Upper Bounds for Bayes Error

G. M. Dilshan Godaliyadda
Class: ECE 662
Spring 2014
Student Lecture

1 Introduction

This lecture is dedicated to explain upper bounds for Bayes Error. In order to motivate the problem we will first discuss what Bayes rule is, and how it is used to classify continuously valued data. Then we will present the probability of error that results from using Bayes rule.

When Bayes rule is used the resulting probability of error is the smallest possible error, and therefore becomes a very important quantity to know. Although, in many cases computing this error is intractable and therefore having an upper bound to the Bayes error becomes useful. The **Chernoff Bound** is one such upper bound which is reasonably easy to compute in many cases. Therefore we will present this bound, and then discuss a few results that can be attained for certain types of distributions.

2 Bayes rule for classifying continuously valued data

Let's assume that the features x are continuously valued and that $x \in \mathbb{R}^N$. We will also assume that the data belongs to two different classes, ω_1 and ω_2. Then our objective is to classify the data in to these two classes.

We can classify x to be in ω_1 if the probability of class 1 given the feature vector x is larger than the probability of class 2 given the feature vector x. Hence,

$$Prob(\omega_1|x) \geq Prob(\omega_2|x) \tag{1}$$

then that particular feature vector x belongs in class 1, and vise versa. Although, in many cases calculating $\rho(\omega_i|x)$ is impossible, or extremely difficult. Therefore, we use Bayes theorem to simplify our problem.

Bayes theorem states,

$$\rho(\omega_i|x) = \frac{\rho(x|\omega_i)Prob(\omega_i)}{\rho(x)}. \tag{2}$$

Here, $\rho(x|\omega_i)$ is the probability density of x given it belongs to class i, $Prob(\omega_i)$ known as the prior probability is the probability of class i, and $\rho(x)$ is the probability density of x. Using equation (2), inequality (1) can be re-written as,

$$\frac{\rho(x|\omega_1)Prob(\omega_1)}{\rho(x)} \geq \frac{\rho(x|\omega_2)Prob(\omega_2)}{\rho(x)}$$

$$\rho(x|\omega_1)Prob(\omega_1) \geq \rho(x|\omega_2)Prob(\omega_2). \tag{3}$$

3 Probability of Error when using Bayes rule

When using a certain method for classifying data, it is important to know the probability of making an error, because the ultimate goal of classification is to minimize this error. When we use Bayes rule to classify data, the probability of error is,

$$Prob\left[Error\right] = \int_{\mathbb{R}^N} \rho\left(Error, x\right)\, \mathrm{d}x \tag{4}$$

Then using the definition of conditional probability we can rewrite this equation as,

$$Prob\left[Error\right] = \int_{\mathbb{R}^N} Prob\left(Error|x\right) \rho\left(x\right)\, \mathrm{d}x \tag{5}$$

Now let us look at an example to understand how we can write this integral in terms of $Prob\left(\omega_1|x\right)$ and $Prob\left(\omega_1|x\right)$.

From equation (2) we know that $Prob\left(\omega_i|x\right) \propto \rho(x|\omega_i)Prob(\omega_i)$, and therefore $\rho(x|\omega_i)Prob(\omega_i)$ and $Prob\left(\omega_i|x\right)$ will have the same shape. Therefore, we plot $\rho(x|\omega_1)Prob(\omega_1)$ and $\rho(x|\omega_2)Prob(\omega_2)$ on the same plot as shown in Figure 1.

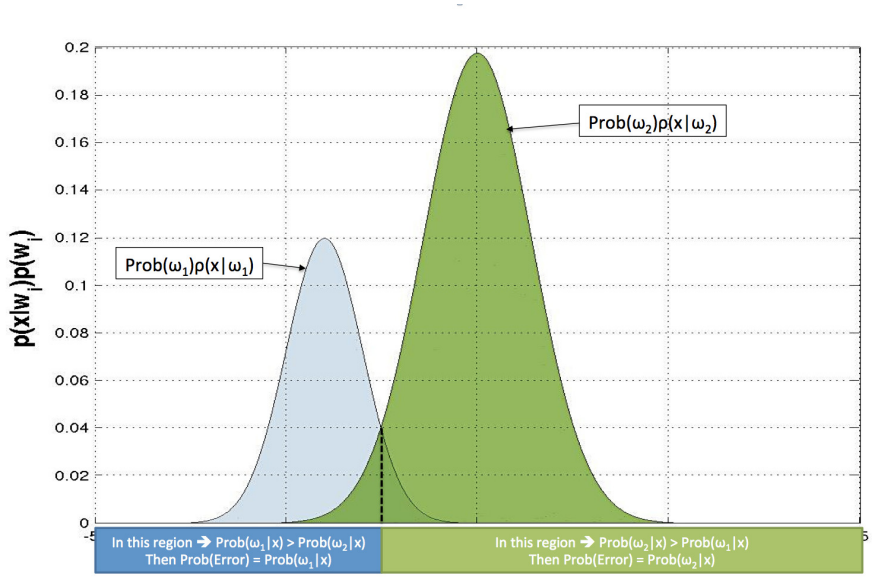


Figure 1: $\rho(x|\omega_1)Prob(\omega_1)$ and $\rho(x|\omega_2)Prob(\omega_2)$

From this we can see that errors happen in the regions where the two curves intersect. For example, if we consider a point in the region where $\rho(x|\omega_1)Prob(\omega_1)$ and $\rho(x|\omega_2)Prob(\omega_2)$ intersect, and $\rho(x|\omega_1)Prob(\omega_1) \geq \rho(x|\omega_2)Prob(\omega_2)$, it will be classified as Class 1.

Now since $Prob\left(\omega_i|x\right) \propto \rho(x|\omega_i)Prob(\omega_i)$, our expression for the probability of error can be written as,

$$Prob\left[Error\right] = \int_{\mathbb{R}^N} \min\left(Prob\left(\omega_1|x\right), Prob\left(\omega_2|x\right)\right)\, \mathrm{d}x \tag{6}$$

It is important to note that the above example is a trivial case when there is just one intersection between the curves. There can be any number of intersections in the general

case, but even still the same equation holds. In this case the probability of error will be the summation of the integrals over the lesser of the $Prob\left(\omega_i|x\right)$'s in the different regions, which once more is given by equation (6).

By using Bayes Theorem, we can simplify further so that,

$$Prob\left[Error\right] = \int_{\mathbb{R}^N} \min\left(\rho\left(x|\omega_1\right) Prob\left(\omega_1\right), \rho\left(x|\omega_2\right) Prob\left(\omega_2\right)\right) \, \mathrm{d}x \tag{7}$$

In most cases solving either integral is intractable, and therefore the need for an approximate answer becomes necessary.

Of course our goal will always be to get the smallest possible probability of error. Although, we know Bayes error is almost impossible to calculate in many cases. So instead what we can do is find an upper bound to the Bayes error.

For example let us assume that we want the probability of error to be lower than 0.05. Now if we find an upper bound to Bayes error that is less than 0.05, we definitely know that Bayes error is less than 0.05. In the next section we will explore a single parameter family of upper bounds to Bayes error, the **Chernoff Bound.**

4 Chernoff Bound

To formulate the Chernoff bound we use the following mathematical property:

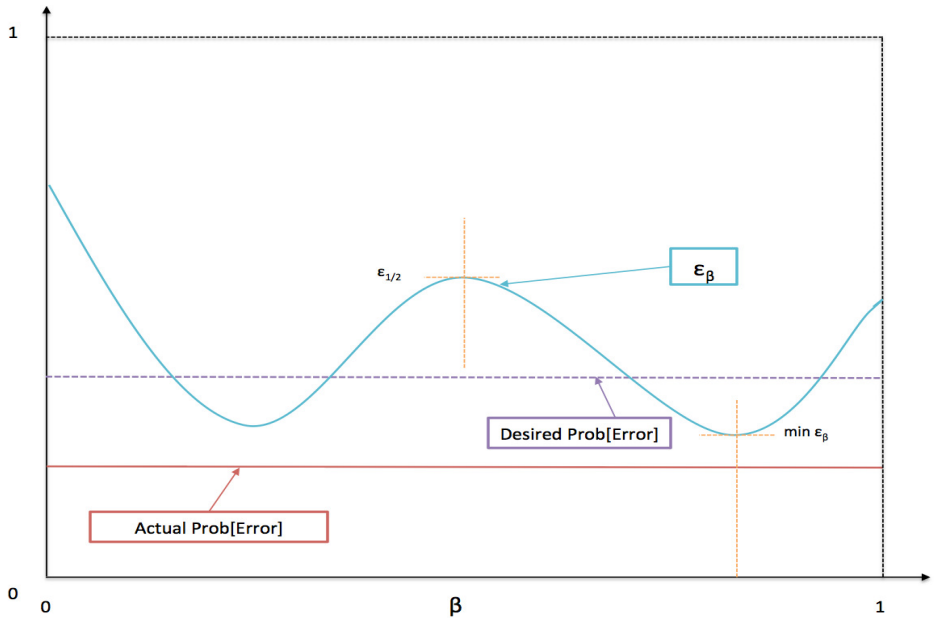
If $a, b \in \mathbb{R}_{\geq 0}$, and $0 \leq \beta \leq 1$ then,

$$min\left\{a, b\right\} \leq a^{\beta} b^{1-\beta} \tag{8}$$

Since equation (7) has probabilities and they are non-negative, we can use inequality (8) and get the following upper bound for the probability of error,

$$\begin{aligned} Prob\left[Error\right] &\leq \int_{\mathbb{R}^N} \left[Prob\left(\omega_1|x\right)\right]^{\beta} \left[Prob\left(\omega_2|x\right)\right]^{1-\beta} \mathrm{d}x & (9) \\ &\leq \int_{\mathbb{R}^N} \left[\rho\left(x|\omega_1\right) Prob\left(\omega_1\right)\right]^{\beta} \left[\rho\left(x|\omega_2\right) Prob\left(\omega_2\right)\right]^{1-\beta} \mathrm{d}x & (10) \end{aligned}$$

This family of upper bounds parameterized by β is known as the **Chernoff bound** (ε_β).



If needed it is possible to find the smallest error bound by solving,

$$\varepsilon_{min} = \min_{\beta \in [0,1]} \varepsilon_\beta. \tag{11}$$

Although, this is not necessarily an easy problem to solve, because the optimization problem might not be convex and might have local minimums. For the purpose of evaluating

a classification method one can try with different values for β, and compare the least value of ε_β with the desired error probability.

In the special case where $\beta = \frac{1}{2}$, the Chernoff bound, $\varepsilon_{\frac{1}{2}}$, is known as the **Bhattacharyya bound**.

4.1 Chernoff bound for Normally distributed data

When $\rho(x|\omega_i)$ is Normally distributed and has mean μ_i and covariance Σ_i, the Chernoff bound has the following closed form solution,

$$\varepsilon_\beta = Prob(\omega_1)^\beta Prob(\omega_2)^{1-\beta} e^{-f(\beta)} \tag{12}$$

where,

$$f(\beta) = \frac{\beta(1-\beta)}{2}(\mu_2-\mu_1)^t \left[\beta\Sigma_1 + (1-\beta)\Sigma_2\right]^{-1}(\mu_2-\mu_1) + \frac{1}{2}\ln\left(\frac{|\beta\Sigma_1 + (1-\beta)\Sigma_2|}{|\Sigma_1|^\beta |\Sigma_2|^{1-\beta}}\right).$$

In this case the Bhattacharyya bound becomes,

$$\varepsilon_{\frac{1}{2}} = \sqrt{Prob(\omega_1)Prob(\omega_2)}\, e^{f(\frac{1}{2})} \tag{13}$$

where,

$$f\left(\frac{1}{2}\right) = \frac{1}{8}(\mu_2-\mu_1)^t \left[\frac{\Sigma_1+\Sigma_2}{2}\right]^{-1}(\mu_2-\mu_1) + \frac{1}{2}\ln\left(\frac{|\frac{\Sigma_1+\Sigma_2}{2}|}{|\Sigma_1|^{\frac{1}{2}}|\Sigma_2|^{\frac{1}{2}}}\right).$$

Notice that if $\mu_2 = \mu_1$, then the first term in $f(\frac{1}{2})$ disappears. This term basically describes the separability due to the "distance" between the two mean values. Here the "distance" is described by metric defined by the matrix $\left[\frac{\Sigma_1+\Sigma_2}{2}\right]^{-1}$.

The second term disappears if the covariance matrices are equal, i.e. $\Sigma_1 = \Sigma_2$. Also, in this case it can be shown that, $\varepsilon_{\frac{1}{2}} = \varepsilon_{min}$.

Proof:
We want to minimize ε_β, which means we want maximize $f(\beta)$.

$$f(\beta) = \frac{\beta(1-\beta)}{2}(\mu_2-\mu_1)^t \left[\beta\Sigma_1 + (1-\beta)\Sigma_2\right]^{-1}(\mu_2-\mu_1)$$

Since this is convex in β we know that there is only one maximum and that maximum is the global maximum. So we take the derivative w.r.t. β, and set it equal to zero to find the maximum.

$$\begin{aligned}\frac{\partial}{\partial\beta}f(\beta) &= 0 \\ &= \frac{(1-2\beta)}{2}(\mu_2-\mu_1)^t \left[\beta\Sigma_1 + (1-\beta)\Sigma_2\right]^{-1}(\mu_2-\mu_1)\end{aligned}$$

if $\mu_1 = \mu_2$, then $f(\beta) = 0$, and $\varepsilon_\beta = constant$. This means $\varepsilon_{min} = \varepsilon_{\frac{1}{2}}$.

if $\mu_1 \neq \mu_2$, then the β that maximizes $f(\beta)$ is $\beta = \frac{1}{2}$. Then $\varepsilon_{min} = \varepsilon_{\frac{1}{2}}$.

5 Summary and Conclusions

In this lecture we have shown that the probability of error ($Prob\,[Error]$) when using Bayes error, is upper bounded by the Chernoff Bound. Therefore,

$$Prob\,[Error] \leq \varepsilon_\beta \tag{14}$$

for $\beta in\,[0, 1]$.

When $\beta = \frac{1}{2}$ then$\varepsilon_{\frac{1}{2}}$ in known as the Bhattacharyya bound.

References

[1]. Duda, Richard O. and Hart, Peter E. and Stork, David G., "Pattern Classification (2^{nd} Edition)," Wiley-Interscience, 2000.