Classification Methods II: Linear and Quadratic Discrimminant Analysis

Rebecca C. Steorts, Duke University

STA 325, Chapter 4 ISL

Agenda

- Linear Discriminant Analysis (LDA)

Classification

- Recall that linear regression assumes the responses is quantitative
- In many cases, the response is qualitative (categorical).

Here, we study approaches for predicting qualitative responses, a process that is known as classification.

Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning the observation to a category, or class.

We have already covered two such methods, namely, KNN regression and logistic regression.

Setup

We have set of training observations $(x_1, y_1), \ldots, (x_n, y_n)$ that we can use to build a classifier.

We want our classifier to perform well on both the training and the test data.

Linear Discrimminant Analysis (LDA)

- Logistic regression involves directly modeling $Pr(Y = k|X = x)$ using the logistic function, given by (4.7) for the case of two response classes.
- We model the conditional distribution of the response Y, given the predictor(s) X.
    - We now consider an alternative and less direct approach to estimating these probabilities.
    - We model the distribution of the predictors X separately in each of the response classes (i.e. given Y), and then use Bayes theorem to flip these around into estimates for $Pr(Y = k|X = x)$.
    - When these distributions are assumed to be normal, it turns out that the model is very similar in form to logistic regression.

Why do we need another method, when we have logistic regression?

LDA

- When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable.
- Linear discriminant analysis (LDA) does not suffer from this problem.
- If n is small and the distribution of the predictors X is approximately normal in each of the classes, the linear discriminant model is again more stable than the logistic regression model.
- Linear discriminant analysis is popular when we have more than two response classes.

Using Bayes' Theorem for Classification

- Wish to classify an observation into one of K classes, $K \geq 2$.
  - Let π_k: overall or prior probability that a randomly chosen observation comes from the kth class.
  - Let $f_k(X) = Pr(X = x | Y = k)$: the density function of X for an observation that comes from the kth class.
    - $f_k(X)$ is relatively large if there is a high probability that an observation in the kth class has $X \approx x$.
    - $f_k(X)$ is small if it is very unlikely that an observation in the kth class has $X \approx x$.
- Then **Bayes theorem** states that

$$P(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{i=1}^{K} \pi_i f_i(x)}. \tag{1}$$

Using Bayes' Theorem for Classification

- Let $p_k(X) = Pr(Y = k|X)$.
- Instead of directly computing $p_k(X)$, we can simply plug in estimates of π_k and $f_k(X)$ into (1).
- Estimating π_k is easy if we have a random sample of Ys from the population.
  - We simply compute the fraction of the training observations that belong to the kth class.
  - However, estimating $f_k(X)$ tends to be more challenging, unless we assume some simple forms for these densities.

Bayes Classifier

- A Bayes classifier classifies an observation to the class for which $p_k(x)$ is largest and has the lowest possible error rate out of all classifiers.
- (This is of course only true if the terms in (1) are all correctly specified.)
- Therefore, if we can find a way to estimate $f_k(X)$, then we can develop a classifier that approximates the Bayes classifier.
- Such an approach is the topic of the following sections.

Intro to LDA

- Assume that $p = 1$, that is, we only have one predictor.
- We would like to obtain an estimate for $f_k(x)$ that we can plug into (1) in order to estimate $p_k(x)$.
- We will then classify an observation to the class for which $p_k(x)$ is greatest.
    - In order to estimate $f_k(x)$, we will first make some assumptions about its form.
    - Assume that $f_k(x)$ is normal or Gaussian with mean μ_k and variance σ_k^2.
    - Assume that $\sigma_1^2 = \cdots = \sigma_k^2 =: \sigma^2$, meaning there is a shared variance term across all K classes.

The normal (Gaussian) density

Recall that the normal density is

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\{\frac{1}{2\sigma^2}(x - \mu_k)^2\} \tag{2}$$

Plugging (2) into (1) yields

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} \exp\{\frac{1}{2\sigma^2}(x - \mu_k)^2\}}{\sum_{i=1}^{K} \pi_i \frac{1}{\sqrt{2\pi}\sigma} \exp\{\frac{1}{2\sigma^2}(x - \mu_i)^2\}}. \tag{3}$$

LDA (continued)

- The Bayes classifier involves assigning an observation X = x to the class for which (3) is largest.
- Taking the log of (3) and rearrange the terms.
- Then we can show that this is equivalent to assigning the observation to the class for which

$$\delta(x)_k = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k) \tag{4}$$

is largest (Exercise).

LDA (continued)

- For instance, if $K = 2$ and $\pi_1 = \pi_2$, then the Bayes classifier assigns an observation to class 1 if $2x(\mu_1 - \mu_2) > (\mu_1^2 - \mu_2^2)$. and to class 2 otherwise.
- In this case, the Bayes decision boundary corresponds to the point where

$$x = \frac{\mu_1^2 - \mu_2^2}{2(\mu_1 - \mu_2)} = \frac{(\mu_1 - \mu_2)(\mu_1 + \mu_2)}{2(\mu_1 - \mu_2)} = \frac{\mu_1 + \mu_2}{2}.$$

Bayes decision boundaries

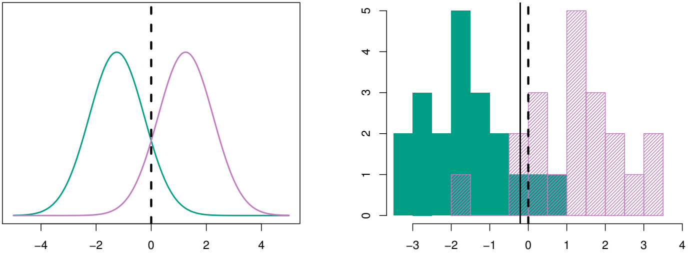


Figure 1: Two one-dimensional normal density functions are shown. The dashed vertical line represents the Bayes decision boundary. Right: 20 observations were drawn from each of the two classes, and are shown as histograms. The Bayes decision boundary is again shown as a dashed vertical line. The solid vertical line represents the LDA decision boundary estimated from the training data.

Bayes decision boundaries

- In practice, even if we are quite certain of our assumption that X is drawn from a Gaussian distribution within each class, we still have to estimate the parameters $\mu_1, \ldots, \mu_k; \pi_1, \ldots, \pi_K$, and σ^2.
- The linear discriminant analysis (LDA) method approximates the Bayes classifier by plugging estimates for π_k, π_k, and σ into (4.13).
- In particular, the following estimates are used:

Next part

$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i \tag{5}$$

$$\hat{\sigma}^2 = \frac{1}{n-K} \sum_{k=1}^{K} \sum_{i:y_i=k} (x_i - \hat{\mu}_k)^2, \tag{6}$$

- where n is the total number of training observations, and n_k is the number of training observations in the kth class.
- $\hat{\mu}_k$ is simply the average of all the training observations from the kth class
- σ^2 can be seen as a weighted average of the sample variances for each of the K classes.

Next part

- Sometimes we have knowledge of the class membership probabilities $\pi_1, \ldots, \pi_k$, which can be used directly.
- In the absence of any additional information, LDA estimates π_k using the proportion of the training observations that belong to the kth class. In other words,

$$\hat{\pi}_k = n_k/n. \tag{7}$$

LDA classifier

The LDA classifier plugs the estimates given in (6) and (7) into (8), and assigns an observation $X = x$ to the class for which

$$\delta_k(x) = x\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log(\hat{\pi}_k) \tag{8}$$

is largest.

The word linear in the classifierÕs name stems from the fact that the discriminant functions $\delta_k(x)$ in (8) are linear functions of x (as opposed to a more complex function of x).

LDA classifier ($p > 1$)

- We now extend the LDA classifier to the case of multiple predictors.
- To do this, we will assume that $X = (X_1, \dots X_p)$ is drawn from a multivariate Gaussian (or multivariate normal) distribution, with a class-specific mean vector and a common covariance matrix.

Iris application

- We motivate this section with a very well known dataset (introduced by R.A. Fisher and available in `R`).
- Illustrate the Iris data through naive linear regression, giving the multivariate form of LDA, and going through some extensions.

Iris application

Suppose that response categories are coded as an indicator variable. Suppose $\mathcal{G}$ has K classes, then $\boldsymbol{Y_1}$ is a vector of 0's and 1's indicating for example whether each person is in class 1.

- The indicator response matrix is defined as $Y = (\boldsymbol{Y_1}, \ldots, \boldsymbol{Y_K})$.
- Y is a matrix of 0's and 1's with each row having a single 1 indicating a person is in class k.
- The i^{th} person of interest has covariate values $\boldsymbol{x_{i1}}, \ldots, \boldsymbol{x_{ip}}$ that will be represented by $X_{N \times p}$.
- Our goal is to predict what class each observation is in given its covariate values.

Iris application

Let's proceed blindly and use a naive method of linear regression. We will do the following:

- Fit a linear regression to each column of Y.
- The coefficient matrix is $\hat{\beta} = (X'X)^{-1}X'Y$.
- $\hat{Y} = X(X'X)^{-1}X'Y$
- The k^{th} column of $\hat{\beta}$ contains the estimates corresponding to the linear regression coefficients that we get from regressing $\boldsymbol{X_1}, \ldots, \boldsymbol{X_p}$ onto $\boldsymbol{Y_K}$.

LDA

Look at $\hat{Y}$ corresponding to the indicator variable for each class k. Assign each person to the class for which $\hat{Y}$ is the largest. More formally stated, a new observation with covariate $\boldsymbol{x}$ is classified as follows:

- Compute the fitted output $\hat{\boldsymbol{Y}}_{\boldsymbol{new}}(\mathbf{X}) = [(1, \mathbf{X})'\hat{\beta}]'$.
- Identify the largest component of $\hat{\boldsymbol{Y}}_{\boldsymbol{new}}(\mathbf{X})$ and classify according to
$$\hat{G}(\mathbf{X}) = \arg\max_{k} \hat{\boldsymbol{Y}}_{\boldsymbol{new}}(\mathbf{X}).$$

We now pose the question, does this approach make sense?

LDA

- The regression line estimates $\mathrm{E}(Y_k|\boldsymbol{X} = \mathbf{X}) = \mathrm{P}(G = k|\boldsymbol{X} = \mathbf{X})$ so the method seems somewhat sensible at first.
- Although $\sum_k \hat{Y}_k(\mathbf{X}) = 1$ for any $\mathbf{X}$ as long as there is an intercept in the model (exercise), $\hat{Y}_k(\mathbf{X})$ can be negative or greater than 1 which is nonsensical to the initial problem statement.
- Worse problems can occur when classes are masked by others due to the rigid nature of the regression model.
- Illustrate "masking effect" with the Iris dataset.

What is masking?

When the number of classes $K \geq 3$, a class may be hidden or masked by others because there is no region in the feature space that is labeled as this class.

Iris data

- The Iris data (Fisher, Annals of Eugenics,1936) gives the measurements of sepal and petal length and width for 150 flowers using 3 species of iris (50 flowers per species).
  - The species considered are setosa, versicolor, and virginica.
  - To best illustrate the methods of classification, we considered how petal width and length predict the species of a flower.
  - We see the dataset partially represented in Table 1.

Sepal L	Sepal W	Petal L	Petal W	Species
5.1	3.5	1.4	0.2	setosa
4.9	3.0	1.4	0.2	setosa
4.7	3.2	1.3	0.2	setosa
⋮	⋮	⋮	⋮	⋮
7.0	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.9	3.1	4.9	1.5	versicolor
⋮	⋮	⋮	⋮	⋮
6.3	3.3	6.0	2.5	virginica
5.8	2.7	5.1	1.9	virginica
7.1	3.0	5.9	2.1	virginica

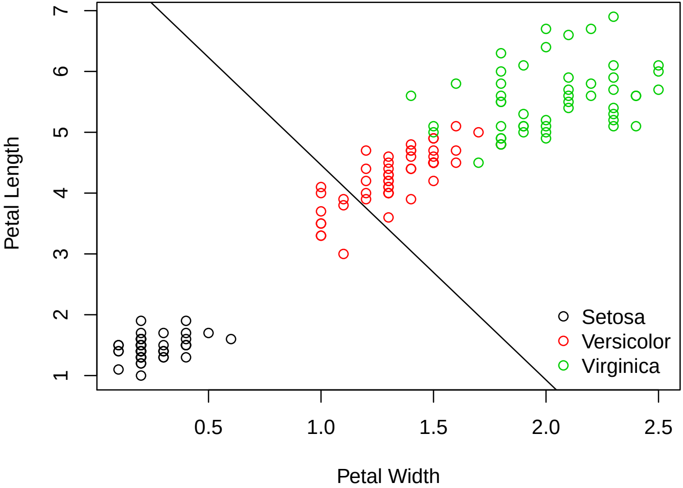
Table 1: Iris data with 3 species (setosa, versicolor, and virginica)

Iris data

We can can see that using linear regression (Figure 2 to predict for different classes can lead to a masking effect of one group or more. This occurs for the following reasons:

1. There is a plane that is high in the bottom left corner (setosa) and low in the top right corner (virginica).
2. There is a second plane that is high in the top right corner (virginica) but low in the bottom left corner (setosa).
3. The third plane is approximately flat since it tries to linearly fit a collection of points that is high in the middle (versicolor) and low on both ends.

Iris data


Petal Length
1
2
3
4
5
6
7
0.5
1.0
1.5
2.0
2.5
Petal Width
○ Setosa
○ Versicolor
○ Virginica

Iris data with LDA

We instead consider LDA, which has the following assumptions:

For each person, conditional on them being in class k, we assume

$$\boldsymbol{X}|G = k \sim N_p(\boldsymbol{\mu_k}, \Sigma).$$

That is,

$$f_k(\mathbf{X}) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{X} - \boldsymbol{\mu_k})'\Sigma^{-1}(\mathbf{X} - \boldsymbol{\mu_k})\right\}.$$

}

LDA assumes $\Sigma_k = \Sigma$ for all k (just as was true in the $p = 1$ setting).

In practice the parameters of the Gaussian distribution are unknown and must be estimated by:

- $\hat{\pi}_k = N_k/N$, where N_k is the number of people of class k
- $\hat{\boldsymbol{\mu}}_{\boldsymbol{k}} = \sum_{i:g_i=k} \boldsymbol{x_i}/N_k$
- $\hat{\Sigma} = \sum_{k=1}^{K} \sum_{i:g_i=k} (\boldsymbol{x_i} - \hat{\boldsymbol{\mu}}_{\boldsymbol{k}})(\boldsymbol{x_i} - \hat{\boldsymbol{\mu}}_{\boldsymbol{k}})'/(N - K),$

LDA Computation

We're interested in computing

$$P(G = k|\boldsymbol{X} = \mathbf{X}) = \frac{P(G = k, \boldsymbol{X} = \mathbf{X})}{P(\boldsymbol{X} = \mathbf{X})}$$

$$= \frac{P(\boldsymbol{X} = \mathbf{X}|G = k)P(G = k)}{\sum_{k=1}^{K} P(\boldsymbol{X} = \mathbf{X}, G = k)}$$

$$= \frac{f_k(\mathbf{X})\pi_k}{\sum_{j=1}^{K} f_j(\mathbf{X})\pi_j}.$$

We will compute $P(G = k|\boldsymbol{X} = \mathbf{X})$ for each class k.

Consider comparing $P(G = k_1|\boldsymbol{X} = \mathbf{X})$ and $P(G = k_2|\boldsymbol{X} = \mathbf{X})$.

LDA Computation

Then

$$\log\left[\frac{P(G=k_1|\mathbf{X}=\mathbf{X})}{P(G=k_2|\mathbf{X}=\mathbf{X})}\right]=\log\left[\frac{f_{k_1}(\mathbf{X})\pi_{k_1}}{f_{k_2}(\mathbf{X})\pi_{k_2}}\right]$$

$$=-\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_{\boldsymbol{k_1}})'\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_{\boldsymbol{k_1}})+\frac{1}{2}(\mathbf{X}-\boldsymbol{\mu}_{\boldsymbol{k_2}})'\Sigma^{-1}(\mathbf{X}-\boldsymbol{\mu}_{\boldsymbol{k_2}})+\log\left[\frac{\pi_{k_1}}{\pi_{k_2}}\right]$$

$$=(\boldsymbol{\mu}_{\boldsymbol{k_1}}-\boldsymbol{\mu}_{\boldsymbol{k_2}})'\Sigma^{-1}\mathbf{X}-\frac{1}{2}\boldsymbol{\mu}'_{\boldsymbol{k_1}}\Sigma^{-1}\boldsymbol{\mu}_{\boldsymbol{k_1}}+\frac{1}{2}\boldsymbol{\mu}'_{\boldsymbol{k_2}}\Sigma^{-1}\boldsymbol{\mu}_{\boldsymbol{k_2}}+\log\left[\frac{\pi_{k_1}}{\pi_{k_2}}\right]$$

Boundary Lines for LDA

- Now let's consider the boundary between predicting someone to be in class k_1 or class k_2.
- To be on the the boundary, we must decide what $\mathbf{X}$ would need to be if we think that a person is equally likely to be in class k_1 or k_2.

This reduces to solving

$$(\boldsymbol{\mu}_{\boldsymbol{k_1}} - \boldsymbol{\mu}_{\boldsymbol{k_2}})'\Sigma^{-1}\mathbf{X} - \frac{1}{2}\boldsymbol{\mu}'_{\boldsymbol{k_1}}\Sigma^{-1}\boldsymbol{\mu}_{\boldsymbol{k_1}} + \frac{1}{2}\boldsymbol{\mu}'_{\boldsymbol{k_2}}\Sigma^{-1}\boldsymbol{\mu}_{\boldsymbol{k_2}} + \log\left[\frac{\pi_{k_1}}{\pi_{k_2}}\right] = 0,$$

which is linear in $\mathbf{X}$.

Boundary Lines for LDA

- The boundary will be a line for two dimensional problems.
- The boundary will be a hyperplane for three dimensional problems.

The linear log-odds function implies that our decision boundary between classes k_1 and k_2 will be the set where

$$P(G = k_1 | \boldsymbol{X} = \mathbf{X}) = P(G = k_2 | \boldsymbol{X} = \mathbf{X}),$$

which is linear in $\mathbf{X}$. In p dimensions, this is a hyperplane.
} \frame{

We can then say that class k_1 is more likely than class k_2 if

$$P(G = k_1 | \boldsymbol{X} = \mathbf{X}) > P(G = k_2 | \boldsymbol{X} = \mathbf{X}) \implies$$

$$\log\left[\frac{P(G = k_1 | \boldsymbol{X} = \mathbf{X})}{P(G = k_2 | \boldsymbol{X} = \mathbf{X})}\right] > 0 \implies$$

$$(\boldsymbol{\mu}_{\boldsymbol{k_1}} - \boldsymbol{\mu}_{\boldsymbol{k_2}})'\Sigma^{-1}\mathbf{X} - \frac{1}{2}\boldsymbol{\mu}'_{\boldsymbol{k_1}}\Sigma^{-1}\boldsymbol{\mu}_{\boldsymbol{k_1}} + \frac{1}{2}\boldsymbol{\mu}'_{\boldsymbol{k_2}}\Sigma^{-1}\boldsymbol{\mu}_{\boldsymbol{k_2}} + \log\left[\frac{\pi_{k_1}}{\pi_{k_2}}\right] > 0 \implies$$

Linear Discrimminant Function

The linear discriminant function $\delta_k^L(\mathbf{X})$ is defined as

$$\delta_k^L(\mathbf{X}) = \boldsymbol{\mu_k'}\Sigma^{-1}\mathbf{X} - \boldsymbol{\mu_k'}\Sigma^{-1}\boldsymbol{\mu_k} + \log(\pi_k).$$

We can tell which class is more likely for a particular value of $\mathbf{X}$ by comparing the classes' linear discriminant functions.

Quadratic Discriminant Analysis (QDA)

We now introduce Quadratic Discriminant Analysis, which handles the following:

- If the Σ_k are not assumed to be equal, then convenient cancellations in our derivations earlier do not occur.
- The quadratic pieces in $\mathbf{X}$ end up remaining leading to quadratic discriminant functions (QDA).
- QDA is similar to LDA except a covariance matrix must be estimated for each class k.

QDA

The quadratic discriminant function $\delta_k^Q(\mathbf{X})$ is defined as

$$\delta_k^Q(\mathbf{X}) = -\frac{1}{2}\log|\Sigma_k| - \frac{1}{2}(\mathbf{X} - \boldsymbol{\mu_k})'\Sigma_k^{-1}(\mathbf{X} - \boldsymbol{\mu_k}) + \log(\pi_k).$$

LDA and QDA seem to be widely accepted due to a bias variance trade off that leads to stability of the models. That is, we want our model to have low variance, so we are willing to sacrifice some bias of a linear decision boundary in order for our model to be more stable.

LDA and QDA on Iris data

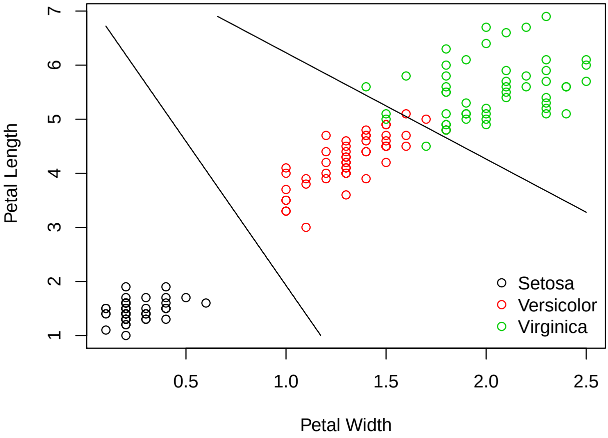


Figure 3: Clustering of Iris data using LDA.

LDA and QDA on Iris data

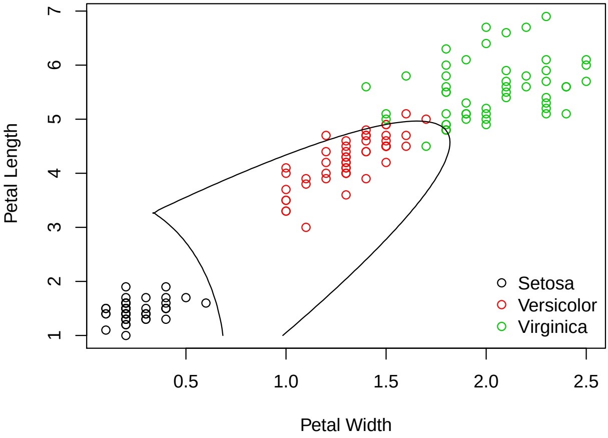


Figure 4: Clustering of Iris data using QDA.

Extensions of LDA and QDA

- We consider some extensions of LDA and QDA, namely Regularized Discriminant Analysis (RDA), proposed by Friedman (1989). This was proposed as compromise between LDA and QDA.
- This method says that we should shrink the covariance matrices of QDA toward a common covariance matrix as done in LDA.
- Regularized covariance matrices take the form

$$\hat{\Sigma}_k(\alpha) = \alpha\hat{\Sigma}_k + (1-\alpha)\hat{\Sigma}, \quad 0 \leq \alpha \leq 1.$$

- In practice, α is chosen based on performance of the model on validation data or by using cross-validation.

Extensions of LDA and QDA

- Fisher proposed an alternative derivation to dimension reduction in LDA that is equivalent to the ideas previously discussed.
- He suggested the proper way to rotate the coordinate axes was by maximizing the variance between classes relative to the variance within the classes.

Extensions

- Let $Z = \boldsymbol{a}'X$ and find the linear combo Z such that the between class variance is maximized wrt within class variance.
- Denote the covariance of the centroids by B.
- Denote the pooled within class covariance of the original data by W.
- BC $\text{Var}(Z) = \boldsymbol{a}'B\boldsymbol{a}$ and WC $\text{Var}(Z) = \boldsymbol{a}'W\boldsymbol{a}$.
- $B + W = T =$ total covariance matrix of X.

Extensions

Fisher's problem amounts to maximizing

$$\max_{\boldsymbol{a}} \frac{\boldsymbol{a}' B \boldsymbol{a}}{\mathbf{a}' W \boldsymbol{a}} \quad (\textit{exercise}).$$

The solution is $\boldsymbol{a} =$ largest eigenvalue of $W^{-1}B$.

Once we find the solution to maximization problem above, denoted by $\boldsymbol{a_1}$, we repeat the process again of maximization except this time the new maximum, $\boldsymbol{a_2}$, must be orthogonal to $\boldsymbol{a_1}$. This process continues and $\boldsymbol{a_k}$ are called the discriminant coordinates. In terms of what we did earlier, the $\boldsymbol{a_k}$'s are equivalent to the $\boldsymbol{x_k''}$'s.
Reduced-Rank LDA is desirable in high dimensions since it leads to a further dimension reduction in LDA.