IDIAP RESEARCH REPORT

A Probabilistic Interpretation of SVMs with an Application to Unbalanced Classification

Yves Grandvalet [1] Johnny Mariéthoz [2]
Samy Bengio [3]

IDIAP–RR 05-26

June 6, 2005

submitted for publication

[1] Heudiasyc, CNRS/UTC,60205 Compiègne cedex, France, grandval@utc.fr
[2] IDIAP Research Institute, CP 592, 1920 Martigny, Switzerland, marietho@idiap.ch
[3] IDIAP Research Institute, CP 592, 1920 Martigny, Switzerland, bengio@idiap.ch

IDIAP Research Report 05-26

A Probabilistic Interpretation of SVMs with an Application to Unbalanced Classification

Yves Grandvalet Johnny Mariéthoz Samy Bengio

June 6, 2005

submitted for publication

Abstract. In this paper, we show that the hinge loss can be interpreted as the neg-log-likelihood of a semi-parametric model of posterior probabilities. From this point of view, SVMs represent the parametric component of a semi-parametric model fitted by a maximum a posteriori estimation procedure. This connection enables to derive a mapping from SVM scores to estimated posterior probabilities. Unlike previous proposals, the suggested mapping is interval-valued, providing a set of posterior probabilities compatible with each SVM score. This framework offers a new way to adapt the SVM optimization problem when decisions result in unequal losses. Experiments on an unbalanced classification loss show improvements over state-of-the-art procedures.

Contents

1 Introduction

In this paper, we show that support vector machines (SVMs) are the solution of a relaxed maximum a posteriori (MAP) estimation problem. This relaxed problem results from fitting a semi-parametric model [7] of posterior probabilities. This model is decomposed into two components: the parametric component, which is a function of the SVM score, and the non-parametric component which we call a nuisance function. Given a proper binding of the nuisance function adapted to the considered problem, this decomposition enables to concentrate on selected ranges of the probability spectrum. The estimation process can thus allocate model capacity to the neighborhoods of decision boundaries.

The connection to semi-parametric models provides a probabilistic interpretation of SVM scores, which may have several applications, such as estimating confidences over the predictions, or dealing with unbalanced losses (which occur in domains such as diagnosis, intruder detection, etc). Several mappings relating SVM scores to probabilities have already been proposed [9, 8], but they are subject to arbitrary choices, which are avoided here by their integration to the nuisance function.

The paper is organized as follows. Section 2 presents the semi-parametric modeling approach. Section 3 shows how we can reformulate SVM within this framework. Section 4 proposes several outcomes of this formulation, including a new method to handle unbalanced losses, which is then tested empirically in Section 5. Finally, Section 6 briefly concludes the paper.

2 Semi-Parametric Classification

We address the binary classification problem of estimating a decision rule from a learning set $\mathcal{L}_n = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$, where the ith example is described by the pattern $\mathbf{x}_i \in \mathcal{X}$ and the associated response $y_i \in \{-1, 1\}$. In the framework of maximum likelihood estimation, classification can be addressed either via generative models, *i.e.* models of the joint distribution $P(X, Y)$, or via discriminative methods modeling the conditional $P(Y|X)$.

2.1 Complete and Marginal Likelihood, Nuisance Functions

Let $p(1|\mathbf{x};\boldsymbol{\theta})$ denote the model of $P(Y = 1|X = \mathbf{x})$, $p(\mathbf{x};\boldsymbol{\psi})$ the model of $P(X)$ and t_i the binary response variable such that $t_i = 1$ when $y_i = 1$ and $t_i = 0$ when $y_i = -1$. Assuming independent examples, the complete log-likelihood can be decomposed as

$$L(\boldsymbol{\theta}, \boldsymbol{\psi}; \mathcal{L}_n) = \sum_{i=1}^{n} t_i \log(p(1|\mathbf{x}_i;\boldsymbol{\theta})) + (1 - t_i) \log(1 - p(1|\mathbf{x}_i;\boldsymbol{\theta})) + \log(p(\mathbf{x}_i;\boldsymbol{\psi})) \ , \tag{1}$$

where the two first terms of the right-hand side represent the marginal or conditional likelihood, that is, the likelihood of $p(1|\mathbf{x};\boldsymbol{\theta})$.

For classification purposes, the parameter $\boldsymbol{\psi}$ is not relevant, and may thus be qualified as a nuisance parameter [4]. When $\boldsymbol{\theta}$ can be estimated independently of $\boldsymbol{\psi}$, maximizing the marginal likelihood provides the estimate returned by maximizing the complete likelihood with respect to $\boldsymbol{\theta}$ and $\boldsymbol{\psi}$. In particular, when no assumption whatsoever is made on $P(X)$, maximizing the conditional likelihood amounts to maximize the joint likelihood [5]. The density of inputs is then considered as a nuisance function.

2.2 Semi-Parametric Models

Again, for classification purposes, estimating $P(Y|X)$ may be considered as too demanding. Indeed, taking a decision only requires the knowledge of $\operatorname{sign}(2P(Y = 1|X = \mathbf{x}) - 1)$. We may thus con-

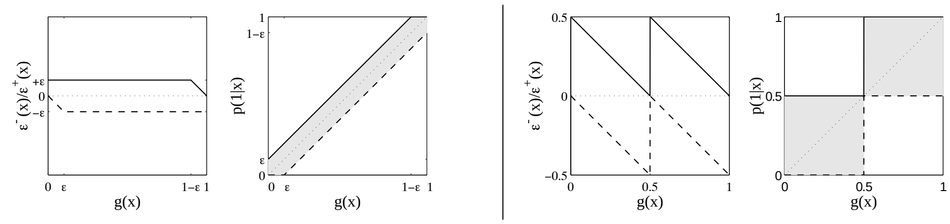

Figure 1: Two examples of $\varepsilon^-(\mathbf{x})$ (dashed) and $\varepsilon^+(\mathbf{x})$ (plain) *vs.* $g(\mathbf{x})$ and resulting ϵ-tube of possible values for the estimate of $P(Y=1|X=\mathbf{x})$ (gray zone) *vs.* $g(\mathbf{x})$.

sider looking for the decision rule minimizing the empirical classification error, but this problem is intractable for non-trivial models of discriminant functions.

Here, we briefly explore how semi-parametric models [7] may be used to reduce the modelization effort as compared to the standard likelihood approach. For this, we consider a two-component semi-parametric model of $P(Y=1|X=\mathbf{x})$, defined as $p(1|\mathbf{x};\boldsymbol{\theta}) = g(\mathbf{x};\boldsymbol{\theta}) + \varepsilon(\mathbf{x})$, where the parametric component $g(\mathbf{x};\boldsymbol{\theta})$ is the function of interest, and where the non-parametric component ε is a constrained nuisance function. Then, we address the maximum likelihood estimation of the semi-parametric model $p(1|\mathbf{x};\boldsymbol{\theta})$

$$\left\{ \begin{array}{ll} \displaystyle\min_{\boldsymbol{\theta},\varepsilon} & -\displaystyle\sum_{i=1}^{n} t_i \log(p(1|\mathbf{x}_i;\boldsymbol{\theta})) + (1-t_i)\log(1-p(1|\mathbf{x}_i;\boldsymbol{\theta})) \\ \text{s. t.} & p(1|\mathbf{x};\boldsymbol{\theta}) = g(\mathbf{x};\boldsymbol{\theta}) + \varepsilon(\mathbf{x}) \\ & 0 \leq p(1|\mathbf{x};\boldsymbol{\theta}) \leq 1 \\ & \varepsilon^-(\mathbf{x}) \leq \varepsilon(\mathbf{x}) \leq \varepsilon^+(\mathbf{x}) \end{array} \right. \tag{2}$$

where ε^- and ε^+ are user-defined functions, which place constraints on the non-parametric component ε. According to these constraints, one pursues different objectives, which can be interpreted as either weakened or focused versions of the original problem of estimating precisely $P(Y|X)$ on the whole range $[0,1]$.

At the one extreme, when $\varepsilon^- = \varepsilon^+$, one recovers a parametric maximum likelihood problem, where the estimate of posterior probabilities $p(1|\mathbf{x};\boldsymbol{\theta})$ is simply $g(\mathbf{x};\boldsymbol{\theta})$ shifted by the baseline function ε. At the other extreme, when $\varepsilon^-(\mathbf{x}) \leq -g(\mathbf{x})$ and $\varepsilon^+(\mathbf{x}) \geq 1 - g(\mathbf{x})$, $p(1|\cdot;\boldsymbol{\theta})$ perfectly explains (interpolates) any training sample for any $\boldsymbol{\theta}$, and the optimization problem in $\boldsymbol{\theta}$ is ill-posed. Note that the optimization problem in ε is always ill-posed, but this is not of concern as we do not wish to estimate the nuisance function.

Generally, as ε is not estimated, the estimate of posterior probabilities $p(1|\mathbf{x};\boldsymbol{\theta})$ is only known to lie within the interval $[g(\mathbf{x};\boldsymbol{\theta}) + \varepsilon^-(\mathbf{x}), g(\mathbf{x};\boldsymbol{\theta}) + \varepsilon^+(\mathbf{x})]$. In what follows, we only consider functions ε^- and ε^+ expressed as functions of the argument $g(\mathbf{x})$, for which the interval can be recovered from $g(\mathbf{x})$ alone. We also require $\varepsilon^-(\mathbf{x}) \leq 0 \leq \varepsilon^+(\mathbf{x})$, in order to ensure that $g(\mathbf{x};\boldsymbol{\theta})$ is an admissible value of $p(1|\mathbf{x};\boldsymbol{\theta})$.

Two simple examples are displayed in Figure 1. The two first graphs represent ε^- and ε^+ designed to estimate posterior probabilities up to precision ϵ, and the corresponding ϵ-tube of admissible estimates knowing $g(\mathbf{x})$. The two last graphs represent the same functions for ε^- and ε^+ defined to focus on the only relevant piece of information regarding decision: estimating where $P(Y|X)$ is above 1/2. [1]

[1] Of course, this naive attempt to minimize the training classification error is doomed to failure. As shown in the following section, this reformulation of the classification error does not affect its convexity, hence the optimization problem typically remains NP-hard.

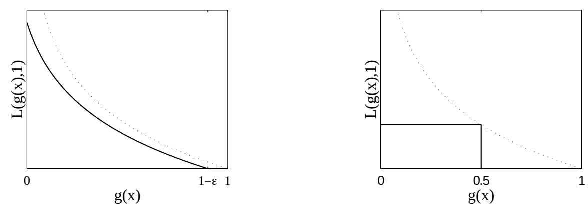

Figure 2: Losses for positive examples (plain) and neg-log-likelihood of $g(\mathbf{x})$ (dotted) *vs.* $g(\mathbf{x})$. Left: for the function ε^+ displayed on the left-hand side of Figure 1; right: for the function ε^+ displayed on the right-hand side of Figure 1.

2.3 Estimation of the Parametric Component

The definitions of ε^- and ε^+ affect the estimation of the parametric component. Regarding $\boldsymbol{\theta}$, when the values of $g(\mathbf{x};\boldsymbol{\theta}) + \varepsilon^-(\mathbf{x})$ and $g(\mathbf{x};\boldsymbol{\theta}) + \varepsilon^+(\mathbf{x})$ lie within $[0, 1]$, the semi-parametric estimation problem (2) is equivalent to the following relaxed maximum likelihood problem

$$\begin{cases} \min\limits_{\boldsymbol{\theta},\boldsymbol{\varepsilon}} & -\sum\limits_{i=1}^{n} t_i \log(g(\mathbf{x}_i;\boldsymbol{\theta}) + \varepsilon_i) + (1 - t_i) \log(1 - g(\mathbf{x}_i;\boldsymbol{\theta}) - \varepsilon_i) \\ \text{s. t.} & \varepsilon^-(\mathbf{x}_i) \leq \varepsilon_i \leq \varepsilon^+(\mathbf{x}_i) \quad i = 1, \ldots, n \end{cases} \tag{3}$$

where $\boldsymbol{\varepsilon}$ is an n-dimensional vector of slack variables. The problem is qualified as relaxed compared to the the maximum likelihood estimation of posterior probabilities by $g(\mathbf{x}_i;\boldsymbol{\theta})$, because modeling posterior probabilities by $g(\mathbf{x}_i;\boldsymbol{\theta}) + \varepsilon_i$ is a looser objective.

The monotonicity of the objective function with respect to ε_i implies that the constraints $\varepsilon^-(\mathbf{x}_i) \leq \varepsilon_i$ and $\varepsilon_i \leq \varepsilon^+(\mathbf{x}_i)$ are saturated at the solution of (3) for $t_i = 0$ or $t_i = 1$ respectively. Thus, the loss in (3) is the neg-log-likelihood of the lower or the upper bound on $p(1|\mathbf{x}_i;\boldsymbol{\theta})$ respectively, and the optimization problem with respect to $\boldsymbol{\theta}$ reduces to

$$\min_{\boldsymbol{\theta}} - \sum_{i=1}^{n} t_i \log(g(\mathbf{x}_i;\boldsymbol{\theta}) + \varepsilon^+(\mathbf{x}_i)) + (1 - t_i) \log(1 - g(\mathbf{x}_i;\boldsymbol{\theta}) - \varepsilon^-(\mathbf{x}_i)) \ , \tag{4}$$

where we assumed that g, ε^- and ε^+ are defined such that $\varepsilon^-(\mathbf{x}) \leq \varepsilon^+(\mathbf{x})$, $0 \leq g(\mathbf{x}) + \varepsilon^-(\mathbf{x}) \leq 1$ and $0 \leq g(\mathbf{x}) + \varepsilon^+(\mathbf{x}) \leq 1$.

Figure 2 displays the losses for positive examples corresponding to the choices of ε^- and ε^+ depicted in Figure 1 (the losses are symmetrical around 0.5 for negative examples). Note that the convexity of the objective function with respect to g depends on the choices of ε^- and ε^+. One can show that, providing ε^+ and ε^- are respectively concave and convex functions of g, then the loss (4) is convex in g.

When $\varepsilon^-(\mathbf{x}) \leq 0 \leq \varepsilon^+(\mathbf{x})$, $g(\mathbf{x})$ is an admissible estimate of $P(Y = 1|\mathbf{x})$. However, the relaxed loss (4) is optimistic, below the neg-log-likelihood of g. This optimism usually results in a non-consistent estimation of posterior probabilities, a common situation in semi-parametric modeling [4].[2] This lack of consistency should not be a concern here, since the non-parametric component is purposely introduced to address a looser estimation problem. We should therefore restrict consistency requirements to the primary goal of having posterior probabilities in the ϵ-tube $[g(\mathbf{x}) + \varepsilon^-(\mathbf{x}), g(\mathbf{x}) + \varepsilon^+(\mathbf{x})]$.

3 Semi-Parametric Formulation of SVMs

Several authors pointed the close relationship between the SVM criterion and the MAP approach to Gaussian processes (see [9] and references therein). However, this similarity is not perfect, prevent-

[2] That is, $g(\mathbf{x})$ does not converge towards $P(Y = 1|X = \mathbf{x})$ as the sample size goes to infinity.

ing the unambiguous interpretation of SVM scores as posterior probabilities. Here, we resolve this difficulty thanks to the additional degrees of freedom provided by semi-parametric modelling.

3.1 SVMs and Gaussian Processes

In its primal Lagrangian formulation, the SVM optimization problem reads

$$\min_{f,b} \frac{1}{2}\|f\|^2_{\mathcal{H}} + C\sum_{i=1}^{n}[1-y_i(f(\mathbf{x}_i)+b)]_+ \quad , \tag{5}$$

where $\mathcal{H}$ is a reproducing kernel Hilbert space with norm $\|\cdot\|_{\mathcal{H}}$, C is a regularization parameter and $[f]_+ = \max(f,0)$.

The penalization term in (5) can be interpreted as a Gaussian prior on f, with a covariance function proportional to the reproducing kernel of $\mathcal{H}$ [9]. Then, the interpretation of the hinge loss as a marginal log-likelihood requires to identify an affine function of the last term of (5) with the two first terms of (1). We thus look for two constants c_0 and $c_1 \neq 0$, such that, for all values of $f(\mathbf{x})+b$, there exists a value $0 \leq p(1|\mathbf{x}) \leq 1$ such that

$$\begin{cases} p(1|\mathbf{x}) &= \exp -(c_0 + c_1[1-(f(\mathbf{x})+b)]_+) \\ 1-p(1|\mathbf{x}) &= \exp -(c_0 + c_1[1+(f(\mathbf{x})+b)]_+) \end{cases} \quad . \tag{6}$$

The system (6) has a solution over the whole range of possible values of $f(\mathbf{x})+b$ if and only if $c_0 = \log(2)$ and $c_1 = 0$. Thus, the SVM optimization problem does not implement the MAP approach to Gaussian processes.

To proceed with a probabilistic interpretation of SVMs, one should thus depart from standard Gaussian processes. [9] proposed a normalized probability model, but the normalization functional was chosen arbitrarily, and the consequences of this choice on the probabilistic interpretation was not evaluated. In what follows, we derive an imprecise mapping, with interval-valued estimates of probabilities, representing the set of all admissible semi-parametric formulations of SVM scores.

3.2 SVMs and Semi-Parametric Models

With the semi-parametric models introduced in Section 2.2, one has to identify an affine function of the hinge loss with the two terms of (4). Compared to the previous situation, the identification is eased, as one has the freedom to define the slack functions ε^- and ε^+. The identification problem is now

$$\begin{cases} g(\mathbf{x})+\varepsilon^+(\mathbf{x}) = \exp -(c_0 + c_1[1-(f(\mathbf{x})+b)]_+) \\ 1-g(\mathbf{x})-\varepsilon^-(\mathbf{x}) = \exp -(c_0 + c_1[1+(f(\mathbf{x})+b)]_+) \\ \text{s.t.} \quad 0 \leq g(\mathbf{x})+\varepsilon^-(\mathbf{x}) \leq 1 \\ \qquad 0 \leq g(\mathbf{x})+\varepsilon^+(\mathbf{x}) \leq 1 \\ \qquad \varepsilon^-(\mathbf{x}) \leq \varepsilon^+(\mathbf{x}) \end{cases} \quad . \tag{7}$$

Provided $c_0 = 0$ and $0 < c_1 \leq \log(2)$, there are functions g, ε^- and ε^+ such that the above problem has a solution. Hence, semi-parametric models provide a set of probabilistic interpretations fully compatible with SVM scores. The subsets of solutions indexed by c_1 are nested, in the sense that, for any $\mathbf{x}$, the length of the uncertainty interval, $\varepsilon^+(\mathbf{x})-\varepsilon^-(\mathbf{x})$, is monotonically decreasing in c_1. In other words, the interpretation of SVM scores as posterior probabilities gets tighter as c_1 increases.

The most restricted subset of admissible interpretations, with the shortest uncertainty intervals, obtained for $c_1 = \log(2)$, is represented in the left-hand side of Figure 3. The loss incurred by a positive example is represented on the central graph, where the gray zone represents the neg-log-likelihood of all admissible solutions of $g(\mathbf{x})$. Note that the hinge loss is proportional to the neg-log-likelihood of the upper posterior probability $g(\mathbf{x})+\varepsilon^+(\mathbf{x})$, which is the loss for positive examples in the semi-parametric model in (4). Conversely, the hinge loss for negative examples is reached for $g(\mathbf{x})+\varepsilon^-(\mathbf{x})$.

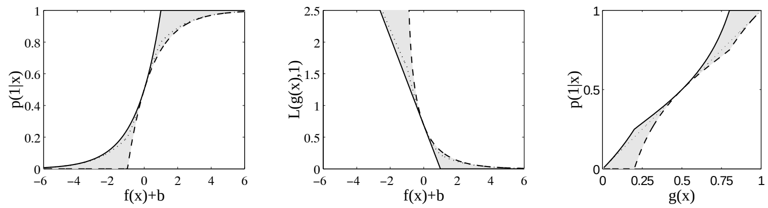

Figure 3: Left: lower (dashed) and upper (plain) posterior probabilities $[g(\mathbf{x}) + \varepsilon^-(\mathbf{x}), g(\mathbf{x}) + \varepsilon^+(\mathbf{x})]$ *vs.* SVM scores $f(\mathbf{x}) + b$; center: corresponding neg-log-likelihood of $g(\mathbf{x})$ for positive examples *vs.* $f(\mathbf{x}) + b$. right: lower (dashed) and upper (plain) posterior probabilities *vs.* $g(\mathbf{x})$, for g defined in (8).

An important observation, that will be useful in Section 4.2 is that the neg-log-likelihood of any admissible functions $g(\mathbf{x})$ is tangent to the hinge loss at $f(\mathbf{x}) + b = 0$.

The solution is unique in terms of the admissible interval $[g + \varepsilon^-, g + \varepsilon^+]$, but many definitions of $(\varepsilon^-, \varepsilon^+, g)$ solve (7). For example, g may be defined as

$$g(\mathbf{x}; \theta) = \frac{2^{-[1-(f(\mathbf{x})+b)]_+}}{2^{-[1+(f(\mathbf{x})+b)]_+} + 2^{-[1-(f(\mathbf{x})+b)]_+}} \ , \tag{8}$$

which is essentially the posterior probability model proposed by [9], represented dotted in the first two graphs of Figure 3.

The last graph of Figure 3 displays the mapping from $g(\mathbf{x})$ to admissible values of $p(1|\mathbf{x})$ which results from the choice described in (8). Although the interpretation of SVM scores does not require to specify g, it may worth to list some features common to all options. First, $g(\mathbf{x}) + \varepsilon^-(\mathbf{x}) = 0$ for all $g(\mathbf{x})$ below some threshold $g_0 > 0$, and conversely, $g(\mathbf{x}) + \varepsilon^+(\mathbf{x}) = 1$ for all $g(\mathbf{x})$ above some threshold $g_1 < 1$. These two features are responsible for the sparsity of the SVM solution. Second, the estimation of posterior probabilities is accurate at 0.5, and the length of the uncertainty interval regarding $p(1|\mathbf{x})$ monotonically increases in $[g_0, 0.5]$ and then monotonically decreases in $[0.5, g_1]$. These observations corroborate that the training objective of SVMs is intermediate between the accurate estimation of posterior probabilities on the whole range $[0, 1]$ and the minimization of the classification risk.

4 Outcomes of the Probabilistic Interpretation

This section gives two consequences of our probabilistic interpretation of SVMs. Further outcomes, still reserved for future research are listed in Section 6.

4.1 Pointwise Posterior Probabilities from SVM Scores

[8] proposed to estimate posterior probabilities from SVM scores by fitting a logistic function over the SVM scores. The only logistic function compatible with the most stringent interpretation of SVMs in the semi-parametric framework,

$$g(\mathbf{x}; \theta) = \frac{1}{1 + 4^{-(f(\mathbf{x})+b))}} \ , \tag{9}$$

is identical to the posterior probability model proposed by [9] (8) when $f(\mathbf{x}) + b$ lies in the interval $[-1, 1]$.

Other logistic functions are compatible with the looser interpretations obtained by letting $c_1 < \log(2)$, but their use as pointwise estimates becomes even more questionable, since the associated interval of admissible posterior probabilities is wider. In particular, these looser interpretations do not ensure that $f(\mathbf{x}) + b = 0$ corresponds to $g(\mathbf{x}) = 0.5$. Then, the decision function based on the estimated posterior probabilities by $g(\mathbf{x})$ may differ from the SVM decision function.

Thus, pointwise estimates of posterior probabilities derived from SVM scores should be interpreted with caution, as they require an arbitrary choice of $g(\mathbf{x})$. As discussed by [11], one should not expect accuracy from these estimates, even asymptotically, except at $f(\mathbf{x}) + b = 0$, where the estimated posterior probability converges towards 0.5.

4.2 Unbalanced Classification Losses

SVMs are known to perform well regarding misclassification error, but they also have been recognized to provide skewed decision boundaries for unbalanced classification losses, where the losses associated with incorrect decisions differ according to the true label. The mainstream approach used to address this problem consists in using different losses for positive and negative examples [6, 10], *i.e.*

$$\min_{f,b} \frac{1}{2}\|f\|^2_{\mathcal{H}} + C^+ \sum_{\{i|y_i=1\}} [1 - (f(\mathbf{x}_i) + b)]_+ + C^- \sum_{\{i|y_i=-1\}} [1 + (f(\mathbf{x}_i) + b)]_+ \quad , \tag{10}$$

where the coefficients C^+ and C^- are constants, whose ratio is equal to the ratio of the losses ℓ_{FN} and ℓ_{FP} pertaining to false negatives and false positives, respectively [3].[3] Bayes' decision theory defines the optimal decision rule by positive classification when $P(y = 1|\mathbf{x}) > P_0$, where $P_0 = \frac{\ell_{\mathrm{FP}}}{\ell_{\mathrm{FP}}+\ell_{\mathrm{FN}}}$. We may thus rewrite $C^+ = C \cdot (1 - P_0)$ and $C^- = C \cdot P_0$. With such definitions, the optimization problem may be interpreted as an upper-bound on the classification risk defined from ℓ_{FN} and ℓ_{FP}. However, the machinery of Section 3.2 unveils a major problem: the SVM decision function provided by $\mathrm{sign}(f(\mathbf{x}_i) + b)$ is not consistent with the probabilistic interpretation of SVM scores.

We address this problem by deriving another criterion, by requiring that the neg-log-likelihood of any admissible functions $g(\mathbf{x})$ is tangent to the hinge loss at $f(\mathbf{x}) + b = 0$. This leads to the following problem:

$$\begin{aligned} \min_{f,b} \frac{1}{2}\|f\|^2_{\mathcal{H}} \quad + \quad C \Bigg(& \sum_{\{i|y_i=1\}} [-\log(P_0) - (1 - P_0)(f(\mathbf{x}_i) + b)]_+ + \\ & \sum_{\{i|y_i=-1\}} [-\log(1 - P_0) + P_0(f(\mathbf{x}_i) + b)]_+ \Bigg) \quad . \end{aligned} \tag{11}$$

This loss differs from (10) whatever C^+ and C^- may be, in the respect that the margin for positive examples are smaller than the one for negative examples when $P_0 < 0.5$. In particular, (10) does not affect the SVM solution for separable problems, while in (11), the decision boundary moves towards positive support vectors when P_0 decreases. Note that The analogue of Figure 3, displayed on Figure 4, shows that one recovers the characteristics of the standard SVM loss, except that the focus is now on the posterior probability P_0 defined by Bayes' decision rule.

5 Experiments with Unbalanced Classifications Losses

It is straightforward to modify standard SVM packages to implement (11). For experimenting with difficult unbalanced two-class problems, we used a subset of the Forest database[4] which is currently the largest available UCI dataset. We consider the subproblem of discriminating the positive class Krummholz (20510 examples, originally labeled 7) against the negative class Spruce/Fir (211840 examples, originally labeled 1). The ratio of negative/positive examples is higher than 10, a feature commonly encountered with unbalanced classification losses.

[3] False negatives/positives respectively designate positive/negative examples incorrectly classified.
[4] Available at `http://kdd.ics.uci.edu/databases/covertype/covertype.data.htm`.

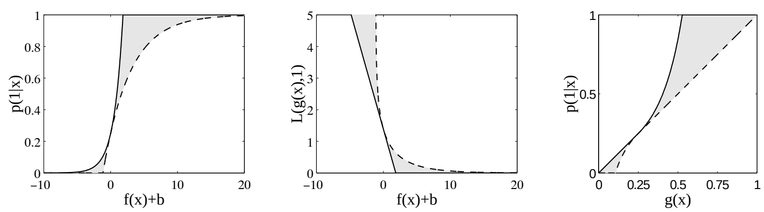

Figure 4: Left: lower (dashed) and upper (plain) posterior probabilities $[g(\mathbf{x})+\varepsilon^-(\mathbf{x}), g(\mathbf{x})+\varepsilon^+(\mathbf{x})]$ *vs.* SVM scores $f(\mathbf{x})+b$ obtained from (11) with $P_0 = 0.25$; center: corresponding neg-log-likelihood of $g(\mathbf{x})$ for positive examples *vs.* $f(\mathbf{x})+b$. right: lower (dashed) and upper (plain) posterior probabilities *vs.* $g(\mathbf{x})$, for g defined by $\varepsilon^+(\mathbf{x}) = 0$ for $f(\mathbf{x})+b \leq 0$ and $\varepsilon^-(\mathbf{x}) = 0$ for $f(\mathbf{x})+b \geq 0$.

We randomly selected from the original dataset 3 000 examples for the positive class and 30 0000 for the negative class. Each set of examples was further divided into three equal parts corresponding to the training, validation and test sets.

The performance is measured by the weighted risk function $R = \frac{1}{n}(N_{\mathrm{FN}}\ell_{\mathrm{FN}} + N_{\mathrm{FP}}\ell_{\mathrm{FP}})$, where N_{FN} and N_{FP} are the number of false negatives and false positives, respectively. The loss ℓ_{FP} was set to one, and ℓ_{FN} was successively set to 1, 10 and 100, in order to penalize more and more heavily errors from the under-represented class.

All approaches were tested using SVMs with a Gaussian kernel. All hyper-parameters were tuned on the validation set for each of the ℓ_{FN} values. The bias b was also tuned on the validation set, to improve the results for the baseline and C^+/C^- optimizers, which do not estimate this parameter correctly for $\ell_{\mathrm{FN}} \neq \ell_{\mathrm{FP}}$. Table 1 compares the risk R obtained with the three approaches for the different values of ℓ_{FN}.

Table 1: Errors for 3 different criteria and for 3 different models over the Forest database

ℓ_{FN}	Baseline, problem (5)	C^+/C^-, problem (10)	P_0, problem (11)
1	2.52	2.52	2.52
10	11.69	10.73	10.91
100	45.23	36.81	30.67

The first line, with $\ell_{\mathrm{FN}} = 1$ corresponds to the standard classification error, where all criteria are equivalent in theory and in practice. For $\ell_{\mathrm{FN}} = 10$, the models obtained by optimizing C^+/C^- (10) and P_0 (11) give similar results, better than the baseline. Finally, for the highly unbalanced loss, the novel approach provides statistically significantly better results (where significance was tested at the 5% level, using a bootstrap-based statistical test similar to [1]). The new optimization criterion can thus outperform standard approaches for highly unbalanced problems.

6 Conclusion

This paper introduced a semi-parametric model for classification which provides an interesting viewpoint on SVMs. The non-parametric component provides an intuitive means of transforming the likelihood into a decision-oriented criterion. This framework was used here to propose a new parameterization of the hinge loss, dedicated to unbalanced classification problems, yielding significant improvements over the classical procedure.

Among other prospectives, we plan to apply the same framework to investigate hinge-like criteria for decision rules including a reject option, where the classifier abstains when a pattern is ambiguous.

We also aim at defining losses encouraging sparsity in probabilistic models, such as kernelized logistic regression. We could thus build sparse probabilistic classifiers, providing an accurate estimation of posterior probabilities on a (limited) predefined range of posterior probabilities. In particular, we could derive decision-oriented criteria for multi-class probabilistic classifiers. For example, minimizing classification error only requires to find the class with highest posterior probability, and this search does not require precise estimates of probabilities outside the interval $[1/K, 1/2]$, where K is the number of classes.

7 Acknowledgments

This research has been partially carried out in the framework of the Swiss NCCR project (IM)2. It was also supported in part by the IST Program of the European Community, under the PASCAL Network of Excellence, IST-2002-506778, funded in part by the Swiss OFES. All experiments were performed using the *Torch* package [2].

References

[1] R. M. Bolle, N. K. Ratha, and S. Pankanti. Error analysis of pattern recognition systems - the subsets bootstrap. *Computer Vision and Image Understanding*, 93:1–33, 2004.

[2] R. Collobert, S. Bengio, and J. Mariéthoz. Torch: a modular machine learning software library. Technical Report IDIAP-RR 02-46, IDIAP, 2002.

[3] Y. Lin, Y. Lee, and G. Wahba. Support vector machines for classification in non-standard situations. *Machine Learning*, 46:191–202, 2002.

[4] B. G. Lindsay. Nuisance parameters. In S. Kotz, C. B. Read, and D. L. Banks, editors, *Encyclopedia of Statistical Sciences*, volume 6. Wiley, 1985.

[5] G. J. McLachlan. *Discriminant analysis and statistical pattern recognition.* Wiley, 1992.

[6] K. Morik, P. Brockhausen, and T. Joachims. Combining statistical learning with a knowledge-based approach - a case study in intensive care monitoring. In *Proceedings of the Sixteenth International Conference on Machine Learning (ICML-99)*, 1999.

[7] D. Oakes. Semi-parametric models. In S. Kotz, C. B. Read, and D. L. Banks, editors, *Encyclopedia of Statistical Sciences*, volume 8. Wiley, 1988.

[8] J. C. Platt. Probabilities for SV machines. In A. J. Smola, P. L. Bartlett, B. Schölkopf, and D. Schuurmans, editors, *Advances in Large Margin Classifiers*, pages 61–74. MIT Press, 2000.

[9] P. Sollich. Probabilistic methods for support vector machines. In S. A. Solla, T. K. Leen, and K.-R. Müller, editors, *Advances in Neural Information Processing Systems 12*, pages 349–355. MIT Press, 2000.

[10] K. Veropoulos, C. Campbell, and N. Cristianini. Controlling the sensitivity of support vector machines. In T. Dean, editor, *Proceedings of the International Joint Conference on Artificial Intelligence*, pages 55–60. Morgan Kaufmann, 1999.

[11] T. Zhang. Statistical behavior and consistency of classification methods based on convex risk minimization. *Annals of Statistics*, 32(1):56–85, 2004.