On the Relationship between Classification Error Bounds and Training Criteria in Statistical Pattern Recognition

Hermann Ney

Lehrstuhl für Informatik VI – Computer Science Department
RWTH Aachen – University of Technology, 52056 Aachen, Germany
ney@informatik.rwth-aachen.de

Abstract. We present two novel bounds for the classification error that, at the same time, can be used as practical training criteria. Unlike the bounds reported in the literature so far, these novel bounds are based on a strict distinction between the true but unknown distribution and the model distribution, which is used in the decision rule. The two bounds we derive are the squared distance and the Kullback-Leibler distance, where in both cases the distance is computed between the true distribution and the model distribution. In terms of practical training criteria, these bounds result in the squared error criterion and the mutual information (or equivocation) criterion, respectively.

1 Introduction

The classification error is the most important performance criterion in any pattern recognition task. The goal of this work to establish a direct relationship between practical training criteria and exact upper bounds for the classification error. There are three novel contributions of this paper:

- All the considerations will be based on the model-based classification error as opposed to the Bayes classification error. The Bayes error is only of theoretical value, because it requires the true but unknown distribution. Instead, we will use the model distribution in the decision rule whose parameters have to be learned from training data.
- Since the classification error is difficult to handle, we will derive two upper bounds that are more convenient for mathematical analysis.
- Using these bounds, we derive two practical training criteria which are well known in pattern recognition and show that they are related to upper bounds of the model-based classification error.

The concept of using the classification error directly as training criterion is widely known in pattern recognition [7, pp.46/47], [9, pp.106/107]. However, these studies always use the Bayes classification error. In addition, upper bounds are reported, but again only for the Bayes classification error [7, pp.46/47], [10]. In [6], the model-based classification error is studied, but only for two-class

F.J. Perales et al. (Eds.): IbPRIA 2003, LNCS 2652, pp. 636–645, 2003.

problems. Vapnik's framework of empirical risk minimization [6, pp. 187],[18] is more concerned with statistical fluctuations from one sample set to another sample set, and the reference error rate is *not* the Bayes classification error. To the best of our knowledge, the exact mathematical dependence between the model-based classification error and the possible training criteria has not been studied before.

2 Model-Based Decision Rule and Classification Error

2.1 Classification Task and True Distribution

In statistical pattern recognition, we consider the observation (or feature) vector $x \in \mathcal{X} \subset \mathrm{I\!R}^D$ and the class index $c = 1, ..., C$ to be random variables with a joint distribution:

$$\begin{aligned} \text{pair of random variables:} \quad & (x, c) \\ \text{with true distribution:} \quad & pr(x,c) \ = \ pr(x)\, pr(c|x) \end{aligned} \tag{1}$$

The classification task is to determine for each observation vector x the associated class index c. For such a task, the minimum classification error is obtained for the Bayes decision rule in which the class posterior distribution $pr(c|x)$ plays a crucial role. We will refer to it simply as the true distribution.

2.2 Model Distribution and Associated Decision Rule

In all practical applications, the true distribution $pr(c|x)$ is not known, and we can use only a so-called model distribution $p_\vartheta(c|x)$ instead. For such a model distribution, the functional dependence of the class index c on the observation vector x is fully specified using some unknown parameter set ϑ. The choice of this functional dependence is very much application specific. A large number of widely used techniques in pattern recognition fit into this interpretation. Examples are artificial neural networks or any type of discriminant functions, decision tree (CART) approaches, the single Gaussian and Gaussian mixture classifiers and maximum entropy (or log-linear) models. In case of observation vectors over a time axis, Hidden Markov models are typically used.

To be more exact, the model distribution $p_\vartheta(c|x)$ is a posterior distribution over the classes $c = 1, ..., C$:

$$\begin{aligned} \text{model distribution:} \quad & p_\vartheta(c|x) \\ \text{with:} \quad & 0 \ \le \ p_\vartheta(c|x) \qquad \sum_c p_\vartheta(c|x) = 1 \end{aligned} \tag{2}$$

We interpret it as the score of the hypothesis that the observation x has been generated by the class c, and thus it is a natural requirement to normalize these scores in such a way that, for each observation x, they sum up to unity. Note that, for non-negative scores $p_\vartheta(c, x)$, we can always satisfy this constraint by simple re-normalization.

To find the unknown class identity of an observation x, we define the model-based decision rule:

$$\text{decision rule } c_\vartheta(\cdot): \qquad c_\vartheta : \mathcal{X} \rightarrow \{1, ..., C\}$$
$$x \rightarrow c_\vartheta(x) := \arg\max_c \Big\{ p_\vartheta(c|x) \Big\} \tag{3}$$

In order to avoid an awkward notation, we use only the parameter ϑ as index on the decision rule to express the dependence on the *full* model distribution $p_\vartheta(c|x)$. We use the attribute *model-based* to distinguish this decision rule from the *Bayes* decision rule where the true but unknown distribution $pr(c|x)$ is needed. In the following, the goal will be to study whether and how the classification error of the model-based decision rule will get close to the minimum classification error.

2.3 Model-Based Classification Error

When using such a decision rule $x \rightarrow c_\vartheta(x)$, we have a classification error count $e(x, c)$ for a joint event (x, c):

$$e(x, c) := 1 - \delta(c_\vartheta(x), c) \tag{4}$$

where $\delta(\cdot, \cdot)$ denotes the Kronecker delta. The *local* classification error $E_\vartheta\{e|x\}$ is the x-conditional expectation of $e(x, c)$, which is obtained by using the true class posterior distribution $pr(c|x)$ in the point x of the observation space:

$$\begin{aligned} E_\vartheta\{e|x\} &:= \sum_c pr(c|x) \cdot \Big[1 - \delta(c_\vartheta(x), c)\Big] \\ &= 1 - pr(c_\vartheta(x)|x) \end{aligned} \tag{5}$$

The *global* classification error $E_\vartheta\{e\}$ is obtained by integrating over the whole space of observations x:

$$E_\vartheta\{e\} = \int_x dx \; pr(x) \; E_\vartheta\{e|x\} \tag{6}$$

Ideally, we would like to directly minimize this classification error in order to learn the unknown parameter set ϑ. However, the direct solution to this optimization problem is extremely difficult for two reasons: First, there are two extreme nonlinearities, namely the maximum operations and the Kronecker delta. Second, we have to compute the expectation over the true distribution $pr(x)$ which however is unknown and for which only a training sample is available.

3 Bounds for Local Classification Error

In this section, we will derive bounds for the local classification error when the decision rule Eq.(3) is used with *any type* of model $p_\vartheta(c|x)$. We will start with the x-conditional, i.e. *local*, classification error and consider the global classification error later.

3.1 Principle

It is well known that the global minimum of the error rate is obtained by the Bayes decision rule:

$$x \rightarrow c_*(x) := \arg\max_c \Big\{ pr(c|x) \Big\} \tag{7}$$

i.e. when the true (but unknown) posterior distribution $pr(c|x)$ is used as model distribution $p_\vartheta(c|x)$. The associated local Bayes classification error $E_*\{e|x\}$ is:

$$E_*\{e|x\} = 1 - pr(c_*(x)|x) \tag{8}$$

Therefore, the Bayes error is also the absolute minimum of *any* model $p_\vartheta(c|x)$ (for a *fixed* type of observations x), and we will consider the *difference* between the model-based classification error $E_\vartheta(e|x)$ and the Bayes classification error $E_*\{e|x\}$. In the following, we will derive an inequality of the form:

$$E_\vartheta\{e|x\} - E_*\{e|x\} \leq \alpha \cdot \Big|\Big| pr(\cdot|x) - p_\vartheta(\cdot|x) \Big|\Big| \tag{9}$$

where we have a positive constant α and we use a suitable norm $||\cdot||$ of a C-dimensional difference vector between the true distribution $pr(c|x)$ and the model distribution $p_\vartheta(c|x)$. Depending on the type of norm $||\cdot||$, we will refer to these bounds as l_1, l_2 and l_∞ bounds.

3.2 Basic Inequality

Using the basic definitions introduced so far, we can write down the following sequence of equations and inequalities:

$$\begin{aligned}
& E_\vartheta\{e|x\} - E_*\{e|x\} \; := \\
& := \Big[1 - pr(c_\vartheta(x)|x)\Big] - \Big[1 - pr(c_*(x)|x)\Big] \\
& = pr(c_*(x)|x) - pr(c_\vartheta(x)|x) \\
& \leq pr(c_*(x)|x) - pr(c_\vartheta(x)|x) + p_\vartheta(c_\vartheta(x)|x) - p_\vartheta(c_*(x)|x) \qquad (10) \\
& = \Big[pr(c_*(x)|x) - p_\vartheta(c_*(x)|x)\Big] + \Big[p_\vartheta(c_\vartheta(x)|x) - pr(c_\vartheta(x)|x)\Big] \\
& \leq \Big|pr(c_*(x)|x) - p_\vartheta(c_*(x)|x)\Big| + \Big|pr(c_\vartheta(x)|x) - p_\vartheta(c_\vartheta(x)|x)\Big| \qquad (11)
\end{aligned}$$

Here, the first inequality Eq.(10) is true because, by the definition of the model-based decision rule $x \rightarrow c_\vartheta(x)$, we must have for any class c:

$$p_\vartheta(c|x) \leq \max_{\tilde{c}} \; \{p_\vartheta(\tilde{c}|x)\} \; \equiv \; p_\vartheta(c_\vartheta(x)|x) \tag{12}$$

The second inequality Eq.(11) results simply from the application of the triangle inequality.

3.3 Local Bounds

From the inequality Eq.(11), we immediately obtain what will be referred to as l_1 bound:

$$E_\vartheta\{e|x\} - E_*\{e|x\} \leq \sum_c |pr(c|x) - p_\vartheta(c|x)| \tag{13}$$

It is easy to verify that this bound is also correct in the special case: $c_\vartheta(x) = c_*(x)$. In addition, we can also immediately establish the l_∞ bound (or maximum bound) and the l_2 bound:

$$\begin{aligned} E_\vartheta\{e|x\} - E_*\{e|x\} &\leq 2 \cdot \max_c \Big\{ |pr(c|x) - p_\vartheta(c|x)| \Big\} && (14) \\ &\leq 2 \cdot \sqrt{\sum_c [pr(c|x) - p_\vartheta(c|x)]^2} && (15) \end{aligned}$$

We would like to emphasize that each of these three *local* bounds is tight in the following sense. When the model distribution $p_\vartheta(c|x)$ approaches the true distribution $pr(c|x)$, the bound goes to zero so that the model-based classification error $E_\vartheta\{e|x\}$ approaches the Bayes classification error $E_*\{e|x\}$.

4 Bounds for Global Classification Error

In this section, we will establish bounds for the *global* classification error that have similar properties as the bounds for the local classification error.

4.1 From Local to Global Bounds

We consider the difference between the model-based classification error $E_\vartheta\{e|x\}$ and the Bayes classification error $E_*\{e|x\}$:

$$E_\vartheta\{e|x\} - E_*\{e|x\} \leq g(x)$$

where $g(x)$ stands for one of the local bounds derived so far. We move from local to global bounds by integrating over the whole space of observations using the true probability (density) distribution $pr(x)$:

$$\begin{aligned} E_\vartheta\{e\} - E_*\{e\} &= \int dx\, pr(x) \Big(E_\vartheta\{e|x\} - E_*\{e|x\} \Big) \\ &\leq \int dx\, pr(x)\, g(x) && (16) \end{aligned}$$

In carrying out the integration, the local inequality is preserved and we obtain a global inequality. Now it turns out that, in order to arrive at useful bounds, it is helpful to consider the *squared* difference:

$$\begin{aligned} \Big(E_\vartheta\{e\} - E_*\{e\} \Big)^2 &\leq \left(\int dx\, pr(x)\, g(x) \right)^2 \\ &\leq \int dx\, pr(x)\, g^2(x) && (17) \end{aligned}$$

The second inequality is true because for any function $x \to g(x)$ we have the inequality:

$$\left(\int dx\, pr(x)\, g(x) \right)^2 \leq \int dx\, pr(x)\, g^2(x) \tag{18}$$

$$\begin{aligned} \text{since:} \qquad 0 \;\leq\; Var\{g(x)\} &:= E\{[g(x) - E\{g(x)\}]^2\} \\ &= E\{g^2(x)\} - E^2\{g(x)\} \end{aligned}$$

where $E\{\cdot\}$ denotes the statistical expectation using the distribution $pr(x)$. The ultimate justification for considering the *squared* difference in the classification error will be the usefulness of the practical training criteria to be presented in Section 5.

4.2 Squared Distance Bound

We start with the local bound Eq.(15) and immediately obtain the global bound using Eq.(17):

$$\Big(E_\vartheta\{e\} - E_*\{e\} \Big)^2 \leq 4 \cdot \int dx\, pr(x) \sum_c [pr(c|x) - p_\vartheta(c|x)]^2 \tag{19}$$

This global bound will be called squared error bound because it is based on the squared difference between the true distribution $pr(c|x)$ and the model distribution $p_\vartheta(c|x)$.

4.3 Kullback-Leibler Bound

To derive this bound, we make use of the Pinsker inequality for two probability distributions p_c and q_c (with normalization $\sum_c p_c = \sum_c q_c = 1$) [5, p. 300],[17]:

$$\frac{1}{2} \left(\sum_c |p_c - q_c| \right)^2 \leq - \sum_c p_c \log \frac{q_c}{p_c} \tag{20}$$

The term on the right-hand side of this inequality is known as the Kullback-Leibler distance (or relative entropy) between the two distributions p_c and q_c [5, p. 18]. It was originally introduced in the context of statistics and information theory *without* any link to the classification error rate. We use the Kullback-Leibler distance as a distance between the true distribution $pr(c|x)$ and the model distribution $p_\vartheta(c|x)$.

Inserting the local bound Eq.(13) into Eq.(17), we obtain the global bound:

$$\begin{aligned} \Big(E_\vartheta\{e\} - E_*\{e\} \Big)^2 &\leq \int dx\, pr(x) \left(\sum_c |pr(c|x) - p_\vartheta(c|x)| \right)^2 \\ &\leq -2 \cdot \int dx\, pr(x) \sum_c pr(c|x) \log \frac{p_\vartheta(c|x)}{pr(c|x)} \end{aligned} \tag{21}$$

Each of the two global bounds Eqs.(19) and (21) is *tight*: When the model distribution approaches the true distribution, the bound goes to zero, and so does the difference between model-based classification error and Bayes classification error.

5 Empirical Training Criteria

In this section, we will show how each of the global bounds can be used *directly* as training criterion to learn the unknown parameter set ϑ from a set of training data.

5.1 From Error Bounds to Empirical Training Criteria

The approach is based on re-writing the inequality for each of the classification error bounds in the form:

$$\Big(E_\vartheta\{e\} - E_*\{e\}\Big)^2 \le \int dx \sum_c pr(x,c)\; h_\vartheta(x,c) \tag{22}$$

with a suitable function $h_\vartheta(x,c)$. To obtain a practical training criterion, we apply two steps:

- For the classification error $E_\vartheta\{e\}$ to approach the Bayes error $E_*\{e\}$, we tighten the bound on the right-hand side by minimizing it over the unknown parameter set ϑ.
- Now, of course, the true distribution $pr(x,c)$ is not known, and we have only access to a *representative sample*, i.e. a set of labelled observations from the task for which we want to design our pattern classification system:

  $$(x_n, c_n),\;\; n = 1, ..., N$$

  i.e. observation x_n with class label c_n. Using this set of labelled training data, we define the *empirical* distribution

  $$pr(x,c) = \frac{1}{N}\sum_{n=1}^{N} \delta(x,x_n)\,\delta(c,c_n)$$

  where, for continuous-valued observations x, $\delta(x,x_n)$ is the Dirac delta function rather than the Kronecker delta.

The training criterion for determining the optimum parameter set $\hat{\vartheta}$ can now be written as:

$$\begin{aligned}\hat{\vartheta} &:= \arg\min_{\vartheta}\left\{\int dx \sum_c pr(x,c)\; h_\vartheta(x,c)\right\}\\ &= \arg\min_{\vartheta}\left\{\frac{1}{N}\sum_{n=1}^{N} h_\vartheta(x_n,c_n)\right\} \;=\; \arg\min_{\vartheta}\left\{\sum_{n=1}^{N} h_\vartheta(x_n,c_n)\right\}\end{aligned} \tag{23}$$

If, in addition to determining the optimum parameter set $\hat{\vartheta}$, we want to estimate the classification error using this method, we have to be careful and avoid too optimistic an estimate [8, p. 248]. In other words, the approach presented here does not address the problem of *overfitting*.

5.2 Squared Error Criterion

To derive the squared error criterion, we use the following identity [13]:

$$\sum_c [pr(c|x) - p_\vartheta(c|x)]^2 \; =$$

$$= \sum_c pr(c|x) \sum_{c'} [p_\vartheta(c'|x) - \delta(c', c)]^2 \; - \; \left(1 - \sum_c pr^2(c|x)\right) \qquad (24)$$

This identity has been re-discovered several times in the context of statistical pattern recognition and artificial neural networks. The earliest reference (using a different framework of notation) we know is [15]. Inserting this identity into Eq.(19) and dropping the terms independent of ϑ, we arrive at the following training criterion for the unknown parameter set ϑ:

$$\hat{\vartheta} = \arg\min_{\vartheta} \left\{ \sum_{n=1}^{N} \sum_{c'} \; [p_\vartheta(c'|x_n) - \delta(c', c_n)]^2 \right\} \qquad (25)$$

This is the standard training criterion used for neural networks and other types of discriminant functions, namely the sum of the squared differences between the actual network output and the desired output for each output node [7, p. 290]. If the model distribution is non-parametric, i.e. has enough degrees of freedom, the global optimum can be really attained (on the training data), and the model distribution is then identical to the true distribution. This is the case for decision trees [3] with a non-parametric model distribution for the discrete-valued observations x. The minimum values of the training criterion is then the second term (with a positive sign) on the right-hand side of Eq.(24), which is referred to as Gini criterion.

5.3 Kullback-Leibler Criterion

From the Kullback-Leibler bound, we obtain the practical training criterion by simply separating the model distribution $p_\vartheta(c|x)$ and dropping the constant terms:

$$\hat{\vartheta} = \arg\max_{\vartheta} \left\{ \sum_{n=1}^{N} \; \log \, p_\vartheta(c_n|x_n) \right\} \qquad (26)$$

This is the general form of a maximum likelihood criterion. Here, we have the likelihood of the class posterior distribution $p_\vartheta(c|x)$ as opposed to the class conditional distribution $p_\vartheta(x|c)$. This criterion has become popular in the context of so-called discriminative training and is referred to under different names: *conditional maximum likelihood* [4, 12] and *maximum mutual information* [1, 14]. In the framework of information theory, the training criterion can be interpreted as the empirical expectation of the model-based equivocation, which, in the special case of constant class probabilities, is equivalent to mutual information. In the context of decision trees [3], the criterion is called *entropy* criterion.

6 Discussion

We have derived two novel bounds for the model-based classification error: the squared distance bound and the Kullback-Leibler bound, both of which result in widely used practical training criteria. Although both these quantities have been used before in statistical pattern recognition, they were not known to provide *strict bounds* for the model-based classification error.

It is interesting to note that, in a Bayesian framework independent of the classification error, some authors [2, pp.67-81] have analyzed criteria for estimating unknown probability distributions and have considered two specific criteria that have attractive properties. These two criteria are identical to the two training criteria that we have derived here. They are referred to as the *quadratic* and the *logarithmic* scoring function.

The bounds we have presented are based on the *square* of the difference between the model-based classification error and the Bayes classification error. The open question is how this is related to approaches where the smoothed classification error is used directly as training criterion [11, 16].

Acknowledgment

Part of this work was done when the author was a visiting scientist at the UP of Valencia in 2002. In particular, the author would like to thank Francisco Casacuberta, Alfons Juan and Enrique Vidal for many helpful discussions.

References

[1] L. R. Bahl, P. F. Brown, P. V. de Souza, R. L. Mercer: Maximum Mutual Information Estimation of Hidden Markov Model Parameters. IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Tokyo, April, 1986. 644

[2] J. M. Bernardo, A. F. M. Smith: Bayesian Theory. J. Wiley & Sons, Chichester 1994. 644

[3] L. Breiman, J. H. Friedman, R. A. Ohlsen, C. J. Stone: Classification And Regression Trees. Wadsworth, Belmont, CA, 1984. 643, 644

[4] F. Casacuberta: Maximum Mutual Information and Conditional Maximum Likelihood Estimation of Stochastic Regular Syntax-Directed Translation Schemes. Int. Coll. on Grammatical Inference, Montpellier, France, Sep. 1996, pp. 282-291 in L. Miclet, C. de la Higuera (eds.), Lecture Notes in Computer Science, Springer, Berlin 1996. 644

[5] T. M. Cover, J. A. Thomas: Elements of Information Theory. John Wiley & Sons, New York, NY, 1991. 641

[6] L. Devroye, J. Györfi, G. Lugosi: A Probabilistic Theory of Pattern Recognition. Springer, New York, 1996. 636, 637

[7] R. O. Duda, P. E. Hart, D. G. Stork: Pattern Classification. 2nd ed., J. Wiley & Sons, New York, NY, 2001. 636, 643

[8] B. Efron, R. J. Tibshirani: An Introduction to the Bootstrap. Chapman & Hall, New York, 1993. 643

[9] K. Fukunaga: Introduction to Statistical Pattern Recognition. Academic Press, New York, 1972. 636

[10] M. E. Hellman, J. Raviv: Probability of Errors, Equivocation and the Chernoff Bound. IEEE Trans. on Information Theory, Vol. IT-16, No. 4, pp. 368-372, July 1970. 636

[11] B.-H. Juang, S. Katagiri: Discriminative Learning for Minimum Error Classification. IEEE Transactions on Signal Processing, Vol. 40, No. 12, pp. 3043-3054, Dec. 1992. 644

[12] A. Nadas, D. Nahamoo, M. Picheny: On a Model-Robust Training Method for Speech Recognition. IEEE Trans. on Acoustics, Speech and Signal Processing, Vol. 36, No. 9, pp. 1432-1435, Sep. 1988. 644

[13] H. Ney: On the Probabilistic Interpretation of Neural Net Classifiers and Discriminative Training Criteria. IEEE Trans. on Pattern Analysis and Machine Intelligence, Vol. PAMI-17, No. 2, pp. 107-119, Feb. 1995. 643

[14] Y. Normandin, R. Cardin, R. De Mori: High-Performance Connected Digit Recognition Using Maximum Mutual Information Estimation. IEEE Trans. on Speech and Audio Processing, vol. 2, no. 2, pp. 299-311, April 1994. 644

[15] J. D. Patterson, B. F. Womack: An Adaptive Pattern Classification Scheme. IEEE Trans. on Systems, Science and Cybernetics, Vol.SSC-2, pp.62-67, Aug. 1966. 643

[16] R. Schlüter, H. Ney: Model-based MCE Bound to the True Bayes' Error. IEEE Signal Processing Letters, Vol. 8, No. 5, pp. 131-133, May 2001. 644

[17] F. Topsoe: Some Inequalities for Information Divergence and Related Measures of Discrimination. IEEE Trans. on Information Theory, to appear, 2003. 641

[18] A. N. Vapnik, V. Y. Chervonenkis: On the Uniform Convergence of Relative Frequencies of Events to their Probabilities. Theory of Probability and Its Applications, Vol. 16, pp. 264-280, 1971. 637