Classification rules in the unknown mixture parameter case: relative value of labeled and unlabeled samples.

Vittorio Castelli Thomas M. Cover

Information Systems Laboratory, Department of Electrical Engineering, 4055 Stanford University, Stanford, California 94305.

Abstract — **We investigate the relative value of labeled and unlabeled samples in constructing classification rules. We observe a training set Q composed of l labeled and u unlabeled samples coming from two classes. Let sample from class 1 be distributed according to $f_1(\cdot)$, samples from class 2 according to $f_2(\cdot)$, and let η be the probability that a sample is in class 1. Assume that $f_1(\cdot)$ and $f_2(\cdot)$ are known and that η is unknown. We want to classify a new sample X_0. The relative value of labeled and unlabeled observations in reducing the probability of error is equal to $I_l(\eta)/I_u(\eta)$, the ratio of the Fisher informations of the labeled and unlabeled samples. Moreover labeled samples are not necessary in order to construct a decision rule.**

However, if $f_1(\cdot)$ and $f_2(\cdot)$ are given, but it is not known whether observations from class 1 are distributed according to $f_1(\cdot)$ or according to $f_2(\cdot)$, then the labeled samples are necessary and exponentially more valuable than unlabeled samples.

I. Summary

We observe a training set Q composed of labeled samples $\{(X_1,\theta_1),\cdots,(X_l,\theta_l)\}$ and unlabeled samples $\{X_1',\cdots,X_u'\}$. Let the labels $\{\theta_i\}_{i=1}^{l}$ be i.i.d. Bernoulli(η) random variables on the set $\{1,2\}$, let the observations $\{X_i\}_{i=1}^{l}$ be independent random variables distributed according to $f_{\theta_i}(\cdot)$ and let the unlabeled samples $\{X_j'\}_{j=1}^{u}$ be i.i.d. according to the mixture distribution $\eta f_1(\cdot)+\overline{\eta} f_2(\cdot)$. Assume that $f_1(\cdot)$ and $f_2(\cdot)$ are known densities, absolutely continuous with respect to each other and three times differentiable. Let the mixing parameter η be unknown. Let (X_0,θ_0) be a new sample independently distributed as the observations in the training set.

We want to infer the classification θ_0 from the observation X_0. We consider rules based on $f_1(\cdot)$, $f_2(\cdot)$ and Q and we wish to minimize the overall probability of error in classifying X_0. We restrict attention to admissible tests (those not uniformly dominated) and, in this class, we consider rules with asymptotically efficient properties. Bayes tests with respect to smooth priors $h(\eta)$ for the parameter η are typical examples.

Theorem 1 *The Bayes solution to the classification problem has the form*

$$\text{Decide } \theta_0=1 \quad \text{if} \quad \frac{f_1(x_0)}{f_2(x_0)} > \frac{1-E[\eta\mid Q]}{E[\eta\mid Q]} \triangleq \frac{1-\hat{\eta}}{\hat{\eta}}.$$

$$\text{Decide } \theta_0=2 \quad \text{otherwise.}$$

Now let η_0 be the true value of the mixing parameter. The difference between the probability of error $R_{l,u}$ of the test and the Bayes Risk R^* can be expressed as

$$\Delta R_{l,u} \triangleq R_{l,u}-R^* = c\,E(\hat{\eta}-\eta_0)^2 + o\left(\frac{1}{l+u}\right)$$

where c depends on $f_1(x)$, $f_2(x)$ and η_0, but not on $h(\eta)$, l and u. □

It can be shown, by extending Theorem 6.7.1 in Lehmann [1], that $\hat{\eta}$ is an asymptotically efficient estimator of η and that the second moment of $\sqrt{l+u}(\hat{\eta}-\eta_0)$ converges to $(l+u)\left[l\,I_l(\eta_0)+u\,I_u(\eta_0)\right]^{-1}$, where

$$I_l(\eta_0)=\frac{1}{\eta_0\overline{\eta}_0}, \qquad I_u(\eta_0)=E\left[\left(\frac{f_1(X)-f_2(X)}{\eta_0 f_1(X)+\overline{\eta}_0 f_2(X)}\right)^2\right]$$

are the Fisher informations associated with the labeled observations and with the unlabeled observations respectively. Thus labeled samples are I_l/I_u times more valuable than unlabeled samples.

Now consider a different problem and assume that $f_1(\cdot)$ and $f_2(\cdot)$ are known, but it is not known whether samples from class 1 are distributed according to $f_1(\cdot)$ or according to $f_2(\cdot)$. To make the statement precise, define a new random variable Z, let $Z=1$ if $f_1(\cdot)$ is the distribution of class 1, $Z=2$ if the opposite holds, and let $\Pr\{Z=1\}=\Pr\{Z=2\}=1/2$ and let all the remaining assumptions hold.

Theorem 2 *For any smooth prior $h(\eta)$, if $l^3u^{-1}\to 0$,*

$$\Delta R_{l,u} \triangleq R_{l,u}-R^* = O\left(u^{-1}\right)+exp(-l(D+o(1)))$$

where $D=-\log\left(2\sqrt{\eta_0\overline{\eta}_0}\int\sqrt{f_1(x)f_2(x)}dx\right)$. □

II. Conclusions

In the first case the additional probability of error $\Delta R_{l,u}$ is due to errors in estimating the true value of the mixing parameter η_0, which result in errors in the boundaries of the decision regions. The labeled samples are I_l/I_u times more valuable than the unlabeled samples in reducing the extra risk $\Delta R_{l,u}$. However a decision rule can be constructed using unlabeled samples only.

In the second case Z is unknown and thus not only the boundaries but also the labels of the decision regions must be inferred from the training set. Since only the labeled samples carry information about Z we need them to construct a classification rule. The probability of error in labeling the decision regions converges exponentially to zero in the number of labeled samples. If the number of unlabeled samples u grows faster than $\exp(Dl)$, the additional probability of error is asymptotically equivalent to the probability of labeling the decision regions incorrectly, namely $\Delta R_{l,u}\sim\exp(-Dl)$. Conversely, if $u\exp\{-Dl\}\to 0$, $\Delta R_{l,u}$ is determined by the error in estimating η_0 from the data and is asymptotically equivalent to $c\,(I_{unl}(\eta_0)\,u)^{-1}$. We conclude that in the second case labeled samples are necessary and exponentially more valuable than unlabeled samples in constructing a classification rule.

References

[1] Erich L. Lehmann. *Theory of point estimation.* Wiley series in mathematical probability. John Wiley & Sons, Inc., 1983.

0 - 7803-2015-8/94/$4.00 ©1994 IEEE