CSCE 970: Pattern Recognition	**Spring Semester, 1999**
Topic [5]: [Class Separability Measures]	
Lecturer: Stephen Scott	*Scribe: [Xiangdong Lou]*

1 Introduction

The emphasis of previous lectures was on techniques referring to the discrimination properties of idividual features. However, after successfully removing all the individual features known to be lack of discriminatory power, can we guarantee that we have gotten the most concise set of features for classification? For most situations, the answer will be no. Due to the existence of the correlations among the various features, the total discriminatory power of a group of features may not equal to the sum of each of the individual ones, which means that we can still prune away some of the features to make a more concise feature groups. In order to decide which subet should be pruned away and which subset should be kept, we need some quantative measurement for the discriminatory power of a group of features (feature vector). Measuring the discrimination effectiveness of feature vectors will be our topic of this lecture series. Three different classes of feature vector separability measurement will be introduced in the lectures, namely, *divergence*,*Bhattachayya distance*,and *scatter matrics*.

2 Divergence

2.1 Difinition

The idea of *divergence* is somehow rooted on that of Bayes classifer. By studying why Bayesian classifer call tell the class of each tesing data based on a certain feature vector, we try to find out a separability measurement related to the feature vector.

Recall the Bayes rule introduced in the Chapter 2 of text book. Given two classes w_1 and w_2 and a feature vector x, we select w_1 if

$$P(w_1|x) > P(w_2|x) \tag{1}$$

or equivalently if

$$\frac{p(x|w_1)}{p(x|w_2)} \geq \frac{P(w_2)}{P(w_1)} \tag{2}$$

When given the sampling data of class w_1 and w_2, the value of $P(w_2)$ and $P(w_1)$ will be fixed. Therefore the value of $\frac{P(x|w_1)}{P(x|w_2)}$ conveys the underlying discriminating information of class w_1 with respect to w_2. Considering most of the time, $P(x|w_1)$ and $P(x|w_2)$ will be in a normal distribution form, a convenience way to express this information is by defining a function $D_{12}(x)$ as:

$$D_{12}(x) = ln\frac{p(x|w_1)}{p(x|w_2)} \tag{3}$$

Clearly, for completely overlapped classes we get $D_{12}(x) = 0$. Since X takes different values, it is natural to consider the mean value over the class w_1, that is,

$$D_{12} = \int_{-\infty}^{+\infty} p(x|w_1) ln \frac{p(x|w_1)}{p(x|w_2)}\, dx \tag{4}$$

Similarily, the same arguments hold for the w_2 and we define

$$D_{21} = \int_{-\infty}^{+\infty} p(x|w_2) ln \frac{p(x|w_2)}{p(x|w_1)}\, dx \tag{5}$$

Finally, the mutual separability measure between class w_1 and class w_2 can be expressed by *divergence*, which is given by the sum of D_{12} and D_{21} , e.g.

$$d_{12} = D_{12} + D_{21} \tag{6}$$

For multiclass problem, the *divergence* is computed for every class pair w_i, w_j :

$$\begin{aligned} d_{ij} &= D_{ij} + D_{ji} \\ &= \int_{-\infty}^{+\infty} (p(x|w_i) - p(x|w_j)) ln \frac{p(x|w_i)}{p(x|w_j)}\, dx \end{aligned}$$

and the average class separability can be computed using the avearge *divergence*

$$d = \sum_{i=1}^{M} \sum_{j=1}^{M} P(w_i)P(w_j)d_{ij} \tag{7}$$

2.2 Property of Divergence

Divergence has the following easily shown properties:

$$d_{ij} \geq 0$$

$$d_{ii} = 0$$

$$d_{ij} = d_{ji}$$

Furthermore, if the components of the feature vector are statistically independent, it can be shown that

$$\begin{aligned} d_{ij}(x) &= d_{ij}(x_1, x_2, \ldots, x_l) = \sum_{r=1}^{l} d_{ij}(x_r) \\ &= \sum_{r=1}^{l} \int_{-\infty}^{+\infty} (p(x_r|w_i) - p(x_r|w_j)) ln \frac{p(x_r|w_i)}{p(x_r|w_j)}\, dx_r \end{aligned}$$

2.3 Divergence under Guassian Distribution

In general, it is not easy to computer the *divergence*. However, If the density functions $p(x|w_i)$ and $p(x|w_j)$ are Gaussians $N(\mu_i, \sigma_i)$ and $N(\mu_j, \sigma_j)$ respectively, the computation of the divergence is simplified and it is not difficult to show

$$d_{ij} = \frac{1}{2}\text{trace}\Sigma_i^{-1}\Sigma_j + \Sigma_j^{-1}\Sigma_i - 2I + \frac{1}{2}(\mu_i - \mu_m)^T(\Sigma_i^{-1} + \Sigma_j^{-1})(\mu_i - \mu_j) \tag{8}$$

Equation (8) becomes even simpler when one-dimensional features are concerned,

$$d_{ij} = \frac{1}{2}(\frac{\sigma_j^2}{\sigma_i^2} + \frac{\sigma_i^2}{\sigma_j^2} - 2) + \frac{1}{2}(\mu_i - \mu_j)^2(\frac{1}{\sigma_i^2} + \frac{1}{\sigma_j^2}) \tag{9}$$

Equation (9) gives more intuitions about the property of the divergence. The divergence of two class depends on not only the difference between the means of the two classes but also the their variances. That is to day, given to classes with equal mean values, the divergence defined on these two classes can still be large, provided their variances differ significantly. Thus class separation is still possible even if the means coincide. One example is given in the Fige 1.

Now consider the special situation when $\sigma_i = \sigma_j = \sigma$, we find the even simplified expression of d_{ij} as following:

$$d_{ij} = (\mu_i - \mu_j)^T\sigma^{-1}(\mu_i - \mu_j) \tag{10}$$

Notice now d_{ij} is nothing but the Mahalanobis distance between the corresponding mean vectors. Recall from the chapter 2 of text book, there is strong relationship between Mahalanobis distance d_m and Bayesian optimal error P_B, e.g.

$$P_B = \int_{(\frac{1}{2})d_m}^{+\infty} \frac{1}{\sqrt{2\pi}}\exp(-z^2/2)\,dz \tag{11}$$

The larger the distance of Mahalanobis, the smaller the optimal error of the Bayesian classifer (problem 2.9 of the text book gives the exact equation). Furthermore, given the above condition and equal probability of the two classes $p(w_i) = p(w_j)$, we know the optimal error can be achieved by using the minimal distance classifer. All those indicate that the divergence is really a quite good candidate for the class separability measures. Unfortunately, such a direct relation of the divergence with the Bayes error is not possible for more general distribution. Even worse, in their papers, Swai[1] and Rich[2] pointed out that the specific dependence of the divergency on the difference of the mean vectors in equation (8) may lead to misleading results, in the sense that small variation in the difference of the mean values can produce large changes in the divergence, which, however, are not reflected in eqation (11). Regarding to this, a variation of the divergence, called transformed divergence:

$$d'_{ij} = 2(1 - exp(-d_{ij}/8)) \tag{12}$$

is suggested. Effort has been made to search for a separability measures with a closer relationship to the Bayes error for more general cases, the following section will introduce one of those measure.

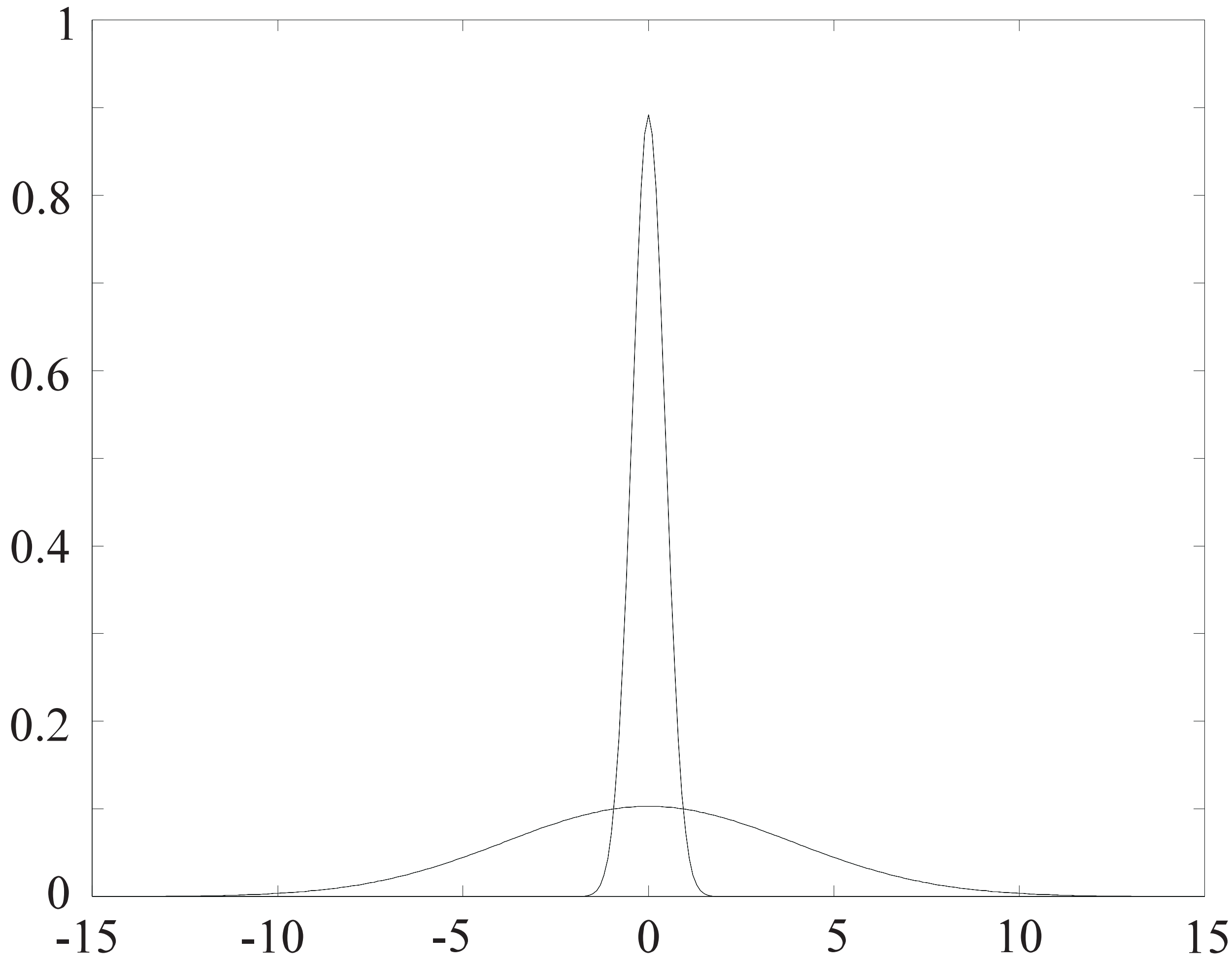

Figure 1: Two separable classes with same mean but different variances.

3 Chernoff Bound and Bhattacharyya distance

As mentioned in the end of previous section, we want to find a separability measure more closely related to Bayes minimum error. Recall from Chapter 1 of the text book, the minumun achievable error of a Bayes for two classes w_1, w_2 can be written as:

$$P_e = \int_{-\infty}^{+\infty} \min\left[P(w_i)p(x|w_i), p(w_j)p(x|w_j)\right]\, dx \tag{13}$$

In the general case, the analytic computation of the above integral is not possible. However, considering the following inequality:

$$\min[a,b] \leq a^s b^{1-s} \text{ for } a, b \geq 0, \text{ and } 0 \leq s \leq 1 \tag{14}$$

an upper bound of eqation (13) can be derived, which is:

$$P_e \le P(w_i)^s P(w_j)^{1-s} \int_{-\infty}^{+\infty} p(x|w_i)^s p(x|w_j)^{1-s}\, dx \equiv \epsilon_{CB} \tag{15}$$

ϵ_{CB} is known as the Chernoff bound. Now our task to find the minimal error is transformed to find the minimal ϵ_{CB} with regarding to the value of s. Such a bound is the Chernoff bound for general s. However, to simplfy the computation, let's restrict the value of s to be 1/2. Under this restriction, equation (15) becomes as:

$$P_e \le sqrtP(w_i)P(w_j) \int_{-\infty}^{+\infty} sqrtp(x|w_i)p(x|w_j)\, dx \equiv \epsilon_{CB} \tag{16}$$

If the pdf's of the two class follow Guassian distribution $N(\mu_i, \sigma_i)$ and $N(\mu_j, \sigma_j)$, by a little bit of algebra the integral in the equation (16) can be removed. As a result, the Chernoff bound (with s=1/2) can be expressed as:

$$\epsilon_C B = sqrtP(w_i)P(w_j)exp(-B_{ij})\text{where}B_{ij} = \frac{1}{8}(\mu_i - \mu_j)^T \left(\frac{\sigma_i + \sigma_j}{2}\right)^{-1} (\mu_i - \mu_j) + \frac{1}{2} ln \frac{\left|\frac{\sigma_i + \sigma_j}{2}\right|}{\sqrt{|\sigma_i|\,|\sigma_j|}} \tag{17}$$

B_{ij} is known as the Bhattacharyya distance. Like *divergence*, the value of B_{ij} depends on both the mean distance of the two classes and the covariance difference. Given two class of data samples, $p(w_i)$ and $p(w_j)$ are independent of feature subsets used, therefore, the value of $\epsilon_C B$ will decrease when the Bhattacharyya distance value increases. That is to say, we get the smaller minimal error bound when using larger B_{ij} value. Therefore, Bhattacharyya distance is also a good candidate for class separability measures.

It can be shown (problem 5.9 of the text book) that if $\sigma_i = \sigma_j$, the Bhattacharyya distance B_{ij} corresponds to the optimal Chernoff bound. In fact, it is equal to the 1/8 times Mahalanobis distance.

Although Bhattacharyya distance is based on the Gaussian distribution, empirically, the Bhattacharyya distance is still a useful class separability measures when the distribution of data is not in Gaussian form.

4 Scatter Matrices

When the data distribution $p(x|w_i)$ and $p(x|w_j)$ are not Guassian, both divergence d_{ij} and Bhattacharyya distance B_{ij} are difficult to be computed. In this sections, we should introduce a set of simpler measures, built upon information related to the way feature vector samples are scattered in the l-dimensional feature vector space. Intuitively, we want to find a set of $l - D$ feature vectors, based on which, if we draw the scatter graph of the data in the l dimension, we will find smaller intra-class distance, and larger inter-class distance. To this end, we should first define 3 matrix to quantify those distances, and from where we can define our third class separability measure.

4.1 Within-class scatter matrix

Let define a within-class scatter matrix as

$$
\begin{aligned}
S_w &= \sum_{i=1}^{M} P(w_i)\sigma_i \\
&= \sum_{i=1}^{M} P(w_i)E\left[(x-\mu_i)(x-\mu_j)^T\right]
\end{aligned}
$$

If the dimension of the feature vector is l, then S_i is a $l * l$ matrix. Obviously, trace $\{S_i\}$ is a measure of the average, over all classes, variance of the features, which gives us a measure of how close sample data in each class are clustered.

4.2 Between-class scatter matrix

The between-class matrix is defined as following:

$$
S_b = \sum_{i=1}^{M} P(w_i)(\mu_i - \mu_o)(\mu_i - \mu_o)^T \tag{18}
$$

where, μ_o is the global mean of all the classes, which is defined as

$$
\mu_o = \sum_{i=1}^{M} P(w_i)\mu_i
$$

It is easily to see that trace $\{S_b\}$ is a measure of the average (over all classes) distance of the mean of each individual class from the global mean.

4.3 Mix-class scatter matrix

Compared to the above two matrices, the mix-class scatter matrix is somehow a comprehensive one, which is defined to be the covariance matrix of the feature vector with respect to the global mean:

$$
S_m = E\left[(x-\mu_0)(x-\mu_0)^T\right] \tag{19}
$$

Trace $\{S_m\}$ is the sum of the variance of the features around the global mean. It is not difficult to show (Problem 5.10 of text book) that

$$
S_m = S_w + S_b
$$

4.4 Distance measure based on scatter matrices

Given the difinition of the above 3 scatter matrices, it is straightforward to see that a new set of class separability measure can be defined.

First, let's define

$$
J_1 = \frac{\mathrm{trace}\{S_m\}}{\mathrm{trace}\{S_w\}} \tag{20}
$$

J_1 meets our intuition for a good class separability measure in a way that J_1 will take large value when smaples in the l-dimensional space are well clustered around their class means, and the clusters of the different classes are well separated. An alternative definition of J_1 is obtained by replacing S_m with S_b in equation (20).

Notice that the scatter matrices S_w S_m are symmetric positive definite and thus their eigenvalues are positive. Therefore, the trace of this matrix is equal to the sum of the eigenvalue, while the determinant is equal to their product. Considering this fact, we can define a new class separability measure by:

$$J_2 = \frac{|S_m|}{|S_w|} = |S_w^{-1} S_m| \tag{21}$$

A variant of J_2 commonly encounted in practec is:

$$J_3 = \text{trace}\{S_w^{-1} S_m\} \tag{22}$$

Fig 2 shows graphically the relations between the magitude of those scatter matrices and the scatter pattern (and degree) of a 3-class sample data with regard to different 2_D feature vectors. It is clear from Fig 2, that largest J_3 value corresponds to the case of distant well-clustered classes and the smalles J_3 value corresponds to the case of closely located classes with large within-class variance.

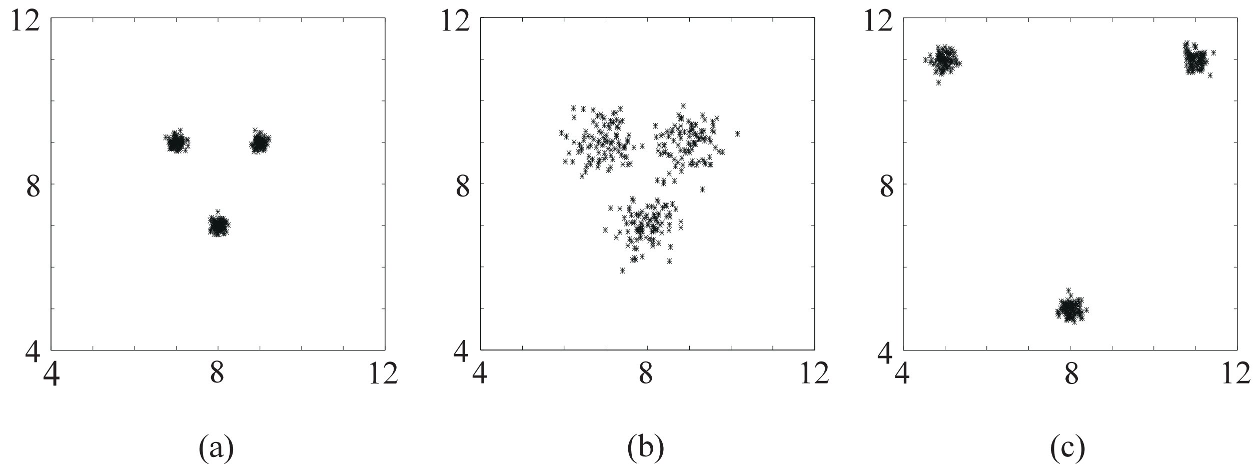

Figure 2: [Classes with (a) small s_w and small S_b, with $J_3 = 164.74$ (b) large S_w and S_b with $J_3 = 12.5$ and (c) small S_w and large S_b, with $J_3 = 620.9$]

For one-dimensional two-class problem, if $P(w_1) = P(w_2)$, then

$$|S_w| = \frac{1}{2}(\sigma_1^2 + \sigma_2^2)$$

$$|S_b| = \frac{1}{4}(\mu_1 - \mu_2)^2$$

We find So-called *Fisher's discriminat retion*, which is propotional to the $|S_b|/|S_w|$ can be defined as:

$$FDR = \frac{(\mu_1 - \mu_2)^2}{\sigma_1^2 + \sigma_2^2}$$

FDR is sometimes used to quantify the separability capability of individual features. FDR takes the similar form as the square of z test statistic, however, in z test, we try to use z value to infer the confident probability of the correntness of our hypothesis, here we use the FDR value to indict the separability of the classes based on a certain feature vector. For the multiclass case, averaging forms of FDR can be used. One possibility is

$$FDR' = \sum_{i}^{M} \sum_{j \neq i}^{M} \frac{(\mu_i - mu_j)^2}{\sigma_i^2 + \sigma_j^2}$$

5 Summary

In this transcription, we discussed how to measure the discriminaton effectiveness of feature vectors. Three classes of separability measures were suggested based on *Divergence*, *Bhattachayya distance*, and *Scatter Matries.* Separability measures based on *Divergence* and *Bhattachayya distanc* have closer relationship with the Bayesian optimal error, but are difficult to compute if the distribution $p(x|w_i)$ and $p(x|w_j)$ of the sample data in the feature vector space are not in the normal form. On the other hand, separability measures based on the Scatter Matrices don't have clear relationship with the Bayesian optimal error, yet, they are easy to be computed.

6 References

[1] Swai P. H., King R. C., "Two effective feature selection criteria for multispect remote sensing" Proceeding of the 1st internatonal conference on Pattern Recognization, pp. 536-540, 1973. [2] Richards J., Remote sensing digital image analysis, 2nd ed., Springer-Verlag, 1995.