ELSEVIER

January 1995

Pattern Recognition Letters 16 (1995) 105–111

Pattern Recognition Letters

On the exponential value of labeled samples ☆

Vittorio Castelli, Thomas M. Cover *

Information Systems Laboratory, Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA

Received 14 July 1994; revised 8 August 1994

Abstract

Consider the problem of classifying a sample X_0 into one of two classes, using a training set Q. Let Q be composed of l labeled samples $\{(X_1, \theta_1), ..., (X_l, \theta_l)\}$ and u unlabeled samples $\{X'_1, ..., X'_u\}$, where the labels θ_i are i.i.d. Bernoulli(η) random variables over the set $\{1, 2\}$, the observations $\{X_i\}_{i=1}^{l}$ are distributed according to $f_{\theta_i}(\cdot)$ and the unlabeled observations $\{X'_j\}_{j=1}^{u}$ are independently distributed according to the mixture density $f_{X'}(\cdot) = \eta f_1(\cdot) + (1-\eta) f_2(\cdot)$. We assume that $f_1(\cdot)$, $f_2(\cdot)$ and η are all unknown. Let $f_1(\cdot)$ and $f_2(\cdot)$ belong to a known family $\mathcal{F}$, and assume that the mixtures of elements of $\mathcal{F}$ are identifiable. Even when the number of unlabeled samples is infinite and the decision regions can therefore be identified, one still needs labeled samples to label the decision regions with the correct classification. Letting $R(l, u)$ denote the optimal probability of error for l labeled and u unlabeled samples, and assuming that the pairwise mixtures of $\mathcal{F}$ are identifiable, we obtain the obvious statements $R(0, u) = R(0, \infty) = 1/2$, $R(1, 0) \leq 2\eta\bar{\eta}$, $R(\infty, u) = R^*$, and then prove $R(1, \infty) = 2R^*(1-R^*)$, where R^* is the Bayes probability of error, and $R(l, \infty) = R^* + \exp\{-\alpha l + o(l)\}$, where the exponent α is given by $-\log(2\sqrt{\eta\bar{\eta}} \int \sqrt{f_1(x) f_2(x)}\, dx)$. Thus the first labeled sample reduces the risk from $1/2$ to $2R^*(1-R^*)$ and subsequent labeled samples in the training set reduce the probability of error exponentially fast to the Bayes risk.

1. Introduction

Machine learning is traditionally divided in two branches: supervised learning and unsupervised learning. While unsupervised learning is usually associated with clustering, the main stream of work in pattern recognition focuses on supervised learning, the construction of classification rules based on labeled examples (Jain and Chandrasekaran, 1992; Raudys and Jain, 1991; Vapnik, 1982; Devroye, 1988). In many practical cases, though, labeled and unlabeled samples may be available at the same time. Some papers treat both labeled and unlabeled observations in the construction of a classification scheme (Tolat and Peterson, 1990; Pao and Sobajic, 1992; Kester, 1985; Greenspan et al., 1991; McLachlan, 1977; McLachlan and Ganesalingam, 1982; O'Neill, 1978; Shahshahani and Landgrebe, 1992).

A problem of general interest is to determine the interplay of the number of labeled samples l and unlabeled samples u in the resulting probability of error $R(l, u)$, with the ultimate goal of finding the relative value of labeled and unlabeled samples. This paper evaluates $R(l, u)$ at the boundaries of the (l, u) region. A following

☆ This work was partially supported by NSF Grant NCR-8914538-02, ARPA contract J-FBI-91-218 and JSEP contract DAAHO4-94-G-0058. Vittorio Castelli was a Rotary Scholar during the academic year 93-94. Some of these results were presented in talks at COLT, Santa Cruz, California, August, 1991, and at the IEEE International Symposium on Information Theory, San Antonio, Texas, February, 1993.

* Corresponding author. Email: cover@isl.stanford.edu

SSDI 0167-8655(94)00074-3

paper will treat the determination of $R(l, u)$ when l and u are in the interior. We show here that labeled samples are exponentially valuable in reducing risk. Roughly speaking, it turns out that unlabeled samples are only polynomially valuable, thereby setting up an anticipated result that labeled samples are exponentially more valuable than unlabeled samples in pattern recognition.

We assume the following framework. Let all the random variables be defined on a probability space (Ω, Ξ, P). The notation "Pr" will be used when we omit measure-theoretical details for sake of clarity. Let the training set Q be composed of l labeled samples $\{(X_1, \theta_1), ..., (X_l, \theta_l)\}$ and ∞ unlabeled samples $\{X'_1, X'_2, ...\}$. Let the class labels $\{\theta_i\}_{i=1}^{l}$ be independent and identically distributed (i.i.d.) random variables with $P\{\theta_i=1\}=\eta$, $P\{\theta_i=2\}=1-\eta\triangleq\bar{\eta}$, and let each observation X_i be independently distributed according to $f_{\theta_i}(x)$. Define $\bar{\theta}_i=1$ if $\theta_i=2$, $\bar{\theta}_i=2$ if $\theta_i=1$. then the labeled samples are distributed according to

$$f(x, \theta)=[\eta f_1(x)]^{\mathbf{1}(\theta=1)}[\bar{\eta} f_2(x)]^{\mathbf{1}(\theta=2)}, \tag{1}$$

where the indicator function $\mathbf{1}(\theta=1)$ is equal to 1 if $\theta=1$, and to 0 otherwise. The unlabeled samples $\{X'_j\}_{i=1}^{\infty}$ appear to be i.i.d. random variables distributed according to the mixture density $\sum_{\theta} f(x, \theta)$ given by

$$f_{X'}(\cdot)\triangleq\eta f_1(\cdot)+\bar{\eta} f_2(\cdot).$$

Let (X_0, θ_0) be a new sample, distributed as the observations in the training set. We wish to guess the classification θ_0 from the observation X_0.

If $f_1(\cdot), f_2(\cdot)$ and η are known, an optimal classifier is given by the Bayes decision rule:

$$\text{decide } \theta_0^*=1 \quad \text{if } f_1(X_0)/f_2(X_0)>\bar{\eta}/\eta,$$

$$\text{decide } \theta_0^*=2 \quad \text{if } f_1(X_0)/f_2(X_0)<\bar{\eta}/\eta.$$

The corresponding probability of error, the Bayes risk R^*, is given by

$$R^*\triangleq\Pr\{\theta_0^*\neq\theta_0\}=\int \min\{\eta f_1(x), \bar{\eta} f_2(x)\}\,dx=\min \Pr\{\tilde{\theta}(X_0)\neq\theta_0\}, \tag{2}$$

where the last minimum is taken over all measurable functions $\tilde{\theta}(\cdot)$ from $\mathbb{R}$ to $\{1, 2\}$.

Thus in this paper we assume that $f_1(\cdot), f_2(\cdot)$ and η are unknown, that the training set contains an infinite number of unlabeled samples, $X'_1, X'_2, ...$, and a finite number of labeled samples, $\{(X_1, \theta_1), ..., (X_l, \theta_l)\}$. A new sample (X_0, θ_0) is to be labeled based on the observation X_0.

2. Labeled and unlabeled samples in classification rules

Since $f_1(\cdot)$ and $f_2(\cdot)$ are both in $\mathcal{F}$, the distribution $f_{X'}(\cdot)=\eta f_1(\cdot)+\bar{\eta} f_2(\cdot)$ of the unlabeled samples belongs to the family of mixtures

$$\mathcal{H}\triangleq\{\zeta g_1(\cdot)+\bar{\zeta} g_2(\cdot): g_1(\cdot), g_2(\cdot)\in\mathcal{F}, g_1(\cdot)\neq g_2(\cdot), \zeta\in(0, 1)\}. \tag{3}$$

We say that $\mathcal{H}$ is *identifiable* if the map $(0, 1)\times\mathcal{F}\times\mathcal{F}\to\mathcal{H}$, induced by the class of probability measures that assign total mass 1 to two distinct elements of $\mathcal{F}$, is one to one.

The theory of mixtures of distributions has been extensively developed by Teicher in a series of papers (Teicher, 1960, 1963). A comprehensive treatment of mixture distributions, as well as a large bibliography, can be found in McLachlan and Basford (1988).

The mixture $f_{X'}(\cdot)\triangleq\eta f_1(\cdot)+\bar{\eta} f_2(\cdot)$ can be estimated from the unlabeled samples. Labeled observations are not needed for this purpose. If the family of mixtures $\mathcal{H}$ is identifiable, the measure on $\mathcal{F}$ that generates $f_{X'}(\cdot)$ can be uniquely recovered from $f_{X'}(\cdot)$ itself. It follows that unlabeled data alone can be used to estimate the single-component distributions and the mixing parameter consistently. Nevertheless, in the current framework,

we cannot construct a classification rule relying only on the unlabeled samples. To see this, consider the case $u=\infty$ and let $l=0$. Rewrite the mixture distribution as

$$f_{X'}(\cdot)=\eta f_1(\cdot)+\bar{\eta} f_2(\cdot)=\zeta g_1(\cdot)+\bar{\zeta} g_2(\cdot)\,,$$

where $g_1(\cdot)$ and $g_2(\cdot)$ are conventional names (for instance call $g_1(\cdot)$ the distribution associated with the larger mixing coefficient). A priori it is not known if $g_1(\cdot)$ is the distribution of samples of class 1 or the distribution of samples of class 2. When we observe an infinite number of labeled samples, we can (almost surely) recover $f_{X'}(\cdot)$, and by identifiability we can decompose the mixture and thus obtain $g_1(\cdot)$, $g_2(\cdot)$ and ζ. [1] We cannot tell, though, whether $g_1(\cdot)=f_1(\cdot)$ and $g_2(\cdot)=f_2(\cdot)$ or if the opposite holds. To formalize this statement, let the random variable Z be defined by

$$Z=1 \quad \text{if } g_1(\cdot)=f_1(\cdot),\ g_2(\cdot)=f_2(\cdot) \text{ and } \zeta=\eta\,,$$

$$Z=2 \quad \text{if } g_1(\cdot)=f_2(\cdot),\ g_2(\cdot)=f_1(\cdot) \text{ and } \zeta=\bar{\eta}\,,$$

and assume $P\{Z=1\}=P\{Z=2\}=\frac{1}{2}$. It is easily checked that Z is independent of the unlabeled samples $\{X'_1, \ldots, X'_u\}$, for all u. By rewriting $\Pr\{\theta_0=i \mid X_0, \{X'_j\}_{j=1}^{\infty}\}$, $i=1, 2$, in terms of the conditional expectation given Z we conclude that

$$R(0,\infty)=R(0,u)=\tfrac{1}{2} \quad \text{for all } u\,. \tag{4}$$

Thus, the probability of error is equal to $1/2$ if there are no labeled samples, even in the presence of an infinite number of unlabeled samples. Incidentally, note that, when the training set is composed of one labeled observation (i.e., $l=1$, $u=0$), the probability of error of the test that decides $\theta_0=\theta_1$, equals $2\eta\bar{\eta}\leqslant 1/2$, the inequality being strict for $\eta\neq 1/2$. Thus $R(1,0)\leqslant 2\eta\bar{\eta}\leqslant 1/2=R(0,u)$, $\forall u$.

The natural interpretation of Eq. (4) in light of the preceding discussion is that from the unlabeled samples we can recover the *decision regions* of an optimum test, but not the corresponding *decisions*. We can obtain the decision regions using $\{X'_j\}_{j=1}^{\infty}$, but we need labeled samples to label them.

3. The value of labeled samples

We now show that the first labeled sample reduces the probability of error to $2R^*(1-R^*)$, which is less than twice the Bayes risk.

Theorem 1. *Let the family of mixtures $\mathcal{H}$ be identifiable. Let the training set Q be composed of an infinite number of unlabeled samples and of one labeled sample. The probability of error of the optimum classification rule is*

$$R(1,\infty)=2R^*(1-R^*)\,.$$

Proof. Let (X_1, θ_1) be the labeled sample in the training set. Since the decision problem involves two simple hypotheses, the likelihood ratio test is optimum, as shown in Lehmann (1991), and can be written as

$$\text{decide } \hat{\theta}_0=1 \quad \text{if} \frac{\Pr\{\theta_0=1 \mid X_0, (X_1,\theta_1), \{X'_j\}_{j=1}^{\infty}\}}{\Pr\{\theta_0=2 \mid X_0, (X_1,\theta_1), \{X'_j\}_{j=1}^{\infty}\}}>1\,,$$

$$\text{decide } \hat{\theta}_0=2 \quad \text{if} \frac{\Pr\{\theta_0=1 \mid X_0, (X_1,\theta_1), \{X'_j\}_{j=1}^{\infty}\}}{\Pr\{\theta_0=2 \mid X_0, (X_1,\theta_1), \{X'_j\}_{j=1}^{\infty}\}}<1\,.$$

[1] Note that, once we agree on a convention for naming the densities, this decomposition is unique.

Conditioning on the unlabeled samples is (a.s.) equivalent to knowing the mixture $f_{X'}(\cdot)$ and, by identifiability, its components. Thus we can remove the conditioning on $\{X'_j\}_{j=1}^{\infty}$ and assume that $\zeta, g_1(\cdot)$ and $g_2(\cdot)$ are known. Trivially,

$$\frac{\Pr\{\theta_0=1 \mid X_0, (X_1,\theta_1)\}}{\Pr\{\theta_0=2 \mid X_0, (X_1,\theta_1)\}} = \frac{f(X_0 \mid \theta_0=1, (X_1,\theta_1))\Pr\{\theta_0=1 \mid (X_1,\theta_1)\}}{f(X_0 \mid \theta_0=2, (X_1,\theta_1))\Pr\{\theta_0=2 \mid (X_1,\theta_1)\}}.$$

The right-hand side can be evaluated by conditioning on Z and taking expectations, giving the numerator

$$\begin{aligned}
&f(X_0 \mid \theta_0=1, (X_1,\theta_1))\Pr\{\theta_0=1 \mid (X_1,\theta_1)\} \\
&\quad = f(X_0 \mid \theta_0=1, (X_1,\theta_1), Z=1)\Pr\{\theta_0=1 \mid (X_1,\theta_1), Z=1\}\Pr\{Z=1 \mid (X_1,\theta_1)\} \\
&\qquad + f(X_0 \mid \theta_0=1, (X_1,\theta_1), Z=2)\Pr\{\theta_0=1 \mid (X_1,\theta_1), Z=2\}\Pr\{Z=2 \mid (X_1,\theta_1)\} \\
&\quad = \zeta g_1(X_0)\Pr\{Z=1 \mid (X_1,\theta_1)\} + \bar{\zeta} g_2(X_0)\Pr\{Z=2 \mid (X_1,\theta_1)\} \\
&\quad = \frac{\zeta g_1(X_0)\zeta_{\theta_1} g_{\theta_1}(X_1)}{\zeta_{\theta_1} g_{\theta_1}(X_1) + \bar{\zeta}_{\theta_1} g_{\bar{\theta}_1}(X_1)} + \frac{\bar{\zeta} g_2(X_0)\bar{\zeta}_{\theta_1} g_{\bar{\theta}_1}(X_1)}{\zeta_{\theta_1} g_{\theta_1}(X_1) + \bar{\zeta}_{\theta_1} g_{\bar{\theta}_1}(X_1)}.
\end{aligned}$$

A similar expression holds for the denominator. Thus, the likelihood ratio becomes

$$\frac{\Pr\{\theta_0=1 \mid X_0, (X_1,\theta_1)\}}{\Pr\{\theta_0=2 \mid X_0, (X_1,\theta_1)\}} = \frac{\zeta g_1(X_0)\zeta_{\theta_1} g_{\theta_1}(X_1) + \bar{\zeta} g_2(X_0)\zeta_{\bar{\theta}_1} g_{\bar{\theta}_1}(X_1)}{\zeta g_1(X_0)\bar{\zeta}_{\theta_1} g_{\bar{\theta}_1}(X_1) + \bar{\zeta} g_2(X_0)\zeta_{\theta_1} g_{\theta_1}(X_1)},$$

and some algebra shows that the test can be rewritten as

$$\begin{aligned}
&\text{decide } \hat{\theta}_0=1 \quad \text{if } \zeta g_1(X_0)[\zeta_{\theta_1} g_{\theta_1}(X_1) - \zeta_{\bar{\theta}_1} g_{\bar{\theta}_1}(X_1)] > \bar{\zeta} g_2(X_0)[\zeta_{\theta_1} g_{\theta_1}(X_1) - \zeta_{\bar{\theta}_1} g_{\bar{\theta}_1}(X_1)], \\
&\text{decide } \hat{\theta}_0=2 \quad \text{if } \zeta g_1(X_0)[\zeta_{\theta_1} g_{\theta_1}(X_1) - \zeta_{\bar{\theta}_1} g_{\bar{\theta}_1}(X_1)] < \bar{\zeta} g_2(X_0)[\zeta_{\theta_1} g_{\theta_1}(X_1) - \zeta_{\bar{\theta}_1} g_{\bar{\theta}_1}(X_1)]. \qquad (5)
\end{aligned}$$

Interpretation. Knowing an infinite number of unlabeled samples is equivalent to identifying the densities $g_1(\cdot)$ and $g_2(\cdot)$ and the mixing parameter ζ. Therefore one can partition the sample space in two sets, $\mathcal{X}_1 = \{x: \zeta g_1(x) > \bar{\zeta} g_2(\cdot)\}$ and $\mathcal{X}_2 = \{x: \zeta g_1(x) < \bar{\zeta} g_2(\cdot)\}$, using the infinite number of unlabeled samples in the training set. Note that $\mathcal{X}_1$ and $\mathcal{X}_2$ are the decision regions induced by the Bayes decision rule: if $Z=1$, the Bayes decision rule decides $\theta_0=1$ when $X_0 \in \mathcal{X}_1$ and $\theta_0=2$ when $X_0 \in \mathcal{X}_2$. If $Z=2$ the decisions are reversed.

Test (5) *is equivalent to a two-stage procedure.*

Stage 1. *Label the components of the partition by deciding that* $Z=1$ *if* $\zeta_{\theta_1} g_{\theta_1}(X_1) > \bar{\zeta}_{\theta_1} g_{\bar{\theta}_1}(X_1)$ *and* $Z=2$ *if the inequality is reversed.*

Stage 2. *Assign the sample* X_0 *to* $\mathcal{X}_1$ *if* $\zeta g_1(X_0) > \bar{\zeta} g_2(X_0)$, *or to* $\mathcal{X}_2$ *if the inequality is reversed. Let* θ_0 *be the label of the component of the partition to which* X_0 *is assigned.*

A classification error occurs when either the first stage or the second stage yields an incorrect answer. When both stages result in wrong answers, the mistakes cancel each other. The interpretation leads to a simple way of calculating the probability of error associated with the optimal test (5). Define the event $A \triangleq \{\text{error in stage 1}\}$.

$$\begin{aligned}
R(1,\infty) &= \Pr\{\hat{\theta}_0 \neq \theta_0\} = \Pr\{\hat{\theta}_0 \neq \theta_0 \mid A\}P\{A\} + \Pr\{\hat{\theta}_0 \neq \theta_0 \mid \bar{A}\}P\{\bar{A}\} \\
&= (1-R^*)R^* + R^*(1-R^*) = 2R^*(1-R^*),
\end{aligned}$$

where $P(A) = R^*$, by the definition of A. Also $\Pr\{\hat{\theta}_0 \neq \theta_0 \mid A\}$ is the probability that stage 2 is correct and is equal to $1-R^*$. Similarly $P(\bar{A}) = 1-R^*$ and $\Pr\{\hat{\theta}_0 \neq \theta_0 \mid \bar{A}\} = R^*$, thus proving the theorem. □

If there are an infinite number of labeled and unlabeled samples, it follows that $f_1(\cdot)$, $f_2(\cdot)$ and η can be determined a.s., and the Bayes decision rule can be used. Thus $R(\infty, \infty) = R^*$. We next ask how $R(l, \infty)$ converges to R^*.

Theorem 2. *Let the number of unlabeled samples u be infinite, assume that the densities in $\mathscr{F}$ have common support and that the family of mixtures $\mathscr{M}$ is identifiable. The probability of error in classifying a new sample X_0 converges exponentially fast to the Bayes risk and satisfies*

$$-\lim_{l\to\infty} \frac{1}{l} \log(R(l, \infty) - R^*) = -\log\left(2\sqrt{\eta\bar{\eta}} \int \sqrt{f_1(x) f_2(x)}\, dx\right). \tag{6}$$

Note that the quantity $-\log(\int \sqrt{f_1(x) f_2(x)}\, dx)$ is the Bhattacharyya distance between the densities $f_1(\cdot)$ and $f_2(\cdot)$.

Proof. The arguments for the first part of the proof of Theorem 1 imply that the optimum test is equivalent to the following two-stage procedure.

Stage 1. Label the partitions by deciding that $Z=1$ if $\prod_{i=1}^{l} \zeta_{\theta_i} g_{\theta_i}(X_i) > \prod_{i=1}^{l} \bar{\zeta}_{\bar{\theta}_i} g_{\bar{\theta}_i}(X_i)$ and by deciding that $Z=2$ if the inequality is reversed.

Stage 2. Assign the sample X_0 to $\mathscr{X}_1$ if $\zeta g_1(X_0) > \bar{\zeta} g_2(X_0)$, or to $\mathscr{X}_2$ if the inequality is reversed. Let θ_0 be the label of the component of the partition to which X_0 is assigned.

Define $A^{(l)} = \{\text{error in stage 1}\}$, where the superscript (l) makes explicit the dependence of the event on the number of labeled samples in the training set. The probability of error is

$$\begin{aligned} R(l, \infty) &= \Pr\{\hat{\theta}_0 \neq \theta_0 \mid A^{(l)}\} \Pr\{A^{(l)}\} + \Pr\{\hat{\theta}_0 \neq \theta_0 \mid \bar{A}^{(l)}\} \Pr\{\bar{A}^{(l)}\} \\ &= \Pr\{\hat{\theta}_0 \neq \theta_0 \mid \bar{A}^{(l)}\} + (\Pr\{\hat{\theta}_0 \neq \theta_0 \mid A^{(l)}\} - \Pr\{\hat{\theta}_0 \neq \theta_0 \mid \bar{A}^{(l)}\}) \Pr\{A^{(l)}\} \\ &= R^* + (1 - 2R^*) \Pr\{A^{(l)}\}. \end{aligned} \tag{7}$$

Thus $R(l, \infty) - R^* = (1 - 2R^*) \Pr\{A^{(l)}\}$, where $A^{(l)}$ can be written in terms of the underlying densities as

$$A^{(l)} = \left\{\prod_{i=1}^{l} \frac{\eta_{\theta_i} f_{\theta_i}(X_i)}{\eta_{\bar{\theta}_i} f_{\bar{\theta}_i}(X_i)} < 1\right\} = \left\{\frac{1}{l} \sum_{i=1}^{l} \log \frac{\eta_{\theta_i} f_{\theta_i}(X_i)}{\eta_{\bar{\theta}_i} f_{\bar{\theta}_i}(X_i)} < 0\right\}.$$

From the distributional properties of the pairs $\{(X_i, \theta_i)\}$ and from the assumption that the densities $f_1(\cdot)$ and $f_2(\cdot)$ have common support, it follows that the quantities $Y_i \triangleq \log(\eta_{\theta_i} f_{\theta_i}(X_i) / (\bar{\eta}_{\theta_i} f_{\bar{\theta}_i}(X_i)))$ are i.i.d. random variables. The expectation of Y_i is positive and is equal to

$$D(f(X, \theta) \,\|\, f(X, \bar{\theta})) = \sum_{\theta=1}^{2} \int \eta_\theta f_\theta(x) \log \frac{\eta_\theta f_\theta(x)}{\bar{\eta}_\theta f_{\bar{\theta}}(x)}\, dx,$$

the relative entropy between the joint distributions $f(X, \theta)$ and $f(X, \bar{\theta})$, where $f(X, \theta)$ is given in Eq. (1), and $f(X, \bar{\theta}) = [\eta f_1(X)]^{\mathbf{1}(\theta=2)} [\bar{\eta} f_2(X)]^{\mathbf{1}(\theta=1)}$.

The large sample behavior of the average of i.i.d. random variables is governed by Cramer's theorem. By a standard large deviations argument (Cover and Thomas, 1991) it is easily seen that

$$-\lim_{l\to\infty}\frac{1}{l}\log \Pr\{A^{(l)}\}=\Lambda^*(0) \tag{8}$$

$$=-\inf_{\zeta\le 0}\Lambda(\zeta) \tag{9}$$

$$=-\inf_{\zeta\le 0}\log\left[\int\left(\frac{\eta f_1(x)}{\bar{\eta} f_2(x)}\right)^{\zeta}\eta f_1(x)\,dx+\int\left(\frac{\bar{\eta} f_2(x)}{\eta f_1(x)}\right)^{\zeta}\bar{\eta} f_2(x)\,dx\right] \tag{10}$$

$$=-\log\left[2\int\sqrt{\eta f_1(x)\bar{\eta} f_2(x)}\,dx\right] \tag{11}$$

where $\Lambda(\zeta)$ is the log-moment generating function of the random variable (X,θ), and $\Lambda^*(x)$ is the corresponding Fenchel–Legendre transform (the rate function). Equality (8) follows directly from Cramer's theorem by the convexity of $\Lambda^*(\cdot)$, (9) follows from Lemma 2.2.5 of Dembo and Zeitouni (1993), (10) follows the definition of logarithmic moment generating function of the random variable $\log(\eta_\theta f_\theta(X)/(\bar{\eta}_\theta f_{\bar{\theta}}(X)))$. The minimization that leads to (1) can be performed by differentiating under the integral sign, and is consistent with the symmetry inherent in the problem. The theorem follows immediately from (7) and (11). □

4. Discussion and conclusions

The identifiability of the family of mixtures $\mathscr{H}$ defined in Eq. (3) enables us to estimate the boundaries of the decision regions from the unlabeled observations. Thus, if we have an infinite number of unlabeled samples, we can find the partition of the sample space that corresponds to the Bayes decision rule with probability one. Nevertheless, from the unlabeled samples $\{X_j'\}_{j=1}^{\infty}$ we cannot infer the decisions associated with the components of the partition. Thus

$$R(0,u)=R(0,\infty)=\tfrac{1}{2}\,.$$

We need labeled samples to label the decision regions. The first labeled sample reduces the probability of error from $R(0,\infty)=1/2$ to

$$R(1,\infty)=2R^*(1-R^*)\,,$$

where R^* is the Bayes probability of error. (It is simply a coincidence that this equality is the same as the upper bound for the nearest neighbor risk.)

When the number of labeled samples is equal to infinity,

$$R(\infty,u)=R(\infty,\infty)=R^*\,.$$

The risk converges exponentially fast to the Bayes risk in the number l of labeled samples according to

$$R(l,\infty)-R^*=\exp\left\{l\log\left(2\sqrt{\eta\bar{\eta}}\int\sqrt{f_1(x)f_2(x)}\,dx\right)+o(l)\right\},$$

where the exponent is the Bhattacharyya distance between the densities $f_1(x)$ and $f_2(x)$ plus a term equal to $\log(2\sqrt{\eta\bar{\eta}})$.

Thus labeled samples have an exponential value in reducing the probability of error. Their essential property, which makes them qualitatively different from unlabeled samples, is the information that they carry about the decisions associated with the decision regions. This property should be exploited in the construction of classification rules when the cost of acquiring labeled observations for the training set is high compared with the cost of obtaining unlabeled observations.

References

Cover, T.M. and J.A. Thomas (1991). *Elements of Information Theory.* Wiley, New York.

Dembo, A. and O. Zeitouni (1993). *Large Deviations, Techniques and Applications.* Jones & Bartlett, Boston, MA.

Devroye, L. (1988). Automatic pattern recognition: a study of the probability of error. *IEEE Trans. Pattern Anal. Machine Intell.* 10 (4), 530–543.

Greenspan, H., R. Goodman and R. Chellappa (1991). Texture analysis via unsupervised and supervised learning. *Internat. Joint Conf. on Neural Networks.* July 8–12, 1991, Seattle, 1, 639–644.

Jain, A.K. and B. Chandrasekaran (1982). Dimensionality and sample size considerations in pattern recognition practice. In: P.R. Krishnaiah and L.N. Kanal, Eds., *Handbook of Statistics,* Vol. 2, 835–855.

Kester, A.D.M. (1985). *Some Large Deviation Results in Statistics.* Centrum voor Wiskunde en Informatica, Amsterdam.

Lehmann, E.L. (1991). *Testing Statistical Hypotheses.* Wadsworth & Brooks/Cole, Belmont, CA.

McLachlan, G.J. (1977). Estimating the linear discriminant function from initial samples containing a small number of unclassified observations. *J. Amer. Statist. Assoc.* 72 (358), 403–406.

McLachlan, G.J. and K.E. Basford (1988). *Mixture Models.* Marcel Dekker, New York.

McLachlan, G.J. and S. Ganesalingam (1982). Updating the discriminant function on the basis of unclassified data. *Communications Statistics – Simulation* 11 (6), 753–767.

O'Neill, T.J. (1978). Normal discrimination with unclassified observations. *J. Amer. Statist. Assoc.* 73 (364), 821–826.

Pao, Y.-H. and D.J. Sobajic (1992). Combined use of supervised and unsupervised learning for dynamic security assessment. *IEEE Trans. Power Systems* 7 (2), 878–884.

Raudys, S.J. and A.K. Jain (1991). Small sample size effects in statistical pattern recognition: recommendations for practitioners. *IEEE Trans. Pattern Anal. Machine Intell.* 13 (3), 252–264.

Shahshahani, B.M. and D.A. Landgrebe (1992). On the asymptotic improvement of supervised learning by utilizing additional unlabeled samples; normal mixture density case. *SPIE* 1766, 143–155.

Teicher, H. (1960). On the mixtures of distributions. *Ann. Math. Statist.* 31, 55–73.

Teicher, H. (1963). Identifiability of finite mixtures. *Ann. Math. Statist.* 38, 1265–1269.

Tolat, V.V. and A.M. Peterson (1990). Nonlinear mapping with minimal supervised learning. *Proc. Hawaii Internat. Conf. on System Science,* 1, Jan. 2–5.

Vapnik, V. (1982). *Estimation of Dependences Based on Empirical Data.* Springer, Berlin.