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 \( n \) unlabeled samples coming from two classes. Let \( \eta \) be the probability that a sample is in class 1. Assume that \( f_1(\cdot) \) and \( f_2(\cdot) \) are known and that \( \eta \) 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 \( I(\eta)/I_0(\eta) \), 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, 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.

1. SUMMARY

We observe a training set \( Q \) composed of labeled samples \( \{(X_1, \theta_1), \ldots, (X_l, \theta_l)\} \) and unlabeled samples \( \{X_1, \ldots, X_n\} \). Let the labels \( \{\theta_1, \ldots, \theta_l\} \) be i.i.d. Bernoulli(\( \eta \)) random variables on the set \( \{1, 2\} \), let the observations \( \{X_1, \ldots, X_l\} \) be independent random variables distributed according to \( f_\theta(\cdot) \) and let the unlabeled samples \( \{X_1, \ldots, X_n\} \) be i.i.d. according to the mixture distribution \( \eta f_1(\cdot) + (1 - \eta) f_2(\cdot) \). 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 \( \eta \) be unknown. Let \( (X_0, \theta_0) \) be a new sample independently distributed as the observations in the training set.

We want to infer the classification \( \theta_0 \) from the observation \( X_0 \). We consider rules based on \( f_1(\cdot) \), \( f_2(\cdot) \) and \( Q \) and 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 \( \eta \) are typical examples.

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

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

Decide \( \theta_0 = 2 \) otherwise.

Now let \( \eta_0 \) be the true value of the mixing parameter. The difference between the probability of error \( R_{\eta_0} \) of the test and the Bayes risk \( R^* \) can be expressed as

\[
\Delta R_{\eta_0} = R_{\eta_0} - R^* = \kappa E(\hat{\eta} - \eta_0)^2 + o\left(\frac{1}{\eta_0^4}\right)
\]

where \( \kappa \) depends on \( f_1(z) \), \( f_2(z) \) and \( \eta_0 \), but not on \( h(\eta) \), \( l \) and \( n \).

It can be shown, by extending Theorem 6.7.1 in Lehmann [1], that \( \hat{\eta} \) is an asymptotically efficient estimator of \( \eta \) and that the second moment of \( \sqrt{l + n}(\hat{\eta} - \eta_0) \) converges to

\[
(l + n) \left( \frac{1}{\eta_0^2} \right) \left( \frac{1}{\eta_0^2} + \frac{1}{\eta_0^2} \right) + \frac{1}{\eta_0^4}
\]

where \( I_1(\eta_0) = \frac{1}{\eta_0^2} \), \( I_2(\eta_0) = E \left( \frac{f_1(x) - f_2(x)}{\eta_0 f_1(x) + (1 - \eta_0) f_2(x)} \right) \)

are the Fisher informations associated with the labeled observations and with the unlabeled observations respectively. Thus labeled samples are \( n/l \) 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 \( n^{-1} \to 0 \),

\[
\Delta R_{\eta_0} \xrightarrow{\Delta} R_{\eta_0} - R^* = O\left(\frac{1}{\eta_0^4} \right) + \exp\left(\frac{-1}{D + D_1(1)}\right)
\]

where

\[
D = -\log\left(2 \frac{1}{\eta_0} \int \sqrt{f_1(x) f_2(x)} dx\right).
\]

II. CONCLUSIONS

In the first case the additional probability of error \( \Delta R_{\eta_0} \) is due to errors in estimating the true value of the mixing parameter \( \eta \), which result in errors in the boundaries of the decision regions. The labeled samples are \( n/l \) times more valuable than the unlabeled samples in reducing the extra risk \( \Delta R_{\eta_0} \). 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 \( n \) grows faster than \( \exp(D) \), the additional probability of error is asymptotically equivalent to the probability of labeling the decision regions incorrectly, namely \( \Delta R_{\eta_0} \xrightarrow{\Delta} 0 \). Conversely, if \( n \exp(-D) \to 0 \), \( \Delta R_{\eta_0} \) is determined by the error in estimating \( \eta_0 \) from the data and is asymptotically equivalent to \( (I_{\eta_0}(\eta_0) n)^{-1} \). We conclude that in the second case labeled samples are necessary and exponentially more valuable than unlabeled samples in constructing a classification rule.

REFERENCES


0 - $7803-2015-8/94$/$4.00 ©1994 IEEE - 111 -