Chapter 4: Linear Methods for Classification
Sections 4.1-4.3

Rebecca Steorts

Department of Statistics
University of Florida

Statistical Learning

February 1, 2010

Outline

General Setup

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.

Linear Regression of an Indicator Matrix

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

Linear Regression of an Indicator Matrix

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 $\hat{\boldsymbol{Y}}_{\boldsymbol{new}}(\boldsymbol{x}) = [(1, \boldsymbol{x})'\hat{B}]'$.
- Identify the largest component of $\hat{\boldsymbol{Y}}_{\boldsymbol{new}}(\boldsymbol{x})$ and classify according to

$$\hat{G}(\boldsymbol{x}) = \arg\max_{k} \hat{\boldsymbol{Y}}_{\boldsymbol{new}}(\boldsymbol{x}).$$

Does this Approach Make Sense?

- The regression line estimates $\mathrm{E}(Y_k|\boldsymbol{X} = \boldsymbol{x}) = \mathrm{P}(G = k|\boldsymbol{X} = \boldsymbol{x})$ so the method seems somewhat sensible at first.
- Although $\sum_k \hat{Y}_k(\boldsymbol{x}) = 1$ for any $\boldsymbol{x}$ as long as there is an intercept in the model (exercise), $\hat{Y}_k(\boldsymbol{x})$ 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.

Iris Data

- This data set (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.

Iris Data

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

Illustration of Masking Effect

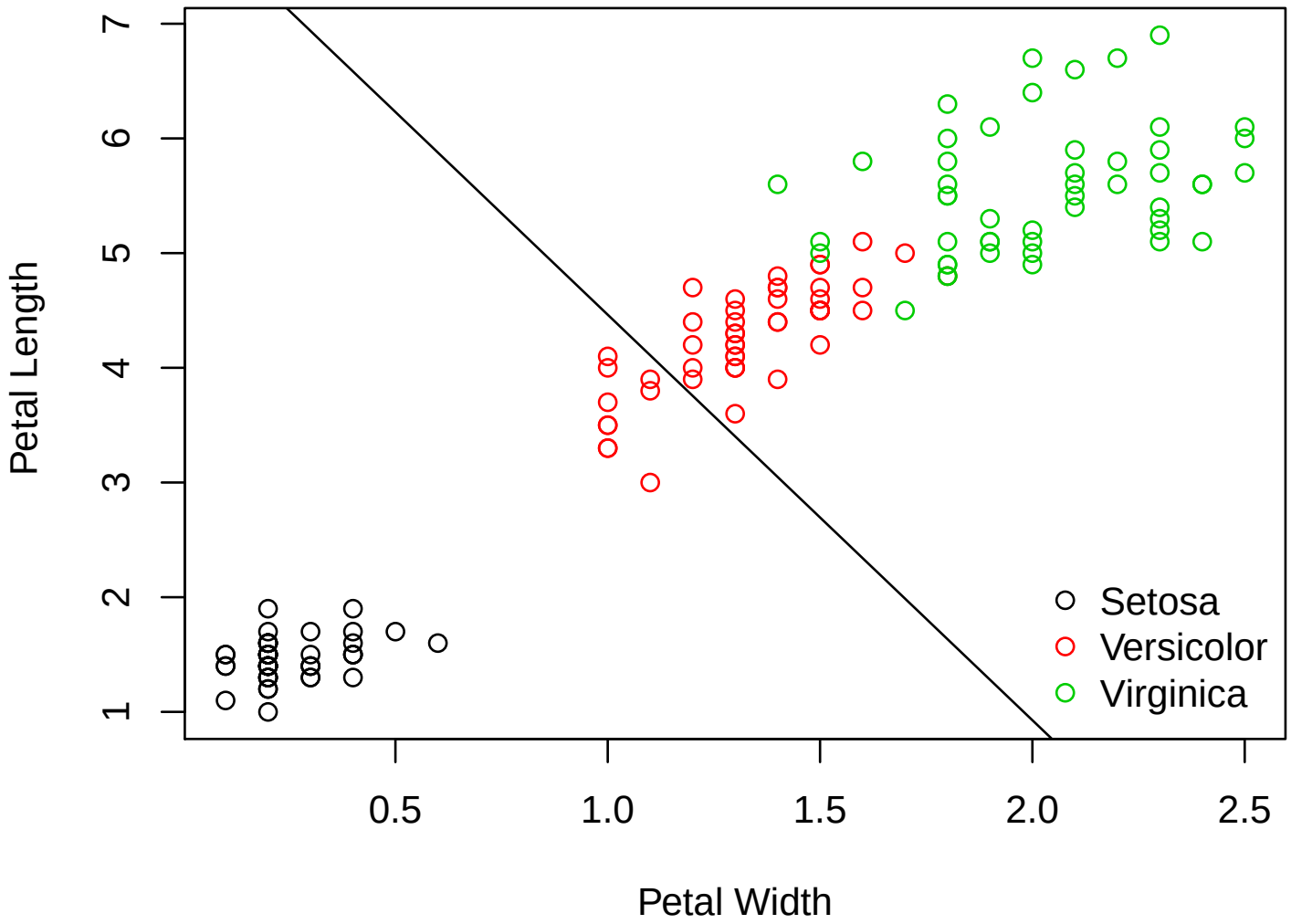


Illustration of Masking Effect

To recap the previous picture, we can see that using linear regression 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.

LDA Model Assumptions

For each person, conditional on them being in class k, we assume $\boldsymbol{X}|G = k \sim N_p(\boldsymbol{\mu_k}, \Sigma_k)$. That is,

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

Linear Discriminant Analysis (LDA) assumes $\Sigma_k = \Sigma$ for all k.

LDA Model Assumptions

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
- $\hat{\boldsymbol{\mu_k}} = \sum_{i:g_i=k} \boldsymbol{x_i}/N_k$
- $\hat{\Sigma} = \sum_{k=1}^{K} \sum_{i:g_i=k} (\boldsymbol{x_i} - \hat{\boldsymbol{\mu_k}})(\boldsymbol{x_i} - \hat{\boldsymbol{\mu_k}})'/(N-K)$,

where $\pi_k = P(G = k)$.

Derivations

We're interested in computing

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

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

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

Derivations

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

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

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

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

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

Derivations

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 $\boldsymbol{x}$ 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

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

which is linear in $\boldsymbol{x}$.

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

Derivations

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

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

which is linear in $\boldsymbol{x}$. In p dimensions, this is a hyperplane.

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

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

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

Derivations

$$(\mu_{k_1} - \mu_{k_2})' \Sigma^{-1} x - \frac{1}{2} \mu_{k_1}' \Sigma^{-1} \mu_{k_1} + \frac{1}{2} \mu_{k_2}' \Sigma^{-1} \mu_{k_2} + \log \left[\frac{\pi_{k_1}}{\pi_{k_2}} \right] > 0 \implies$$

$$\mu_{k_1}' \Sigma^{-1} x - \frac{1}{2} \mu_{k_1}' \Sigma^{-1} \mu_{k_1} + \log(\pi_{k_1}) > \mu_{k_2}' \Sigma^{-1} x - \frac{1}{2} \mu_{k_2}' \Sigma^{-1} \mu_{k_2} + \log(\pi_{k_2}).$$

Definition

The linear discriminant function $\delta_k^L(x)$ is defined as

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

We can tell which class is more likely for a particular value of x by comparing the classes' linear discriminant functions.

QDA

- If the Σ_k are not assumed to be equal, then convenient cancellations in our derivations earlier do not occur.
- The quadratic pieces in $\boldsymbol{x}$ 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.

Definition

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

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

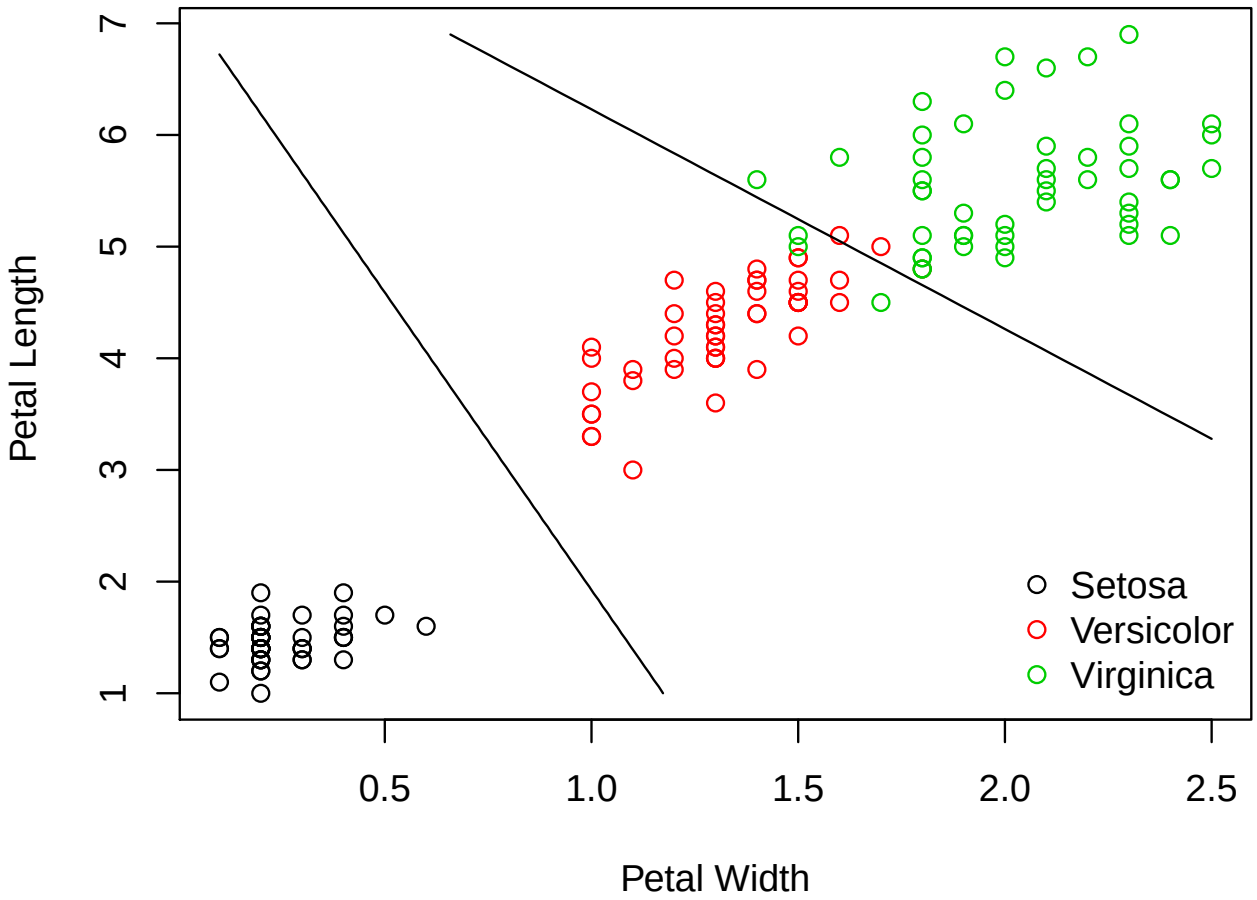
Properties of LDA and QDA

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.

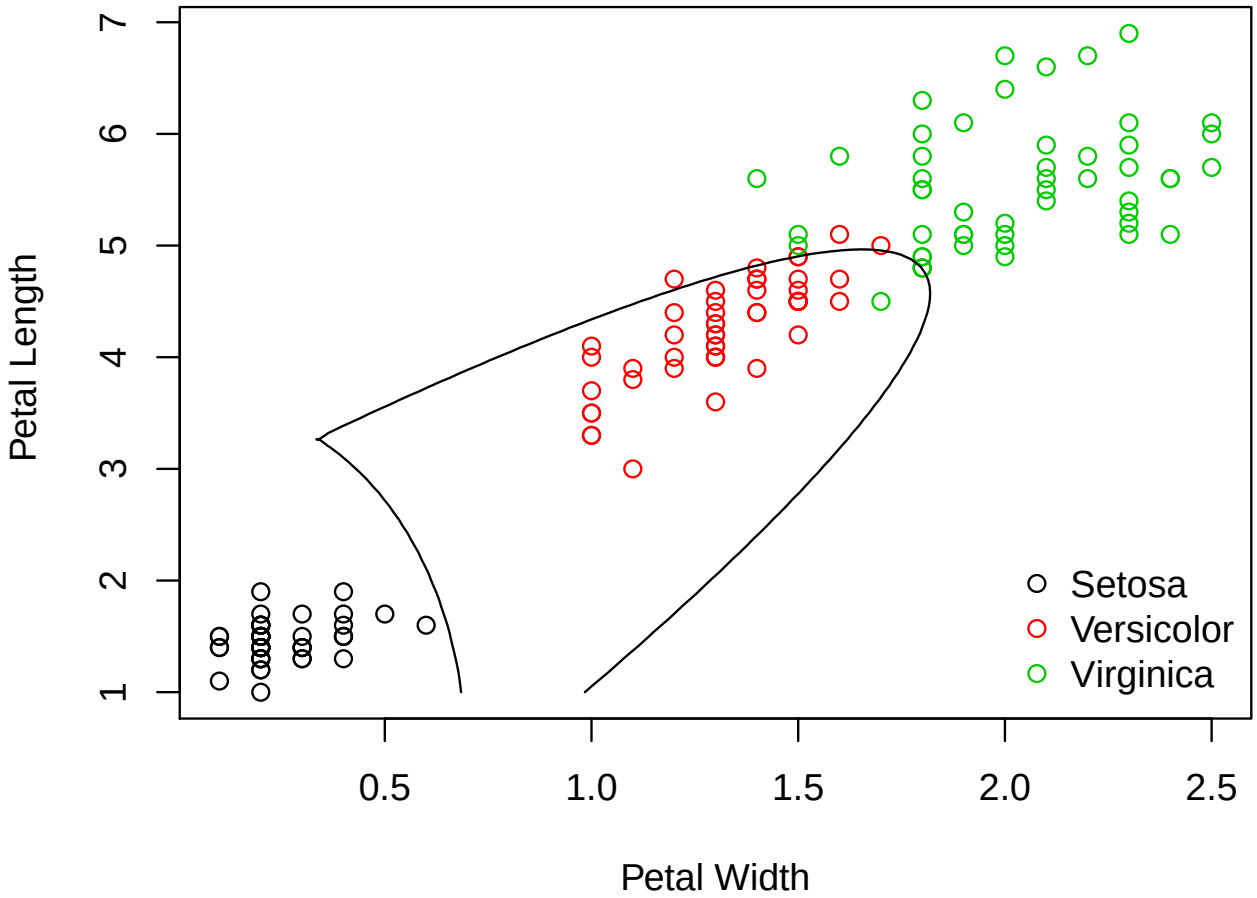
Iris Data Revisited

Returning to the iris data, we now consider predicting what classes each flower will be in using LDA. The following plot is obtained.



Iris Data Revisited

We next consider predicting what class each iris will be put into using QDA. The following plot is obtained.



Regularized Discriminant Analysis

- Friedman (1989) proposed a 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.

Geometric Interpretation of LDA

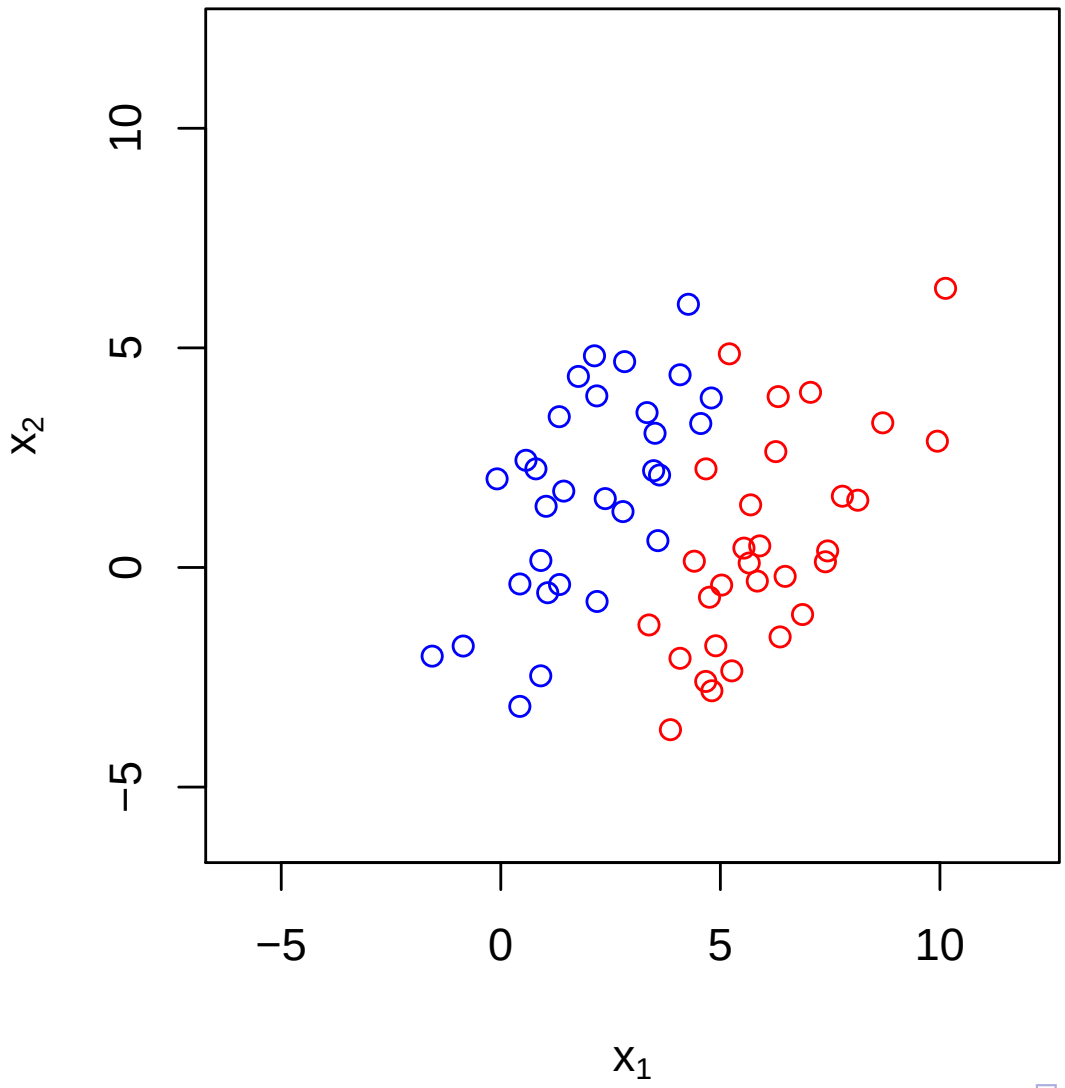
Recall K is the number of classes. Consider the centers of the K classes (centroids) which lie in a subspace of dimension K-1.

- We have a new observation, $\boldsymbol{x}_{new}$. For simplicity, consider p=2 and K=2.
- $\boldsymbol{x}_1''$ is defined to be the line that goes through the centroids after transforming the data to make it more spherical (ensures the line connecting the centroids is also the line that gives the greatest separation between the classes).
- $\boldsymbol{x}_2''$ is defined to be the line that is perpendicular to $\boldsymbol{x}_1''$.
- After the data are transformed we will denote the new axes by $\boldsymbol{x}_1'$ and $\boldsymbol{x}_2'$.

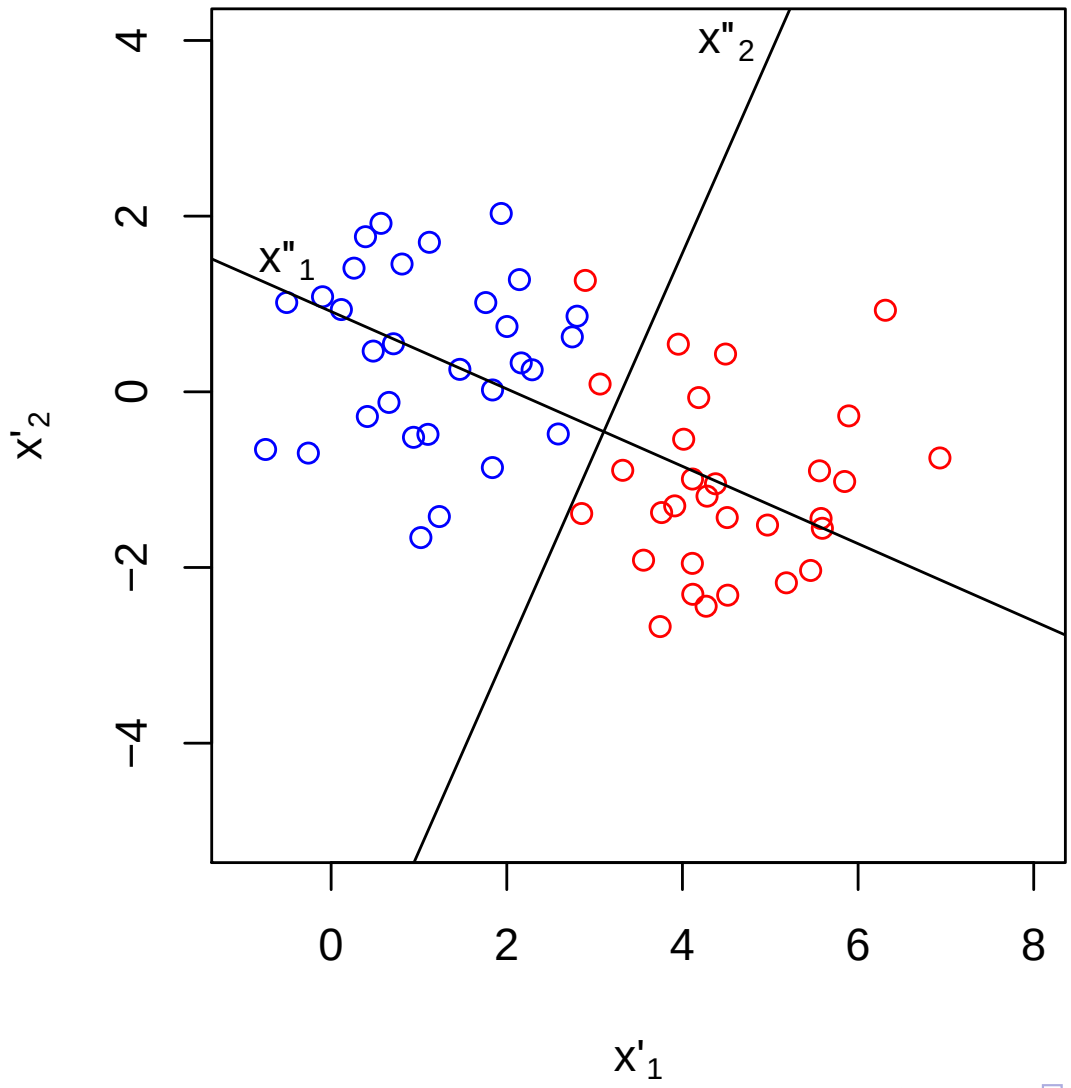
Procedure

1. Estimate the within class covariance matrix, $\hat{\Sigma} =: W$.
2. Transform the data and for each class by multiplying each data point by $W^{-1/2}$. Also, transform $\boldsymbol{x}_{new}$.
3. Plot $\boldsymbol{x}'$'s instead of $\boldsymbol{x}$'s to get new coordinate space.
4. To classify $\boldsymbol{x}_{new}$, classify this point to the class for which $\boldsymbol{x}_{new}$ is closest to that class's centroid.
5. $\boldsymbol{x}_1''$ is the line passing through the centroids and $\boldsymbol{x}_2''$ is the line perpendicular to $\boldsymbol{x}_1''$ in the p=2, K=2 case.

Two Dimensional Example



Two Dimensional Example



Fisher's Method

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.

- Let $Z = \boldsymbol{a}'X$ and find the l.c. 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 Var$(Z) = \boldsymbol{a}'B\boldsymbol{a}$ and WC Var$(Z) = \boldsymbol{a}'W\boldsymbol{a}$.
- $B + W = T$ = total covariance matrix of X.

Fisher's Method

Fisher's problem amounts to maximizing

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

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.

Summary

- Using linear regression to predict for indicator variables is a naive method that can lead to entire classes being masked.
- LDA, QDA, and Regularized Discriminant Analysis procedures avoid this problem by approaching the situation with a more sound theoretical justification.
- Reduced-Rank LDA is desirable in high dimensions since it leads to a further dimension reduction in LDA.