$w_{a,b}(c) = \frac{1}{B(a,b)} c^{a-1}(1-c)^{b-1}.$$

This family includes squared error loss with weight $2w_{a,b}(c)$ and $a = b = 1$ so that

$w(c) = 1$

$$\begin{aligned} L_0^{bs}(p(x)) &= p(x)^2 \\ L_1^{bs}(p(x)) &= (1-p(x))^2. \end{aligned} \tag{8}$$

Generalized 0-1-loss results from $a = b \to \infty$ so that $w(dc) = \delta_c(\,\mathrm{d}c)$

$$\begin{aligned} L_0^{c}(p(x)) &= \mathbb{I}(p(x) > c)c \\ L_1^{c}(p(x)) &= \mathbb{I}(p(x) \leq c)(1-c). \end{aligned} \tag{9}$$

If the weight function is not required to be a probability density, normalization by $B(a,b)$ can be omitted leading to weight functions

$$w'_{a,b}(c) = c^{a-1}(1-c)^{b-1},$$

which lead to log-loss for $a = b = 0$ with $w(c) = (c(1-c))^{-1}$

$$\begin{aligned} L_0^{log}(p(x)) &= -\log(1-p(x)) \\ L_1^{log}(p(x)) &= -\log(p(x)). \end{aligned} \tag{10}$$

and to exponential loss for $a = b = -1/2$ with $w(c) = 1/(c(1-c))^{3/2}$

$$\begin{aligned} L_0^{exp}(p(x)) &= \left(\frac{p(x)}{1-p(x)}\right)^{1/2} \\ L_1^{exp}(p(x)) &= \left(\frac{1-p(x)}{p(x)}\right)^{1/2}, \end{aligned} \tag{11}$$

cf. Buja, Stuetzle, et al. (2005). In general, the weighting with Beta densities yields *beta-loss*, i.e. proper scoring rules of the form

$$\begin{aligned} L_0^{\mathcal{B}(a,b)}(p(x)) &= \frac{B(a+1,b)}{B(a,b)} F_{\mathcal{B}(a,b)}(p(x)) \\ L_1^{\mathcal{B}(a,b)}(p(x)) &= \frac{B(a,b+1)}{B(a,b)} (1 - F_{\mathcal{B}(a,b)}(p(x))), \end{aligned} \tag{12}$$

where $F_{\mathcal{B}(a,b)}$ denotes the cumulative distribution function of a Beta distribution.

The Beta density is a suitable weight as it can be interpreted to reflect reasonable distributional assumptions about c, cf. Buja, Stuetzle, et al. (2005). Assuming that $c_{0,1}$ and $c_{1,0}$ add up to 1, we have $a/b = c/(1-c) = c_{0,1}/c_{1,0}$ so that the ratio of a and b can be chosen according to assumptions about the misclassification cost ratio. In consequence, setting $a = b$ leads to a proper scoring rule approximating 0-1-loss. Values of a, b reflect the amount of certainty about the ratio: the larger a, b the smaller the variance of c which implies stronger certainty about c.

Thus, beta-loss enables the direct evaluation of a probability prediction $p(x)$ in terms of misclassification costs. The aim of classification can be accounted for when evaluating $p(x)$ without requiring the derivation of a classification rule. In contrast to generalized

0-1-loss, beta-loss avoids specification of a fixed cut-off but still captures assumptions about misclassification costs.

The so called ranking loss originally refers to the decision problem of bringing two items into a ranking order. Group membership is understood to pose a natural order of two objects with features X_1 and X_2, i.e. $G_1 > G_2$ means that object X_1 ranks higher that X_2. Given two independently drawn objects (X_1, G_1) and (X_2, G_2) with distribution $F_{G,X}$, the decision problem is to find a ranking rule that ranks the objects according to observable X_1 and X_2. The expected loss of a ranking rule is commonly defined as the probability of ranking two randomly drawn objects incorrectly. For the *bipartite* ranking problem, where G takes values in $\{0, 1\}$, Clémençon, Lugosi, et al. (2008) show that relevant ranking rules are of the form $r(x_1, x_2) = \mathbb{I}(s(x_1) \geq s(x_2))$, which takes value 1 if the first object is to be ranked higher than the second one. The score $s(x)$ is again a strictly increasing transformation of $p(x)$. Finding the Bayes rule r^* in the bipartite ranking problem is equivalent to finding a score $s^*(x)$, or equivalently stating a $p(x)$ that is equal to $\pi(x)$. As shown by Clémençon, Lugosi, et al. (2008), the expected loss $E[L^r(G, s(X))]$ of a ranking rule based on $s(x)$ can be expressed as function of the *Area under the Receiver Operating Characteristic curve* AUC_s of score $s(x)$, which is a widely accepted criterion in classifier evaluation. Precisely,

$$E[L^r(G, s(X))] = 2(1 - AUC_s)\pi_0\pi_1.$$

Thus, the optimal score $s^*(x)$ that minimizes $E[L^r(G, s(X))]$ maximizes the AUC of the score and vice versa. Hence, posterior class probability $\pi(x)$ or any monotone transformation maximizes the AUC, which is the Neyman & Pearson result as stated by Green and Swets (1966) and McIntosh and Pepe (2002).

Using results of Hand (2010), the expected ranking loss can also by expressed by cost-weighted loss. More specifically, we state the following proposition:

Proposition 2.1 (Ranking loss as proper scoring rule)**.** *Let $s(x)$ denote a strictly monotone transformation of a probability forecast $p(x)$ and r a ranking rule of the form $r(x_1, x_2) = \mathbb{I}(s(x_1) > s(x_2))$. Let f_{p_0} and f_{p_1} denote the densities of $p(X)$ in Ω_0 and Ω_1, respectively.*

If the transformation is the identity such that $r(x_1, x_2) = \mathbb{I}(p(x_1) > p(x_2))$, the loss functions employed in expected loss of a ranking problem are proper scoring rules and are equal to cost weighted misclassification losses as defined in (6) *with weight*

$$w(c) = (1 - \pi) f_{p_0}(P_1^{-1}(c)) \left| \frac{\mathrm{d}P_1^{-1}(c)}{\mathrm{d}c} \right| + \pi f_{p_1}(P_1^{-1}(c)) \left| \frac{\mathrm{d}P_1^{-1}(c)}{\mathrm{d}c} \right|. \tag{13}$$

such that

$$L_0^r(p(x)) = \int_0^1 \mathbb{I}(p(x) > c) c \; w_r(c) \, \mathrm{d}c$$
$$L_1^r(p(x)) = \int_0^1 \mathbb{I}(p(x) \leq c)(1 - c) \; w_r(c) \, \mathrm{d}c.$$

Proof. See Appendix A. □

The loss functions defining ranking loss can be interpreted as cost weighted misclassification losses, where the weight is given by (13). In this case, the cost weight depends on the class conditional densities of $p(x)$ (or the score $s(x)$). The expected loss provides an evaluation of a prediction rule $s(x)$ or $p(x)$ without requiring information about misclassification costs. For implementation, the AUC should be preferred to $E[L^r(G, s(X))]$ as is its empirical version is computed more easily.

Preceding results show that decision problems like probability forecasting or ranking are also useful for binary classification problems. From a classification perspective, the loss functions characterizing these decision problems reflect a different state of knowledge about misclassification costs. While the standard loss function for classification problems, the generalized 0-1-loss, quantifies the consequences as fixed costs $c_{0,1}$ and $c_{1,0}$, other loss functions assume less certainty. Employing proper scoring rules with beta weights, probability predictions can be evaluated from a classification perspective without requiring specific values of misclassification costs. As many real life problems suffer from uncertain misclassification consequences, beta-loss seems an appropriate choice for classification problems. Ranking loss can be considered a cost-free criterion as assumptions about misclassification costs can be avoided completely. It might be useful for early stages of classifier development to ignore misclassification consequences and to assess performance independently from any cost assumptions.

3 Learning classification rules from data

The joint distribution $F_{G,X}$ needed to obtain expected loss is unknown in practice. Supervised statistical learning is concerned with constructing optimal prediction rules from data by applying the principle of empirical risk minimization (ERM). Precisely, given a training sample $\mathcal{T} = \{(x_i, g_i)|i = 1...n\}$ of n objects, where $x_1, ..., x_n$ are realizations of n independent and identically distributed random vectors $X_1, ..., X_n$ with distribution function F_X and where g_i denotes the observed output variable of these objects, the aim is to construct an optimal decision rule. The obtained rule serves to predict G for an object outside the training sample with unknown output assuming that it is drawn from the same distribution $F_{G,X}$ as $\mathcal{T}$. Transferring the principles of statistical decision theory, the aim in ERM is to approximate the true Bayes prediction rule f^*. One considers functions $f(x, l)$, $l \in \Lambda$, where Λ is some set of abstract parameters indexing the set of functions. Among this set, one seeks the function that minimizes the empirical version of expected loss, the so called *empirical risk*

$$r(f(x, l)) = \frac{1}{n} \sum_{i=1}^{n} L(g_i, f(x_i, l)), \tag{14}$$

where $x_i = (x_{i1}, ..., x_{iP})'$ denotes the realization of $X = (X_1, ..., X_P)'$ for the ith object. Thereby, the estimated Bayes prediction rule $\hat{f}(x) = f(x, \hat{l})$ is obtained with

$$\hat{l} = \underset{l \in \Lambda}{\operatorname{argmin}}\, r(f(x, l)).$$

As a task of supervised learning, classification is characterized by the empirical risk based on 0-1-loss:

$$r(d(x, l)) = \frac{1}{n} \sum_{i=1}^{n} \mathbb{I}(g_i \neq d(x_i, l)), \tag{15}$$

where $d(x, l)$ denotes the approximating function of a classification rule. However, many classification methods are based on regression, dealing with continuous outcome variables G such that approximating functions are $f(x, l) : \mathcal{X} \mapsto \mathbb{R}$. A monotone transformation $h(f(x, l))$ is commonly imposed to yield a regression function $p(x, l)$ taking values between 0 and 1, which is motivated in two different ways; compare e.g. Cherkassky and Mulier (1998, pp. 308-311). Firstly, the transformed function can be understood as approximating function for $\pi(x)$ and the transformation ensures that the estimate can be interpreted as probability. Proper scoring rules like squared error loss can then be used as suitable loss functions. An alternative interpretation is the aim of avoiding discontinuous objective functions. Therefore, h is understood as an approximation of the indicator function $d(x, l)$ by approaching 0 for smaller and 1 for larger arguments. In addition, the discontinuous loss function is replaced by a continuous one. As noted above, squared error loss resulting from $a = b = 1$ is an approximation to 0-1-loss. In consequence, beta-loss with $a \neq b$ could be used as a continuous approximation of generalized 0-1-loss. The theoretical feature of beta-loss of reflecting uncertainty about misclassification costs therefore translates to a more practical advantage for constructing classification rules from data. It enables direct implementation of unequal misclassification costs with a continuous form of empirical risk.

Due to the special role of $\pi(x)$ shown in the previous section, the remainder focuses on regression based classification. Methods for regression can be differentiated according to the set of functions $f(x, l)$, $l \in \Lambda$ considered for the learning task. Choosing a set of functions faces a trade-off between the necessity of considering functions flexible enough to well approximate the true f^* and of restricting flexibility to avoid overfitting, cf. Hastie, Tibshirani, et al. (2009, Chapter 2). For a large number of methods, $f(x, l)$ can be expressed by a linear combination of *basis expansions* seeking a compromise between restricted parametric and flexible nonparametric approaches. Approximating functions take the form

$$f(x, \beta, \gamma) = \sum_{m=1}^{M} g_m(x, \gamma_m)\beta_m, \tag{16}$$

where both β and γ are estimated from data. Whenever a wide class of approximating

functions is considered the ERM principle cannot be applied directly but needs to be extended to control model complexity and to enable model selection. Constructing a prediction rule from data then subsumes model selection and estimation. Restrictions can also be imposed apriori by omitting γ and employing fixed basis functions $g_m(x)$. This leads to non-adaptive methods that reduce to parametric estimation and do not require model selection, cf. Cherkassky and Mulier (1998, Section 3).

3.1 Estimation with proper scoring rules

The use of alternative loss functions for estimation has already been suggested elsewhere. With a fixed set of basis functions the parameter vector $\beta = (\beta_1, ..., \beta_M)'$ with $\beta \in \mathbb{R}^M$ can be estimated by directly applying the ERM principle. Transforming $f(x, \beta)$ to $p(x, \beta)$ enables the use of proper scoring rules. In general, with $g_i \in \{0, 1\}$, the empirical risk of $p(x, \beta)$ is

$$r(p(x, \beta)) = \frac{1}{n} \sum_{i=1}^{n} g_i L_1(p(x_i, \beta)) + (1 - g_i) L_0(p(x_i, \beta)). \tag{17}$$

Employing log-loss leads to Maximum Likelihood estimation. If $p(x, \beta) = h(\beta' x)$ and h is the logistic function, standard logistic regression results. Squared error loss leads to Brier score (8). As seen in the previous section, the class of proper scoring rules offer alternatives that are more suitable for cost sensitive classification.

Buja, Stuetzle, et al. (2005) suggest the use of beta-loss for estimation. The parameters a and b can be chosen according to the assumptions about misclassification costs so that practical considerations are taken into account by the estimation criterion. Implementations show that estimation is unstable for very large values of a and b, which prevents loss functions that are close to generalized 0-1-loss.

Pepe and Thompson (2000) indirectly proposed to use ranking loss by using the empirical AUC as objective function for estimating β in $f(x, \beta)$. Precisely, $\hat{\beta}$ results from maximizing the empirical AUC. Ma and Huang (2005) propose the *sigmoid* AUC to approximate the indicator function in the empirical AUC thereby avoiding optimization with a discontinuous objective function.

3.2 Model selection with proper scoring rules

Many methods employ a large number or flexible basis expansions necessitating model selection to avoid overfitting. Model complexity depends on the *features* g_m. Restrictions are therefore implemented through parametric penalties, which control the complexity indirectly by imposing constraints on the parameters β or by early stopping; compare e.g. Cherkassky and Mulier (1998, pp. 70-74, 265-267).

Formally, a penalized form of empirical risk is employed as optimization criterion:

$$r_{pen}(f(x, \beta, \gamma)) = r(f(x, \beta, \gamma)) + \lambda J(\beta). \tag{18}$$

The regularization parameter λ determines the strength of the penalty imposed. The

right choice of λ constitutes the model selection problem, i.e. finding a $\hat{\lambda}$ in (18) such that $\hat{f}_{\hat{\lambda}}(x)$ obtained from minimizing the penalized risk for $\hat{\lambda}$ is the best estimate of the true Bayes rule f^*.

Determining the "best" $\hat{f}_{\hat{\lambda}}(x)$ indicates assessment of the estimated prediction rule, which is the reason for the close relationship of model evaluation and selection. Accounting for the purpose of estimated rules to make predictions for an object (X_0, G_0) outside the training sample, prediction error $R_{\mathcal{T}}(\hat{f}(X_0)) = E(L(G_0, \hat{f}(X_0))|\mathcal{T})$ seems a reasonable criterion. It is especially important in *wrapper* methods, which amount to feature selection after applying the estimation step, and in *embedded* methods, which integrate estimation and selection. In contrast, *filter* methods exclude features before applying a particular learning method. See GUYON and ELISSEEFF (2003) for a detailed discussion.

As the true prediction error is unknown, estimation of prediction error $R_{\mathcal{T}}(\hat{f})$ is crucial to the problems of model assessment and selection. Cross-validation provides a very flexible approach to prediction error estimation and thereby to model selection. In general, cross-validation proceeds as follows.

1. Divide $\mathcal{T} = \{(x_i, g_i)|i = 1...n\}$ into K subsets $\mathcal{T}_1, ..., \mathcal{T}_K$.

2. Estimate the prediction rule $f^*_{\mathcal{T}_{-k},\lambda}(x)$ from data in $\mathcal{T}_{-k} = \mathcal{T}\backslash\mathcal{T}_k$ by minimizing (penalized) empirical risk for a certain λ, where λ implies some kind of complexity of f.

3. Use $\hat{f}_{\mathcal{T}_{-k},\lambda}(x)$ to make prediction for objects i in $\mathcal{T}_k$ and calculate the prediction error
$$r_k(\hat{f}_{\mathcal{T}_{-k},\lambda}) = \frac{1}{n_k} \sum_{\{i|(x_i,g_i)\in\mathcal{T}_k\}} L(g_i, \hat{f}_{\mathcal{T}_{-k},\lambda}(x_i)),$$
where n_k denotes the number of objects in $\mathcal{T}_k$.

4. Repeat 2 and 3 for $k = 1, ..., K$ and calculate
$$r_{cv}(\hat{f}_\lambda) = \frac{1}{K} \sum_{k=1}^{K} r_k(\hat{f}_{\mathcal{T}_{-k},\lambda}).$$
which is equal to
$$r_{cv}(\hat{f}_\lambda) = \frac{1}{n} \sum_{i=1}^{n} L(g_i, \hat{f}_{\mathcal{T}_{-k(i)},\lambda}(x_i)),$$
where $\mathcal{T}_{-k(i)}$ indicates the subset $\mathcal{T}_{-k}$ that does not contain object (g_i, x_i), $\mathcal{T}_{-k(i)} = \{\mathcal{T}\backslash\mathcal{T}_k|(g_i, x_i) \in \mathcal{T}_k\}$.

5. Repeat 2 to 4 for different λ.

6. Choose $\hat{\lambda}$ for which $r_{cv}(\hat{f}_\lambda)$ is minimal.

Estimates of prediction error obtained by resampling are not restricted to particular functional forms of L and f. In addition, the loss function L in r_{cv} does not have

to coincide with the loss function used in estimation. When predicting a quantitative output, the aim is to obtain a prediction that is close to the true value so that squared error or log-loss are a common choice for L. Thus, in regression problems these loss functions are used both in estimation and selection steps. If $\hat{p}(x)$ is used to derive a classification rule, correct predictions of class membership are more important than accurately predicting class probability. Thus, classification rules are typically assessed and selected by criteria based on 0-1-loss; compare e.g. CHERKASSKY and MULIER (1998, Chapter 8) or MCLACHLAN (1992, Chapter 10). Hence, the loss functions differ from the one in estimation, where 0-1-loss is usually approximated as described above; cf. CHERKASSKY and MULIER (1998, Chapter 8).

FRIEDMAN (1997) notes that a poor approximation of $\hat{p}(x)$ to $\pi(x)$ according to squared error or log-loss does not necessarily lead to low classification performance of the corresponding classification rule $\hat{d}(x) = \mathbb{I}(\hat{p}(x) > c)$. Instead, it is widely acknowledged that methods providing accurate estimates of class probabilities might perform worse in terms of 0-1-loss than simpler methods with poor estimation performance; see for example FRIEDMAN (1997), HAND and VINCIOTTI (2003). The trade-off that constricts model fit and motivates model selection in regression does therefore not apply to classification directly. When applying regression methods to construct classification rules, this difference should be accounted for in model selection. Therefore we introduce adaptations to suit the classification problem within in the model selection step by taking into account cost considerations.

Employing generalized 0-1-loss and a classification rule derived from an estimate $\hat{p}(x) = p(x, \hat{\beta}, \hat{\gamma})$ of $\pi(x)$, the selection criterion is thus defined as

$$r_{cv}^{c}(\hat{p}_\lambda) = \frac{1}{n}\sum_{i=1}^{n}(1-c)g_i\mathbb{I}(\hat{p}_{\mathcal{T}_{-k(i)},\lambda}(x_i) \leq c) + c(1-g_i)\mathbb{I}(\hat{p}_{\mathcal{T}_{-k(i)},\lambda}(x_i) > c),$$
$$\text{with} \quad c = c_{0,1}/(c_{0,1}+c_{1,0}).$$

But its application is complicated by the fact that $c_{0,1}$ and $c_{1,0}$ and hence c are difficult to quantify in practice. For the case of uncertain misclassification costs with available distributional assumptions regarding the ratio of $c_{0,1}$ and $c_{1,0}$, the beta-loss can be applied instead:

$$r_{cv}^{\mathcal{B}(a,b)}(\hat{p}_\lambda) = \frac{1}{n}\sum_{i=1}^{n} g_i \frac{B(a,b+1)}{B(a,b)}(1 - F_{\mathcal{B}(a,b)}(\hat{p}_{\mathcal{T}_{-k(i)},\lambda}(x_i))) + (1-g_i)\frac{B(a+1,b)}{B(a,b)}F_{\mathcal{B}(a,b)}(\hat{p}_{\mathcal{T}_{-k(i)},\lambda}(x_i)).$$

As noted above, beta-loss has the advantage of being directly applicable to probability predictions instead of classification rules. With suitably chosen weights, it still implies evaluation and selection based on misclassification costs. The choice of a fixed cut-off as necessary with generalized-loss can be avoided, which accounts for uncertainty about unequal misclassification cost. So far, beta-loss based criteria have not been used for model selection.

If no assumptions about misclassification costs can be made, ranking loss was identified as most suitable loss function. In that case, the prediction error is based on the ranking loss L^r. Employing its cost weighted form with weight $w_r(c)$ requires estimation of score densities f_{s_0} and f_{s_1}. Employment of the AUC is more practicable leading to a cross-validatory selection criterion which is to maximized. The pairwise comparison requires separate calculation of the prediction error for each fold k

$$r_k^{auc}(\hat{p}_{\mathcal{T}_{-k},\lambda}) = \frac{1}{n_k} \sum_{i=1}^{n_k} \sum_{j \neq i} \mathbb{I}(g_i > g_j) \cdot \mathbb{I}(\hat{p}_{\mathcal{T}_{-k},\lambda}(x_i) > \hat{p}_{\mathcal{T}_{-k},\lambda}(x_j))$$

to obtain the final estimate

$$r_{cv}^{auc}(\hat{p}_\lambda) = \frac{1}{K} \sum_{k=1}^{K} r_k^{auc}(\hat{p}_{\mathcal{T}_{-k},\lambda}).$$

These criteria can be employed in wrapper and embedded approaches to model selection, where λ takes a different role but the general procedure applies in each case.

3.2.1 Wrapper methods and embedded methods with penalties

Wrapper methods lend themselves to non-adaptive methods with a large number of basis expansions $g_m(x)$ so that model selection constitutes a pure feature selection problem. In that case, the parameter λ can be understood as indexing a certain subset of features. The procedure described above is performed for various feature subsets. Estimated prediction error serves to select the best feature subset so that the cost adapted criteria above, especially beta-loss, can be employed easily. Using the AUC, estimated by cross-validation, for assessing estimated scores $f(x, \hat{\beta})$ is proposed by Ma and Huang (2007). In particular, it is employed within a wrapper approach to feature selection by Huang, Qin, et al. (2011). Wang, Chen, et al. (2011) also employ a cross-validatory estimate of the AUC for feature selection, where the prediction rule is obtained by logistic regression.

Embedded methods with penalties estimate β by minimizing penalized empirical risk r_{pen} as stated in (18). This favors continuous forms of the empirical risk r and the penalty J. The penalty shrinks the parameter towards 0 to control complexity. The loss functions entering empirical risk can be chosen as in Section 3.1 so that cost sensitive adaptations can be implemented. In fact, Ma and Huang (2008) point out that any combination of empirical risk and penalty could be implemented. For example, Zhou, Chen, et al. (2012) use a smoothed version of the empirical AUC. A cost oriented perspective can also adopted by employing beta-loss possibly approximating generalized 0-1-loss.

The model selection step of choosing $\hat{\lambda}$ is done by the cross-validation procedure described above. Hence, cost sensitive adaptations are straightforward using suitable loss functions for prediction error.

3.2.2 Embedded methods with pruning: CART

Another form of embedded methods is due to performing greedy optimization. Basis expansions g_m are added and hidden parameter vectors γ_m are estimated in a stepwise mode, so that feature selection is part of the optimization process to minimize (penalized) empirical risk; cf. e.g. CHERKASSKY and MULIER (1998, Chapter 5). The model selection step consists of deleting added basis expansions denoted as *pruning*. Classification and Regression Trees, imputed to BREIMAN, FRIEDMAN, et al. (1984), is an embedded method that employs pruning. In basis expansions expression the regression function takes the form

$$\begin{aligned} f(x, \beta, T) &= \sum_{m=1}^{M} \beta_m \mathbb{I}(x \in T_m) \\ &= \sum_{m=1}^{M} \beta_m \prod_{p=1}^{P} \mathbb{I}(\gamma_{1pm} \leq x_p \leq \gamma_{2pm}) \end{aligned} \tag{19}$$

such that the basis expansions are $g_m(x, \gamma_{1m}, \gamma_{2m}) = \prod_{p=1}^{P} \mathbb{I}(\gamma_{1pm} \leq x_p \leq \gamma_{2pm})$; cf. e.g. CHERKASSKY and MULIER (1998, Section 5.3).

As the defined regions T_m are disjoint, the predicted output is equal to β_m for an object x in region T_m, so that the region specific prediction rule is $f(x, \beta | T_m) = \beta_m$. The empirical risk is defined as

$$r(f(x, \beta, T)) = \sum_{m=1}^{M} p(T_m) r_m(f(x, \beta | T_m)),$$

where $p(T_m)$ denotes the estimated probability of an object to fall into T_m. With a specified set of regions $T_1, ..., T_M$, minimization of empirical risk subject to β for the complete training sample can be achieved by separately minimizing the region specific risks

$$\begin{aligned} r_m(f(x, \beta | T_m)) = \frac{1}{n_{T_m}} \sum_{\{i | x_i \in T_m\}} L(g_i, f(x_i, \beta | T_m)) = \frac{1}{n_{T_m}} \sum_{\{i | x_i \in T_m\}} L(g_i, \beta_m), \\ \text{with} \quad n_{T_m} = |\{i | x_i \in T_m\}|, \end{aligned}$$

subject to β_m, cf. BREIMAN, FRIEDMAN, et al. (1984, Section 8.4). Prior to estimating β, optimal regions $\hat{T}_m$ have to be determined. These are distinct subsets of observations $x = \{x_i | i = 1, ..., n\}$ such that the empirical risk of (19) is minimal. Binary splits of x are performed stepwise defining new subsets (*nodes*). At each node a new split s, i.e. the variable X_j along with a cut-off a, has to be chosen from all possible splits S. Stopping the splitting process, one arrives at the *terminal nodes* that define the regions $\hat{T}_m$ and thus the final tree $\hat{T}$. The optimal split s^* at each non-terminal node t maximizes the difference between the minimum achievable risk of the current node and the weighted sum of minimum risks of its two descendant nodes. In classification trees, i.e. with G discrete, the classification decision is based on class proportions. The

optimal split depends on the proportion of objects from class 1 in the node, $p(1|\mathtt{t})$, which minimizes the node specific risk. Splitting criteria, called *impurity* measures, can therefore be interpreted as an empirical version of information in node $\mathtt{t}$

$$H_{\mathtt{t}}(p) = \frac{1}{n_{\mathtt{t}}} \sum_{\{i|x_i \in \mathtt{t}\}} L(g_i, p(1|\mathtt{t})).$$

With 0-1-loss, the impurity measure is given by misclassification rate

$$H_{\mathtt{t}}^{0,1}(p) = \min\{p(1|\mathtt{t}), 1 - p(1|\mathtt{t})\},$$

which does not provide a good splitting criterion (Breiman, Friedman, et al. (1984, Chapter 4)). Instead, entropy derived from log-loss is favored:

$$H_{\mathtt{t}}^{log}(p) = -p(1|\mathtt{t}) \log p(1|\mathtt{t}) - (1 - p(1|\mathtt{t})) \log(1 - p(1|\mathtt{t})).$$

Alternatively, a regression related approach with squared error loss leads to the Gini index

$$H_{\mathtt{t}}^{bs}(p) = \frac{1}{n_{\mathtt{t}}} \sum_{x_i \in \mathtt{t}} (g_i - p(1|\mathtt{t}))^2 = p(1|\mathtt{t})(1 - p(1|\mathtt{t})).$$

Splitting criteria based on squared error loss or log-loss imply equal misclassification costs. Typically, unequal misclassification costs are not implemented by the loss functions but by adapting prior probabilities π_g; compare Breiman, Friedman, et al. (1984, p. 28 and Section 4.4) and Therneau and Atkinson (2013). However, for adapting prior probabilities misclassification costs need to be fixed.

As impurity measures are minimum achievable risks, it is straightforward to consider alternative measures by altering the loss function in $H_{\mathtt{t}}(p)$. Buja, Stuetzle, et al. (2005) propose to use information measures $H_{\mathtt{t}}(p)$ based on beta-loss so that

$$H_{\mathtt{t}}^{\mathcal{B}(a,b)}(p) = p(1|\mathtt{t}) L_1^{\mathcal{B}(a,b)}(p(1|\mathtt{t})) + (1 - p(1|\mathtt{t})) L_0^{\mathcal{B}(a,b)}(p(1|\mathtt{t})).$$

In an empirical study, they compare the Gini index to $H^{\mathcal{B}(a,b)}$ with large differences between a and b. They reveal that trees, which are split with beta-loss, show a stronger concentration on one class. This leads to better interpretability but lower prediction performance. However, a direct comparison of these measures is not appropriate as they represent different assumptions about the ratio of misclassification costs. Beta-loss directly induces unequal misclassification costs and can be seen as an alternative to the adaptation of prior class probabilities to account for unequal misclassification costs. Thus, its role as surrogate loss function for generalized 0-1-loss also applies for this case. In contrast, the Gini index is used for equal misclassification costs so that a comparison to splitting with $H^{\mathcal{B}(a,b)}(p)$ is more appropriate with adapted priors. Figure 1 shows impurity $H^{\mathcal{B}(a,b)}(p)$ as a function of p for different choices of a and b. As $a = b$ implies the assumption of equal misclassification costs, the impurity is seen to achieve

its maximum for $p = 0.5$. For $a = b = 1$, the impurity measure is equal to the Gini index. For unequal parameters a and b, the maximum of $H^{\mathcal{B}(a,b)}(p)$ is shifted towards the value of p that is least favorable when $c_{0,1}/c_{1,0} = a/b$. For example, the maximum is 0.4242 for $a = 2$, $b = 3$ and 0.4040 for $a = 20, b = 30$. Thus, approaching generalized 0-1-loss with increasing values of a and b, the maximum is shifted towards $p = 0.4$, which is the least favorable class distribution when $p = p(1|\mathsf{t}) = p(0|\mathsf{t})/1.5$ and $c_{1,0}/c_{0,1} = 1.5$. Hence, the effect of $a/b = 2/3$, implying $c_{0,1}/c_{1,0} = 2/3$, is equivalent to the effect obtained by adapting the priors as described by Therneau and Atkinson (2013).

An impurity measure can also be derived from ranking loss. As ranking loss implies avoidance of a classification decision, its implementation as splitting criterion leading to an actual classification seems inappropriate.

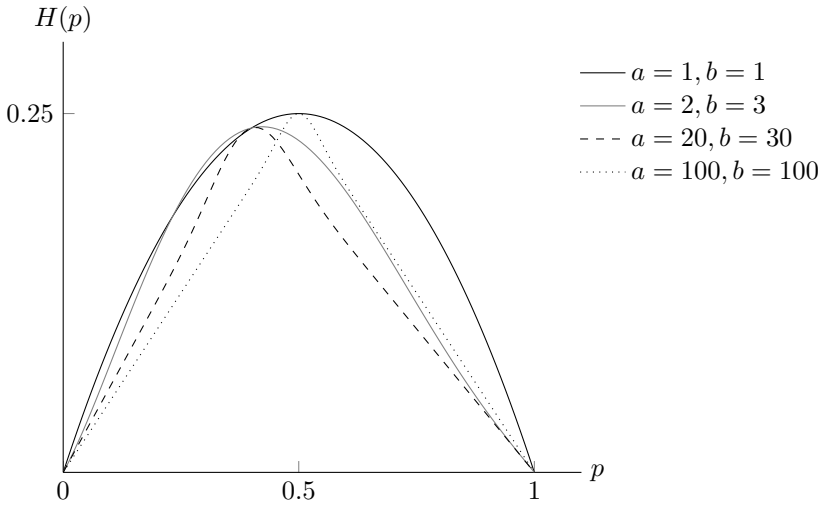

Figure 1: Impurity measure based on beta-loss for different choices of a and b as function of class 1 probability, $H^{\mathcal{B}(a,b)}(p)$.

The class of approximating functions considered with (19) is very wide. The fact that splitting a node always reduces empirical risk of the tree leads to overfitting and requires regularization. The standard approach, suggested by Breiman, Friedman, et al. (1984, Chapter 3), is known as *cost-complexity pruning* and represents another special form of the model selection procedure described above. A grown tree $\hat{T}$ is "pruned" to a smaller subtree by backward elimination of splits. Based on the full tree $\hat{T}$, for each value of λ a subtree is determined, which minimizes a penalized version of the empirical risk

$$r_\lambda(f(x, \beta, \hat{T})) = \sum_{m=1}^{M} p(T_m) r_m(f(x, \beta|\hat{T}_m)) + \lambda M.$$

Hence, the empirical risk of a tree is traded off against the number of its terminal nodes, i.e. the number of basis expansions, which define the tree's complexity. The tree that minimizes $r_\lambda(f(x, \beta, \hat{T}))$ is denoted as $\hat{T}_{\hat{\lambda}}$. The value of λ, for which T_λ achieves the

lowest estimated prediction error is chosen as the optimal solution $\hat{\lambda}$. The estimate of prediction error can be obtained by cross-validation. Therefore, the general procedure of determining $\hat{\lambda}$ as described above applies. Again, squared error loss is used for regression trees so that loss functions for splitting and pruning coincide. In contrast, pruning classification trees, or regression trees that provide class probability predictions $p(x, \beta, T)$, is commonly based on (generalized) 0-1-loss. In addition, optimal pruning of a tree can also be determined according costs sensitive criteria like beta-loss. This is a promising approach to model selection for trees as Breiman, Friedman, et al. (1984) note that the choice of the loss function is more influential in the pruning than in the splitting process.

3.2.3 Embedded methods with early stopping: Gradient-descent boosting

Boosting is a form of ensemble learning, which aims at combining classifiers with low classification performance to a classifier with strongly increased performance. However, the method can also be interpreted as a regression problem with basis expansion functions that employs a stagewise optimization; compare e.g. Hastie, Tibshirani, et al. (2009, Chapter 10) and Berk (2008, Chapter 6). In boosting applications, the basis expansions are often called *base learners*. As before, the optimal prediction rule is estimated by minimizing empirical risk such that

$$\begin{aligned} \hat{f}(x) &= f(x, \hat{\beta}, \hat{\gamma}) \\ \text{with} \quad (\hat{\beta}, \hat{\gamma}) &= \underset{\beta,\gamma}{\operatorname{argmin}} \sum_{i=1}^{n} L\left(g_i, \sum_{m=1}^{M} g_m(x, \gamma_m)\beta_m\right). \end{aligned}$$

In its original version, the AdaBoost algorithm of Freund and Schapire (1995), each of the basis expansions is a classifier trained from iteratively reweighted versions of the training sample. Employing a generalized version, where the base learners provide probability estimates such that $g_m(x) : \mathcal{X} \mapsto [0, 1]$, Friedman, Hastie, et al. (2000) show that AdaBoost amounts to estimating $f(x, \beta, \gamma)$ by minimizing empirical risk derived from exponential loss in its original form,

$$L(g, f(x)) = \exp(-(2g - 1)f(x)),$$

employing a classification rule $d(x) = \mathbb{I}(f(x) > 0)$. This reformulation of the AdaBoost algorithm allows the procedure to be interpreted as stagewise fitting of an additive model by greedy optimization. The interpretation has lead to approaches employing alternative loss functions and forms of g_m.

Precisely, fitting the regression function to the data by forward stagewise additive modeling works as follows (cf. Hastie, Tibshirani, et al. (2009, p. 342) and Friedman (2001)):

1. Initialize $f_0 = 0$,
2. for $m = 1$ to M

a) compute

$$(\beta_m, \gamma_m) = \underset{\beta_0, \gamma_0}{\operatorname{argmin}} \sum_{i=1}^{n} L(g_i, f_{m-1}(x_i) + \beta_0 g_m(x_i, \gamma_0)),$$

b) set $f_m(x) = f_{m-1}(x) + \beta_m g_m(x, \gamma_m)$.

The basis expansion and its coefficient obtained at stage m are not changed in subsequent stages so that $f_M(x) = \hat{f}(x)$.

Variants of boosting emerge from different choices for the loss function L, from considered sets of basis expansions and from methodological differences in the optimization in 2.a); compare e.g. Friedman, Hastie, et al. (2000). A common approach for 2.a), which shall be focused in the remainder, leads to *gradient boosting*, where each step builds on the the negative gradient of L with regard to $f(x_i, \beta, \gamma)$. Mathematical details on the procedure are provided by Friedman (2001).

Generalizing AdaBoost in that way, Friedman, Hastie, et al. (2000), Friedman (2001) and Bühlmann and Hothorn (2007) propose the use of alternative base learners, such as simple regression functions or decision trees. They also employ alternative loss functions to handle classification as well as regression problems. Practical implementation of these ideas for is provided via the `R`-package `mboost` (cf. Hothorn, Bühlmann, et al. (2013)), whose flexibility allows for further adaptations; cf. Hofner, Mayr, et al. (2012).

Again, continuous loss functions are necessary as greedy optimization procedures in boosting employ the derivative of the loss function. According to Bühlmann and Yu (2000), the exponential loss is a very suitable approximation to 0-1-loss. Unequal misclassification costs are usually implemented through altered prior probabilities by giving different weights to objects in the training sample, cf. Landesa-Vázquez and Alba-Castro (2012). A more direct approach is taken by Masnadi-Shirazi and Vasconcelos (2007), who extend the exponential loss by implementing cost-weights in the following form

$$L_0^{cexp}(f(x)) = \mathbb{I}(g = 0) \exp(-c_{0,1}(2g - 1)f(x))$$
$$L_1^{cexp}(f(x)) = \mathbb{I}(g = 1) \exp(-c_{1,0}(2g - 1)f(x)),$$

denoted as cost adjusted exponential loss in the remainder. Masnadi-Shirazi and Vasconcelos (2007) demonstrate that the resulting classifier is actually cost-sensitive. However, they require the misclassification costs or their ratio to be known exactly.

As before, beta-loss can be used to approximate generalized 0-1-loss. Boosting with beta-loss and classification trees as base learners is treated by Shen (2005). Based on an empirical comparison with simulated and real data, he concludes that beta-loss combined with complex base learners is less beneficial in terms of misclassification costs when compared to log-loss or to approaches with altered priors. Subsequent analyses therefore focus on a comparison to cost adjusted exponential loss. Unknown misclassification costs, as reflected by ranking loss, can also provide an optimization

criterion in gradient boosting, which is for example proposed by Komori (2011) and is also implemented in `mboost`. In `mboost`, a smoothed version of $1 - AUC$ is used; cf. Hofner, Mayr, et al. (2012) and Hothorn, Bühlmann, et al. (2013).

Complexity of the estimated regression function depends on the number of iterations in the greedy optimization procedure as these determine the number M of basis functions g_m. Regularization is implemented by an early stopping of the iteration at stage m_{stop}. Thus, determination of the optimal stopping stage amounts to the model selection part in boosting. In terms of the general procedure above, λ now indexes the number of iterations until stopping. Cost sensitive model selection concerning misclassification costs can again be implemented by choosing an appropriate loss function in r_{CV}.

4 Simulative comparison

Simulation studies are performed to investigate the beta-loss for constructing classification rules in CART and gradient-descent boosting. In each method, classification rules are derived from a score or a probability prediction by employing suitable cut-off values, i.e. $\hat{d}(x) = \mathbb{I}(\hat{p}(x) > c^*)$. The classification rules, which are constructed from a training sample, are evaluated by their prediction error for an independent test sample. To evaluate classification performance, prediction error is computed for 0-1-loss (classification error) and generalized 0-1-loss (misclassification costs).

A two-class-problem with $X = (X_1, ..., X_6)'$ is considered. Data are simulated for two scenarios with increasing difficulty in constructing a classification rule. The scenarios are defined by the following class conditional distributions of X, where I denotes the identity matrix:

- Scenario 1:

  $X|G = 0 \sim N_6(\mu_0, I), \quad \mu_0 = (0, 0, 1.5, 1.5, 3, 3)'$
  $X|G = 1 \sim N_6(\mu_1, I), \quad \mu_1 = \mu_0 + (0.5, 1, 0.5, 1, 0.5, 1)'$

- Scenario 2:

  $X|G = 0 \sim N_6(\mu_0, 4 \cdot I), \quad \mu_0 = (0, 0, 0, 0, 0, 0)'$
  $X|G = 1 \sim N_6(\mu_1, I), \quad \mu_1 = (\delta, \delta, \delta, \delta, \delta, \delta)', \ \delta = 1/\sqrt{6}$

Simulations are performed with 500 iterations using the follow design for both scenarios:

- Size of training sample: $n = 100, 200$, with equal class proportions
- Size of test sample: $n_{test} = 5000$, with equal class proportions
- Assumed misclassification costs (where applicable): $c_{0,1} = 10$, $c_{1,0} = 100$

As class conditional covariances are equal in the first scenario, the logistic model holds with $\beta = (-8.625, -0.5, -1.0, -0.5, -1.0, -0.5, -1.0)'$ when $\pi_0 = \pi_1$. The second scenario corresponds to the "ringnorm" data, which pose a more complex classification

task than Scenario 1 and are commonly used as benchmark in classification; cf. Lugosi and Vayatis (2004) and Leisch and Dimitriadou (2013).

Performance is evaluated by test sample prediction errors for 0-1-loss and generalized 0-1-loss with correctly and falsely specified $c_{0,1}$ and $c_{1,0}$. Cut-offs for classification rule construction $\hat{d}(x) = \mathbb{I}(\hat{p}(x) > c^*)$ were chosen to be optimal for misclassification costs implicitly assumed by the choice of the evaluation criterion. Precisely, evaluating a classification rule in terms of classification error, i.e. prediction error with 0-1-loss implies the assumption of equal misclassification costs. Thus, the optimal cut-off for a probability prediction is $c^* = 1/2$. If prediction error is based on generalized 0-1-loss, unequal misclassification costs are assumed. When constructing the rule, misclassification costs are taken as $c_{0,1} = 10$, $c_{1,0} = 100$, which leads to an optimal cut-off of $c^* = 1/11$. To show the effects of falsely specified misclassification costs, evaluation of classification rules is also made by test sample misclassification costs calculated for $c_{0,1} = 11$, $c_{1,0} = 90$ and $c_{0,1} = 9$, $c_{1,0} = 110$.

4.1 CART

The standard procedure of CART is implemented in the `R`-routine `rpart` (cf. Therneau, Atkinson, et al. (2013)). For the two scenarios, trees are constructed to provide probability estimates. Standard splitting criteria, entropy (H^{log}) and Gini index (H^{bs}), provided by `rpart` are compared to a splitting criterion based on beta-loss, $H^{\mathcal{B}(1,10)}$. It is added as an external splitting function; as described by Therneau (2013). As $H^{\mathcal{B}(1,10)}$ implies unequal misclassification costs, a comparison with cost adjusted priors as implemented in the standard routine is most appropriate. Thus, the Gini index with cost adjustment ($H^{bs}(cost)$) is also considered. In addition, pruning is performed with different loss functions. Precisely, pruning criteria are 5-fold cross-validatory estimates of prediction error based on 0-1-loss ($L^{1,1}$), generalized 0-1-loss ($L^{10,100}$), beta-loss ($L^{\mathcal{B}(1,10)}$), and Brier score L^{bs}.

An overall improvement in performance for all criteria is achieved with increased sample size but results do not differ when comparing splitting and pruning criteria. Hence, only the case of $n = 100$ is reported here.

In general, it has to be noted that the constructed trees show very weak classification performance in terms of costs. Test sample errors for generalized 0-1-loss are always higher than those that would be achieved when all subjects were assigned to the more expensive class. Only in some scenarios, CART achieves similar costs in the test sample. The difficulty arises due to imbalanced misclassification costs. None of the approaches, including implementations of appropriate beta-loss, overcomes this problem. However, some comparative results are summarized in the following.

In accordance with Breiman, Friedman, et al. (1984), splitting criteria only differ slightly in classification performance in most cases, while larger differences can be found for pruning criteria In terms of classification error, reported in Tables 1 and 2, splitting criteria have no influence. Pruning with the underlying assumption of unequal misclassification costs leads to low performance . As expected, 0-1-loss is the

most appropriate criterion for pruning when misclassification costs are equal.

	H^{log}	H^{bs}	H^{bs} (cost)	$H^{\mathcal{B}(1,10)}$
$L^{1,1}$	0.322	0.333	0.345	0.345
$L^{10,100}$	0.343	0.350	0.490	0.420
L^{bs}	0.413	0.406	0.418	0.408
$L^{\mathcal{B}(1,10)}$	0.482	0.478	0.471	0.489

Table 1: Test sample classification error of CART for Scenario 1 with $n = 100$.

	H^{log}	H^{bs}	H^{bs} (cost)	$H^{\mathcal{B}(1,10)}$
$L^{1,1}$	0.315	0.325	0.293	0.284
$L^{10,100}$	0.404	0.412	0.476	0.395
L^{bs}	0.408	0.420	0.389	0.338
$L^{\mathcal{B}(1,10)}$	0.443	0.446	0.426	0.477

Table 2: Test sample classification error of CART for Scenario 2 with $n = 100$.

More importantly, classification performance for unequal misclassification costs can be improved by employing corresponding pruning criteria, i.e. criteria based on generalized 0-1-loss or beta-loss. The tree is built assuming $c_{0,1}/c_{1,0} = 1/10$. For correctly specified misclassification costs (Tables 3 and 4), pruning with $L^{\mathcal{B}(1,10)}$ leads to the best predictions when considering all splitting criteria. Costs are slightly lower for splitting with beta-loss. Pruning with generalized 0-1-loss $L^{10,100}$ only leads to desirable predictions when combined with splitting based on the Gini index with cost adjusted priors. In this case, classification performance in terms of misclassification costs is often slightly better than that achieved with beta-loss in splitting and pruning. Results for falsely specified misclassification costs, i.e. when the ratio of costs is over- or underestimated, differ between the scenarios. Tables 5 and 6 report misclassification costs, if the true costs are $c_{0,1} = 9$ and $c_{1,0} = 110$ so that the ratio is overestimated when building the tree. Tables 7 and 8 report the costs for the case of underestimation. In Scenario 1 (Tables 5 and 7), pruning with beta-loss combined with standard splitting criteria leads to minimum misclassification costs. In Scenario 2 (Tables 6 and 8), pruning with generalized 0-1-loss and prior adjusted Gini splitting leads to the lowest cost. However, they are only slightly lower than those achieved by beta-loss splitting and pruning. Thus, accounting for uncertain misclassification costs as implied by beta-loss does not always improve misclassification costs if the classification rule is constructed under incorrect assumptions about the cost ratio. However, on an overall scale, beta-loss is a compatible choice as pruning and splitting criterion for building optimal classification trees when misclassification costs are unequal.

	H^{log}	H^{bs}	H^{bs} (cost)	$H^{\mathcal{B}(1,10)}$
$L^{1,1}$	7.581	7.333	9.279	8.144
$L^{10,100}$	7.074	6.715	5.158	6.259
L^{bs}	5.246	5.147	6.194	6.581
$L^{\mathcal{B}(1,10)}$	5.103	5.054	5.438	5.188

Table 3: Test sample misclassification cost of CART with $c_{0,1} = 10$, $c_{1,0} = 100$ for Scenario 1 with $n = 100$.

	H^{log}	H^{bs}	H^{bs} (cost)	$H^{\mathcal{B}(1,10)}$
$L^{1,1}$	7.586	7.466	7.453	6.996
$L^{10,100}$	6.166	6.026	5.229	5.847
L^{bs}	6.009	5.887	5.922	6.472
$L^{\mathcal{B}(1,10)}$	5.690	5.676	5.639	5.245

Table 4: Test sample misclassification cost of CART with $c_{0,1} = 10$, $c_{1,0} = 100$ for Scenario 2 with $n = 100$.

	H^{log}	H^{bs}	H^{bs} (cost)	$H^{\mathcal{B}(1,10)}$
$L^{1,1}$	7.642	7.355	9.640	8.307
$L^{10,100}$	7.004	6.577	4.699	6.065
L^{bs}	4.803	4.676	6.012	6.447
$L^{\mathcal{B}(1,10)}$	4.624	4.566	5.055	4.735

Table 5: Test sample misclassification cost of CART with $c_{0,1} = 9$, $c_{1,0} = 110$ for Scenario 1 with $n = 100$.

	H^{log}	H^{bs}	H^{bs} (cost)	$H^{\mathcal{B}(1,10)}$
$L^{1,1}$	7.817	7.677	7.709	7.213
$L^{10,100}$	6.008	5.842	4.811	5.684
L^{bs}	5.838	5.659	5.781	6.508
$L^{\mathcal{B}(1,10)}$	5.401	5.376	5.381	4.825

Table 6: Test sample misclassification cost of CART with $c_{0,1} = 9$, $c_{1,0} = 110$ for Scenario 2 with $n = 100$.

	H^{log}	H^{bs}	H^{bs} (cost)	$H^{\mathcal{B}(1,10)}$
$L^{1,1}$	7.620	7.445	8.917	7.981
$L^{10,100}$	7.189	6.893	5.618	6.454
L^{bs}	5.699	5.619	6.377	6.715
$L^{\mathcal{B}(1,10)}$	5.581	5.543	5.821	5.641

Table 7: Test sample misclassification cost of CART with $c_{0,1} = 11$, $c_{1,0} = 90$ for Scenario 1 with $n = 100$.

	H^{log}	H^{bs}	H^{bs} (cost)	$H^{\mathcal{B}(1,10)}$
$L^{1,1}$	7.499	7.513	7.196	6.780
$L^{10,100}$	6.370	6.320	5.647	6.009
L^{bs}	6.242	6.207	6.064	6.435
$L^{\mathcal{B}(1,10)}$	6.005	6.040	5.897	5.665

Table 8: Test sample misclassification cost of CART with $c_{0,1} = 11$, $c_{1,0} = 90$ for Scenario 2 with $n = 100$.

4.2 Gradient-descent boosting

Second, an analysis for component-wise gradient-descent boosting as implemented in the `R`-routine `mboost` is performed. The original predictive variables are used as base learners. The performance of boosting with different loss functions is compared in terms of test sample prediction errors as in the analyses before. Log-loss (*log*), Brier score (*bs*), which have been demonstrated to be approximations to 0-1-loss, are implemented as user written loss functions. Exponential loss (*exp*) is slightly modified to obtain probability estimates instead of class predictions. Cost adjusted exponential loss of Masnadi-Shirazi and Vasconcelos (2007) (*cexp*) and beta-loss $\mathcal{B}(1, 10)$ account for unequal misclassification costs and are also implemented as user written loss functions. Ranking loss (*auc*) is used as implemented in `mboost`. The procedure also includes model selection by early stopping. From a maximum number of steps $m_{max} = 1000$, the optimal value of $\hat{m}_{stop}$ is determined by 5-fold cross-validation as implemented in `mboost`. The loss function used in cross-validatory model selection is equal to the one used in the optimization. Evaluation is based on the same types of prediction error as in previous analysis.

Due to some differences, results for both sample sizes are reported here. Tables 9 and 10 show test sample prediction errors of the estimated probability predictions and classification rules for Scenario 1. Log-loss performs best regardless of the evaluation criterion since boosting with log-loss is comparable to logistic regression if the original variables are taken as base learners; cf. Hofner, Mayr, et al. (2012). Exponential loss and Brier score perform very similar in terms of probability predictions ($L^{\mathcal{B}(1,10)}$

and L^{bs}). But performance of the Brier score in terms of misclassification costs is weaker than for the other two. In general, misclassification costs are also higher when employing cost adjusted exponential or beta-loss as criterion in optimization and early stopping. While *cexp* performs better than beta-loss for $n = 100$, classification error and misclassification costs are lower with beta-loss when the sample size is increased. Surprisingly, *auc* estimates do not perform as well, which can be due to overfitting as no cross-validation could be performed to determine the optimal stopping stage. Instead, the number of boosting iterations was fixed at $m_{stop} = 1000$.

	$L^{1,1}$	$L^{10,100}$	$L^{9,110}$	$L^{11,90}$	$L^{\mathcal{B}(1,10)}$	L^{bs}
log	0.1833	4.0518	3.7776	4.3259	0.3629	0.1291
exp	0.1856	4.1711	3.9235	4.4186	0.3742	0.1315
bs	0.1846	4.2977	3.9039	4.6915	0.3764	0.1330
cexp	0.2119	4.4050	4.0358	4.7742	0.3971	0.1468
$\mathcal{B}(1, 10)$	0.2295	4.5746	4.3304	4.8188	0.4245	0.1604
auc	0.1893	4.6941	4.6565	4.7317	-	-

Table 9: Test sample prediction errors of boosting with early stopping for Scenario 1 with $n = 100$.

	$L^{1,1}$	$L^{10,100}$	$L^{9,110}$	$L^{11,90}$	$L^{\mathcal{B}(1,10)}$	L^{bs}
log	0.1744	3.8825	3.6057	4.1593	0.3475	0.1231
exp	0.1757	3.9278	3.6879	4.1678	0.3530	0.1242
bs	0.1760	4.3098	3.9020	4.7177	0.3730	0.1290
cexp	0.2003	4.3259	3.9382	4.7137	0.3858	0.1392
$\mathcal{B}(1, 10)$	0.1996	4.1271	3.8709	4.3833	0.3769	0.1392
auc	0.1797	4.1451	4.0141	4.2760	-	-

Table 10: Test sample prediction errors of boosting with early stopping for Scenario 1 with $n = 200$.

In the more difficult classification task of Scenario 2, displayed in Tables 11 and 12, employment of cost adjusted exponential loss and beta-loss proves beneficial in terms of misclassification costs. Beta-loss outperforms *cexp* leading to the lowest misclassification costs even if they were misspecified. Especially with increased sample size, differences between beta-loss and other loss functions is striking. For *auc*, increased sample size leads to considerable classification performance with similar misclassification costs as obtained by beta-loss.

	$L^{1,1}$	$L^{10,100}$	$L^{9,110}$	$L^{11,90}$	$L^{\mathcal{B}(1,10)}$	L^{bs}
log	0.3964	4.9454	4.4510	5.4397	0.4785	0.2415
exp	0.3991	4.9655	4.4692	5.4617	0.4825	0.2416
bs	0.3924	4.9367	4.4431	5.4303	0.4749	0.2410
cexp	0.3988	4.8922	4.4036	5.3807	0.4742	0.2570
$\mathcal{B}(1,10)$	0.4124	4.6591	4.2466	5.0716	0.4660	0.2693
auc	0.3665	5.0348	4.7847	5.2849	-	-

Table 11: Test sample prediction errors of boosting with early stopping for Scenario 2 with $n = 100$.

	$L^{1,1}$	$L^{10,100}$	$L^{9,110}$	$L^{11,90}$	$L^{\mathcal{B}(1,10)}$	L^{bs}
log	0.3774	4.9648	4.4683	5.4612	0.4751	0.2351
exp	0.3817	4.9701	4.4731	5.4671	0.4782	0.2360
bs	0.3759	4.9687	4.4718	5.4656	0.4743	0.2349
cexp	0.3691	4.9381	4.4443	5.4319	0.4742	0.2492
$\mathcal{B}(1,10)$	0.4054	4.5256	4.1088	4.9423	0.4516	0.2659
auc	0.3518	4.5747	4.2582	4.8912	-	-

Table 12: Test sample prediction errors of boosting with early stopping for Scenario 2 with $n = 200$.

5 Conclusion

Generalized 0-1-, beta- and ranking loss allow handling of complete, incomplete and no information about misclassification costs, respectively. The loss functions can be used to adapt regression based classifications methods as they enter estimation as well as model selection criteria. Especially, beta-loss provides a suitable model selection criterion approximating generalized 0-1-loss. This enables cost sensitive model selection for methods that provide probability estimates instead of pure classification rules so that uncertainty about misclassification costs can be accounted for methodologically.

Summarizing the analyses above, distinct results regarding the effect of loss functions can be stated for the classification methods. In CART, splitting with a beta-loss based criterion only leads to small improvements while pruning with beta-loss has stronger effects. In the simpler scenario, pruning with beta-loss, especially when combined with beta-loss or Gini splitting, leads to lowest misclassification costs even if these are misspecified when building the tree. In many cases, Gini splitting with cost adjusted priors and pruning with generalized 0-1-loss achieves similar prediction performance in terms of misclassification costs. Hence, it is not clear which of these two combinations is more advantageous. In general, classification performance of CART was very low in both scenarios such that considering other scenarios could lead to more distinct results.

Boosting and early stopping based on beta-loss is more beneficial especially for increased sample size. It reveals the lowest misclassification costs in the more difficult classification task even if these are falsely specified when constructing the classification rule. For the smaller sample size, beta-loss shows classification performance similar to that of cost adjusted exponential loss.

Appendix A

Proof of Proposition 1. The first part directly follows from the result in Clémençon, Lugosi, et al. (2008, Example 1) that the expected loss $E[L^r(G, s(X))]$ of a ranking rule $r(x_1, x_2) = \mathbb{I}(s(x_1) > s(x_2))$ is minimized if s a strictly increasing transformation of $\pi(x)$ and therefore also by $s(x) = \pi(x)$.

The second part derives from calculations based on Hand (2010): With

$$\begin{aligned} AUC_s &= \int_{-\infty}^{\infty} (1 - F_{s_1}(t)) f_{s_0}(t)\, \mathrm{d}t \\ &= \int_{-\infty}^{\infty} \Big(\int_t^{\infty} f_{s_1}(u)\, \mathrm{d}u \Big) f_{s_0}(t)\, \mathrm{d}t, \end{aligned}$$

the expected loss $E[L^r(G, s(X))]$ can be written as

$$\begin{aligned} E[L^r(G, s(X))] &= 2 \left(\int_{-\infty}^{\infty} f_{s_0}(t)\, \mathrm{d}t - \int_{-\infty}^{\infty} \left(\int_t^{\infty} f_{s_1}(u)\, \mathrm{d}u \right) f_{s_0}(t)\, \mathrm{d}t \right) \pi_0 \pi_1 \\ &= 2 \left(\int_{-\infty}^{\infty} \left(1 - \int_t^{\infty} f_{s_1}(u)\, \mathrm{d}u \right) f_{s_0}(t)\, \mathrm{d}t \right) \pi_0 \pi_1 \\ &= \left(\int_{-\infty}^{\infty} \left(\int_u^{\infty} f_{s_0}(t)\, \mathrm{d}t \right) f_{s_1}(u)\, \mathrm{d}u + \int_{-\infty}^{\infty} (F_{s_1}(t)) f_{s_0}(t)\, \mathrm{d}t \right) \pi_0 \pi_1 \\ &= \left(\int_{-\infty}^{\infty} (1 - F_{s_0}(u)) f_{s_1}(u)\, \mathrm{d}u + \int_{-\infty}^{\infty} (F_{s_1}(t)) f_{s_0}(t)\, \mathrm{d}t \right) \pi_0 \pi_1 \\ &= \int_{-\infty}^{\infty} (\pi_0 \pi_1 (1 - F_{s_0}(t)) f_{s_1}(t) + \pi_0 \pi_1 (F_{s_1}(t)) f_{s_0}(t))\, \mathrm{d}t \\ &= \int_{-\infty}^{\infty} (P_1(t) \pi_0 (1 - F_{s_0}(t)) + (1 - P_1(t)) \pi_1 (F_{s_1}(t))) \\ &\quad \cdot (\pi_0 f_{s_0}(t) + \pi_1 f_{s_1}(t))\, \mathrm{d}t, \\ &\text{where} \quad P_1(t) = \frac{\pi f_{s_1}(t)}{(1 - \pi) f_{s_0}(t) + \pi f_{s_1}(t)}. \end{aligned}$$

A variable change from t to c, where $t = P_1^{-1}(c)$, yields

$$\begin{aligned} E[L^r(G, s(X))] = \int_0^1 &\Big(c(1 - \pi)(1 - F_{s_0}(P_1^{-1}(c))) + (1 - c)\pi(F_{s_1}(P_1^{-1}(c))) \Big) \\ &\cdot \Big(\pi_0 f_{s_0}(P_1^{-1}(c)) + \pi_1 f_{s_1}(P_1^{-1}(c)) \Big) (P_1^{-1})'(c)\, \mathrm{d}c, \end{aligned}$$

where $(P_1^{-1})'(c)$ is the first derivative of $P_1^{-1}(c)$. For $s(x) = p(x)$, this is equal to expression (6) with

$$w(c) = (1 - \pi) f_{p_0}(P_1^{-1}(c))(P_1^{-1})'(c) + \pi f_{p_1}(P_1^{-1}(c))(P_1^{-1})'(c) = w_r(c).$$

□

References

Berger, J. O. (1985). *Statistical Decision Theory and Bayesian Analysis*. 2nd ed. Springer: New York.

Berk, R. A. (2008). *Statistical Learning from a Regression Perspective*. Springer: New York.

Breiman, L., Friedman, J. H., Olshen, R. A., and Stone, C. J. (1984). *Classification and Regression Trees*. Chapman & Hall: Boca Raton.

Bühlmann, P. and Hothorn, T. (2007). "Boosting algorithms: Regularization, prediction and model fitting". In: *Statistical Science* 22(4), pp. 477–505.

Bühlmann, P. and Yu, B. (2000). "Discussion on 'Additive logistic regression: A statistical view' by Friedman, J. H., Hastie, T., Tibshirani, R." In: *The Annals of Statistics* 28(2), pp. 377–386.

Buja, A., Stuetzle, W., and Shen, Y. (2005). *Loss functions for binary class probability estimation and classification: Structure and applications*. Working Paper. Wharton School, University of Pennsylvania. URL: http://stat.wharton.upenn.edu/~buja/PAPERS/paper-proper-scoring.pdf (visited on 11/12/2013).

Cherkassky, V. S. and Mulier, F. (1998). *Learning from Data: Concepts, Theory, and Methods*. John Wiley & Sons: New York.

Clémençon, S., Lugosi, G., and Vayatis, N. (2008). "Ranking and empirical minimization of U-statistics". In: *Annals of Statistics* 36(2), pp. 844–874.

Dawid, A. P. (2007). "The geometry of proper scoring rules". In: *Annals of the Institute of Statistical Mathematics* 59(1), pp. 77–93.

Freund, Y. and Schapire, R. E. (1995). "A decision-theoretic generalization of on-line learning and an application to boosting". In: *Computational Learning Theory*. Ed. by Vitányi, P. Vol. 904. Springer, pp. 23–37.

Friedman, J. H. (1997). "On bias, variance, 0/1-loss, and the curse of dimensionality". In: *Data Mining and Knowledge Discovery* 1(1), pp. 55–77.

Friedman, J. H. (2001). "Greedy function approximation: A gradient boosting machine". In: *The Annals of Statistics* 29(5), pp. 1189–1232.

Friedman, J. H., Hastie, T., and Tibshirani, R. (2000). "Additive logistic regression: A statistical view of boosting (with discussion)". In: *The Annals of Statistics* 28(2), pp. 337–407.

Gneiting, T. and Raftery, A. E. (2007). "Strictly proper scoring rules, prediction, and estimation". In: *Journal of the American Statistical Association* 102(477), pp. 359–378.

Green, D. M. and Swets, J. A. (1966). *Signal Detection Theory and Psychophysics*. John Wiley & Sons: New York.

Guyon, I. and Elisseeff, A. (2003). "An introduction to variable and feature selection". In: *Journal of Machine Learning Research* 3, pp. 1157–1182.

Hand, D. J. (1997). *Construction and Assessment of Classification Rules*. John Wiley & Sons: Chichester.

Hand, D. J. (2009). "Measuring classifier performance: A coherent alternative to the area under the ROC curve". In: *Machine Learning* 77(1), pp. 103–123.

Hand, D. J. (2010). "Evaluating diagnostic tests: The area under the ROC curve and the balance of errors". In: *Statistics in Medicine* 29(14), pp. 1502–1510.

Hand, D. J. and Vinciotti, V. (2003). "Local versus global models for classification problems: Fitting models where it matters". In: *The American Statistician* 57(2), pp. 124–131.

Hastie, T., Tibshirani, R., and Friedman, J. H. (2009). *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. 2nd ed. Springer: New York.

Hofner, B., Mayr, A., Robinzonov, N., and Schmid, M. (2012). "Model-based boosting in R: A hands-on tutorial using the R package mboost". In: *Computational Statistics within Clinical Research*, pp. 1–33.

Hothorn, T., Bühlmann, P., Kneib, T., Schmid, M., Hofner, B., Sobotka, F., and Scheipl, F. (2013). *R-Package 'mboost' 2.2-2: Model-based boosting*. CRAN. URL: http://cran.r-project.org/web/packages/mboost/ (visited on 08/14/2013).

Huang, X., Qin, G., and Fang, Y. (2011). "Optimal combinations of diagnostic tests based on AUC". In: *Biometrics* 67(2), pp. 568–576.

Komori, O. (2011). "A boosting method for maximization of the area under the ROC curve". In: *Annals of the Institute of Statistical Mathematics* 63(5), pp. 961–979.

Landesa-Vázquez, I. and Alba-Castro, J. L. (2012). "Shedding light on the asymmetric learning capability of AdaBoost". In: *Pattern Recognition Letters* 33(3), pp. 247–255.

Leisch, F. and Dimitriadou, E. (2013). *R-Package 'mlbench' 2.1-1*. CRAN. URL: http://cran.r-project.org/web/packages/mlbench/mlbench.pdf (visited on 09/06/2013).

Lugosi, G. and Vayatis, N. (2004). "On the Bayes-risk consistency of regularized boosting methods". In: *Annals of Statistics* 32(1), pp. 30–55.

Ma, S. and Huang, J. (2005). "Regularized ROC method for disease classification and biomarker selection with microarray data". In: *Bioinformatics* 21(24), pp. 4356–4362.

Ma, S. and Huang, J. (2007). "Combining multiple markers for classification using ROC". In: *Biometrics* 63(3), pp. 751–757.

Ma, S. and Huang, J. (2008). "Penalized feature selection and classification in bioinformatics". In: *Briefings in Bioinformatics* 9(5), pp. 392–403.

Masnadi-Shirazi, H. and Vasconcelos, N. (2011). "Cost-sensitive boosting". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* 33(2), pp. 294–309.

Masnadi-Shirazi, H. and Vasconcelos, N. (2007). "Asymmetric boosting". In: *Proceedings of the 24th International Conference on Machine Learning*. Ed. by Ghahramani, Z. ACM: New York, pp. 609–619.

McIntosh, M. W. and Pepe, M. S. (2002). "Combining several screening tests: Optimality of the risk score". In: *Biometrics* 58(3), pp. 657–664.

McLachlan, G. J. (1992). *Discriminant Analysis and Statistical Pattern Recognition*. John Wiley & Sons: New York.

Pepe, M. S. and Thompson, M. L. (2000). "Combining diagnostic test results to increase accuracy". In: *Biostatistics* 1(2), pp. 123–140.

Schervish, M. J. (1989). "A general method for comparing probability assessors". In: *The Annals of Statistics* 17(4), pp. 1856–1879.

Shen, Y. (2005). "Loss functions for binary classification and class probability estimation". PhD thesis. Philadelphia (PA): University of Pennsylvania. URL: http://stat.wharton.upenn.edu/~buja/PAPERS/yi-shen-dissertation.pdf (visited on 11/12/2013).

Therneau, T. M. (2013). *User written splitting functions for rpart*. Mayo Clinic. URL: http://cran.r-project.org/web/packages/rpart/vignettes/usercode.pdf (visited on 09/06/2013).

Therneau, T. M. and Atkinson, E. J. (2013). *An introduction to recursive partioning using the rpart routines*. Mayo Foundation. URL: http://cran.r-project.org/web/packages/rpart/vignettes/longintro.pdf (visited on 11/12/2013).

Therneau, T. M., Atkinson, E. J., and Ripley, B. (2013). *R-Package 'rpart' 4.1-1*. CRAN. URL: http://cran.r-project.org/web/packages/rpart/rpart.pdf (visited on 09/06/2013).

Wang, Y., Chen, H., Schwartz, T., Duan, N., Parcesepe, A., and Lewis-Fernández, R. (2011). "Assessment of a disease screener by hierarchical all-subset selection using area under the receiver operating characteristic curves". In: *Statistics in Medicine* 30(14), pp. 1751–1760.

Zhou, X. H., Chen, B., Xie, Y. M., Tian, F., Liu, H., and Liang, X. (2012). "Variable selection using the optimal ROC curve: An application to a traditional Chinese medicine study on osteoporosis disease". In: *Statistics in Medicine* 31(7), pp. 628–635.

Diskussionspapiere 2013
Discussion Papers 2013

01/2013 **Wrede, Matthias:** Rational choice of itemized deductions

02/2013 **Wrede, Matthias:** Fair Inheritance Taxation in the Presence of Tax Planning

03/2013 **Tinkl, Fabian:** Quasi-maximum likelihood estimation in generalized polynomial autoregressive conditional heteroscedasticity models

04/2013 **Cygan-Rehm, Kamila:** Do Immigrants Follow Their Home Country's Fertility Norms?

05/2013 **Ardelean, Vlad and Pleier, Thomas:** Outliers & Predicting Time Series: A comparative study

06/2013 **Fackler, Daniel and Schnabel, Claus:** Survival of spinoffs and other startups: First evidence for the private sector in Germany, 1976-2008

07/2013 **Schild, Christopher-Johannes:** Do Female Mayors Make a Difference? Evidence from Bavaria

08/2013 **Brenzel, Hanna, Gartner, Hermann and Schnabel Claus:** Wage posting or wage bargaining? Evidence from the employers' side

09/2013 **Lechmann, D. S. and Schnabel C.:** Absence from work of the self-employed: A comparison with paid employees

10/2013 **Bünnings, Ch. and Tauchmann, H.:** Who Opts Out of the Statutory Health Insurance? A Discrete Time Hazard Model for Germany

Diskussionspapiere 2012
Discussion Papers 2012

01/2012 **Wrede, Matthias:** Wages, Rents, Unemployment, and the Quality of Life

02/2012 **Schild, Christopher-Johannes:** Trust and Innovation Activity in European Regions - A Geographic Instrumental Variables Approach

03/2012 **Fischer, Matthias:** A skew and leptokurtic distribution with polynomial tails and characterizing functions in closed form

04/2012 **Wrede, Matthias:** Heterogeneous Skills and Homogeneous Land: Segmentation and Agglomeration

05/2012 **Ardelean, Vlad:** Detecting Outliers in Time Series

Diskussionspapiere 2011
Discussion Papers 2011

01/2011 **Klein, Ingo, Fischer, Matthias and Pleier, Thomas:** Weighted Power Mean Copulas: Theory and Application

02/2011 **Kiss, David:** The Impact of Peer Ability and Heterogeneity on Student Achievement: Evidence from a Natural Experiment

03/2011 **Zibrowius, Michael:** Convergence or divergence? Immigrant wage assimilation patterns in Germany

04/2011 **Klein, Ingo and Christa, Florian:** Families of Copulas closed under the Construction of Generalized Linear Means

05/2011 **Schnitzlein, Daniel**: How important is the family? Evidence from sibling correlations in permanent earnings in the US, Germany and Denmark

06/2011 **Schnitzlein, Daniel**: How important is cultural background for the level of intergenerational mobility?

07/2011 **Steffen Mueller**: Teacher Experience and the Class Size Effect - Experimental Evidence

08/2011 **Klein, Ingo:** Van Zwet Ordering for Fechner Asymmetry

09/2011 **Tinkl, Fabian and Reichert Katja:** Dynamic copula-based Markov chains at work: Theory, testing and performance in modeling daily stock returns

10/2011 **Hirsch, Boris and Schnabel, Claus:** Let's Take Bargaining Models Seriously: The Decline in Union Power in Germany, 1992 – 2009

11/2011 **Lechmann, Daniel S.J. and Schnabel**, **Claus** : Are the self-employed really jacks-of-all-trades? Testing the assumptions and implications of Lazear's theory of entrepreneurship with German data

12/2011 **Wrede, Matthias**: Unemployment, Commuting, and Search Intensity

13/2011 **Klein, Ingo**: Van Zwet Ordering and the Ferreira-Steel Family of Skewed Distributions

Diskussionspapiere 2010
Discussion Papers 2010

01/2010 **Mosthaf, Alexander, Schnabel, Claus and Stephani, Jens:** Low-wage careers: Are there dead-end firms and dead-end jobs?

02/2010 **Schlüter, Stephan and Matt Davison:** Pricing an European Gas Storage Facility using a Continuous-Time Spot Price Model with GARCH Diffusion

03/2010 **Fischer, Matthias, Gao, Yang and Herrmann, Klaus:** Volatility Models with Innovations from New Maximum Entropy Densities at Work

04/2010 **Schlüter, Stephan and Deuschle, Carola:** Using Wavelets for Time Series Forecasting – Does it Pay Off?

05/2010 **Feicht, Robert and Stummer, Wolfgang:** Complete closed-form solution to a stochastic growth model and corresponding speed of economic recovery.

06/2010 **Hirsch, Boris and Schnabel, Claus:** Women Move Differently: Job Separations and Gender.

07/2010 **Gartner, Hermann, Schank, Thorsten and Schnabel, Claus:** Wage cyclicality under different regimes of industrial relations.

08/2010 **Tinkl, Fabian**: A note on Hadamard differentiability and differentiability in quadratic mean.

Diskussionspapiere 2009
Discussion Papers 2009

01/2009 **Addison, John T. and Claus Schnabel:** Worker Directors: A German Product that Didn't Export?

02/2009 **Uhde, André and Ulrich Heimeshoff:** Consolidation in banking and financial stability in Europe: Empirical evidence

03/2009 **Gu, Yiquan and Tobias Wenzel:** Product Variety, Price Elasticity of Demand and Fixed Cost in Spatial Models

04/2009 **Schlüter, Stephan:** A Two-Factor Model for Electricity Prices with Dynamic Volatility

05/2009 **Schlüter, Stephan and Fischer, Matthias**: A Tail Quantile Approximation Formula for the Student t and the Symmetric Generalized Hyperbolic Distribution

06/2009 **Ardelean, Vlad:** The impacts of outliers on different estimators for GARCH processes: an empirical study

07/2009 **Herrmann, Klaus:** Non-Extensitivity versus Informative Moments for Financial Models - A Unifying Framework and Empirical Results

08/2009 **Herr, Annika:** Product differentiation and welfare in a mixed duopoly with regulated prices: The case of a public and a private hospital

09/2009 **Dewenter, Ralf, Haucap, Justus and Wenzel, Tobias:** Indirect Network Effects with Two Salop Circles: The Example of the Music Industry

10/2009 **Stuehmeier, Torben and Wenzel, Tobias:** Getting Beer During Commercials: Adverse Effects of Ad-Avoidance

11/2009 **Klein, Ingo, Köck, Christian and Tinkl, Fabian:** Spatial-serial dependency in multivariate GARCH models and dynamic copulas: A simulation study

12/2009 **Schlüter, Stephan:** Constructing a Quasilinear Moving Average Using the Scaling Function

13/2009 **Blien, Uwe, Dauth, Wolfgang, Schank, Thorsten and Schnabel, Claus:** The institutional context of an "empirical law": The wage curve under different regimes of collective bargaining

14/2009 **Mosthaf, Alexander, Schank, Thorsten and Schnabel, Claus:** Low-wage employment versus unemployment: Which one provides better prospects for women?

Diskussionspapiere 2008
Discussion Papers 2008

01/2008 **Grimm, Veronika and Gregor Zoettl:** Strategic Capacity Choice under Uncertainty: The Impact of Market Structure on Investment and Welfare

02/2008 **Grimm, Veronika and Gregor Zoettl:** Production under Uncertainty: A Characterization of Welfare Enhancing and Optimal Price Caps

03/2008 **Engelmann, Dirk and Veronika Grimm:** Mechanisms for Efficient Voting with Private Information about Preferences

04/2008 **Schnabel, Claus and Joachim Wagner:** The Aging of the Unions in West Germany, 1980-2006

05/2008 **Wenzel, Tobias:** On the Incentives to Form Strategic Coalitions in ATM Markets

06/2008 **Herrmann, Klaus**: Models for Time-varying Moments Using Maximum Entropy Applied to a Generalized Measure of Volatility

07/2008 **Klein, Ingo and Michael Grottke:** On J.M. Keynes' "The Principal Averages and the Laws of Error which Lead to Them" - Refinement and Generalisation