Multivariate Data Analysis

(MGT513, BAT531, TIM711)

Lecture 12

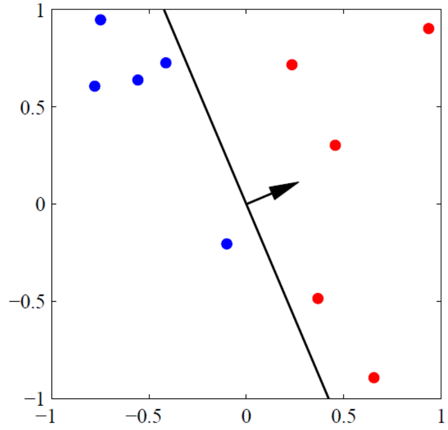
Ch.12 Discriminant Analysis

References

- LCG (textbook) Ch.12 Discriminant Analysis
- An Introduction to Statistical Learning with Applications in R (2nd Edition) by James et al: available online
- Lecture 15: Linear Discriminant Analysis (https://www.doc.ic.ac.uk/~dfg/ProbabilisticInference/IDAPISlides15.pdf)
- Linear and Quadratic Discriminant Analysis: Tutorial (https://arxiv.org/abs/1906.02590)
- Jonathan Taylor's Stats202 Lecture note: Not available anymore

Linear Classification

- Focus on linear classification model: the decision boundary is a linear function of x
  - Defined by (D-1)-dimensional hyperplane
- If the data can be separated exactly by linear decision surfaces, they are called linearly separable
- Implicit assumption: Classes can be modeled well by Gaussians
- Treat classification as a projection problem



From PRML (Bishop, 2006)

Discriminant Analysis

- Goal: To explain possible separation or discrimination between or among groups using independent variables
- Two approaches:
  - Fisher's Discriminant Analysis (FDA)
  - Linear Discriminant Analysis (LDA) / Quadratic Discriminant Analysis (QDA)

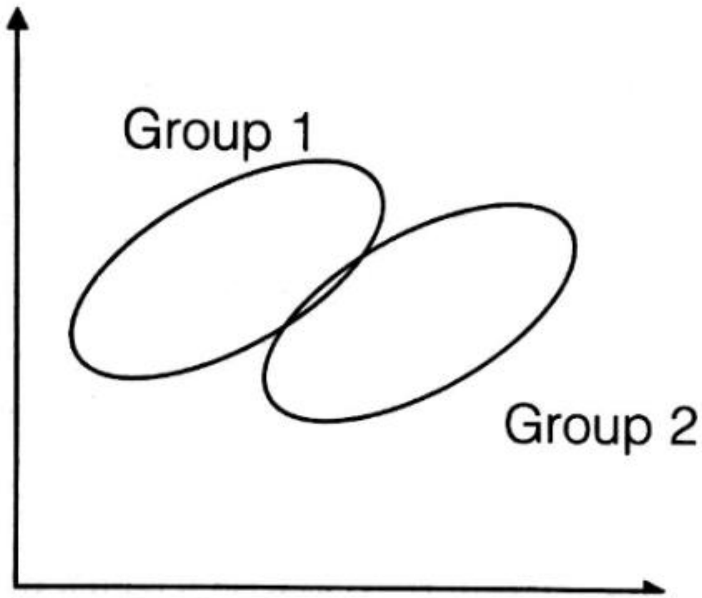
Fisher’s Discriminant Analysis

Fisher's discriminant Analysis (FDA)

- IDEA: Project input vector x to a one-dimensional subspace with basis vector w
- Assume we know the basis vector w, we can compute the projection of any point x onto the one-dimensional subspace spanned by w

Fisher’s discriminant Analysis (FDA)

FIGURE 12.1
Stylized scatter plot showing two groups



Fisher's discriminant Analysis (FDA)

FIGURE 12.2
Using X_1 to discriminate between groups 1 and 2

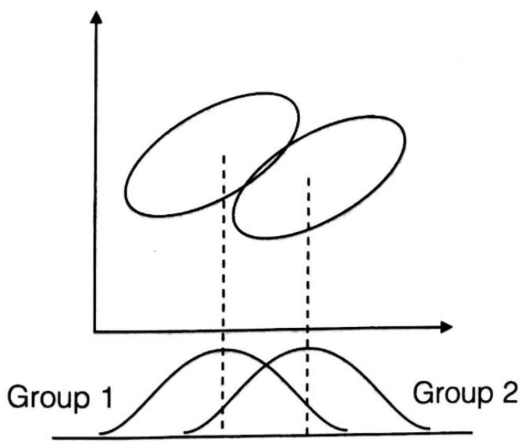
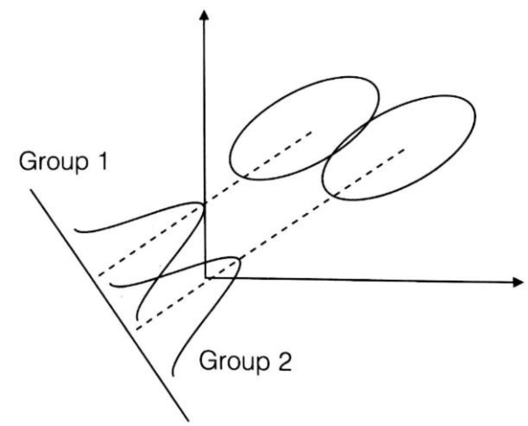
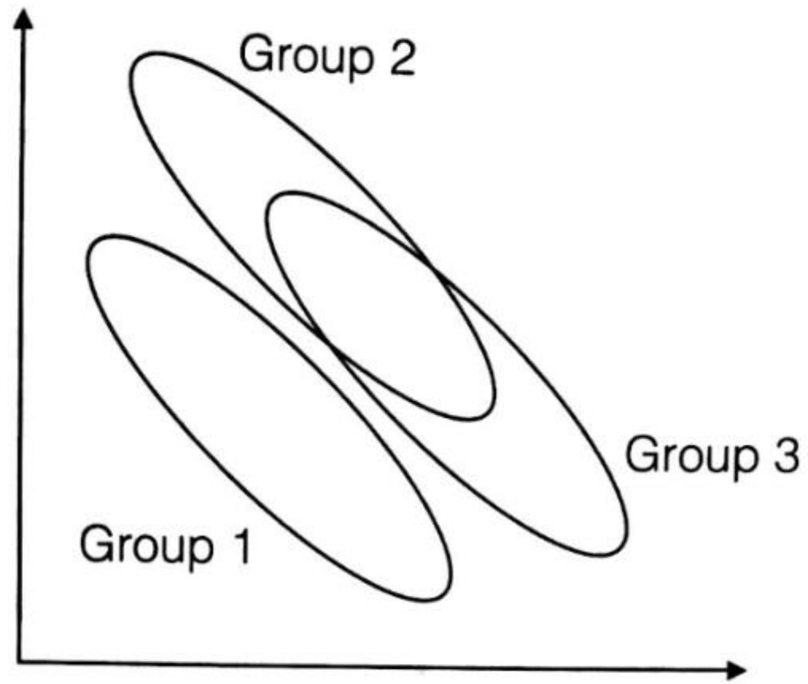


FIGURE 12.3
Using a linear combination of X_1 and X_2 to discriminate between groups 1 and 2



Fisher’s discriminant Analysis (FDA)

FIGURE 12.15
Stylized scatter plot for three-group discriminant analysis problem



Fisher's discriminant Analysis (FDA)

FIGURE 12.16
First discriminant function separates group 1 from groups 2 and 3

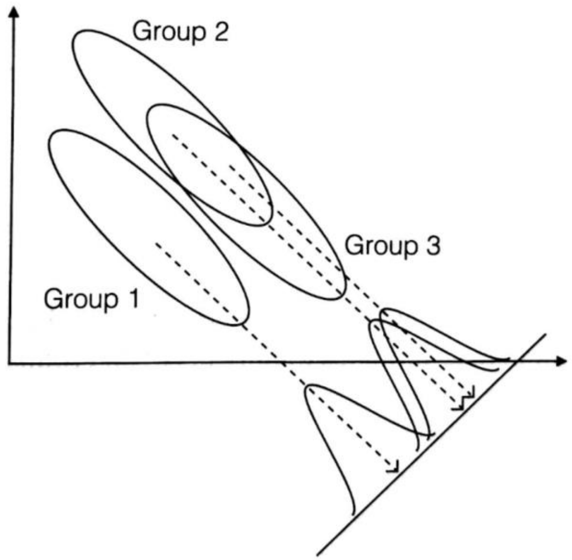
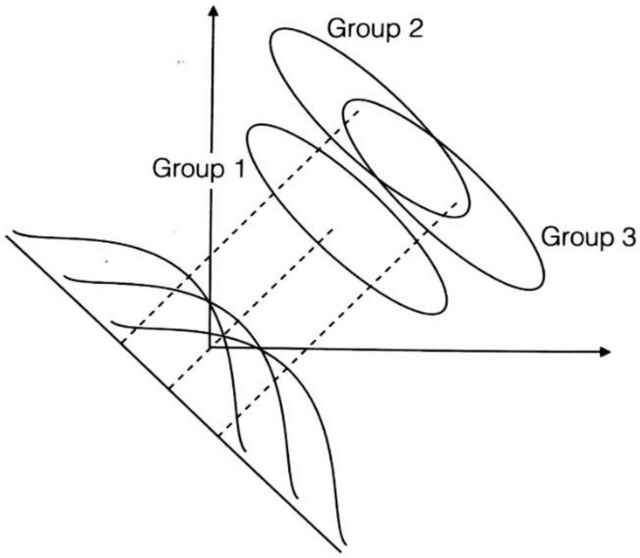


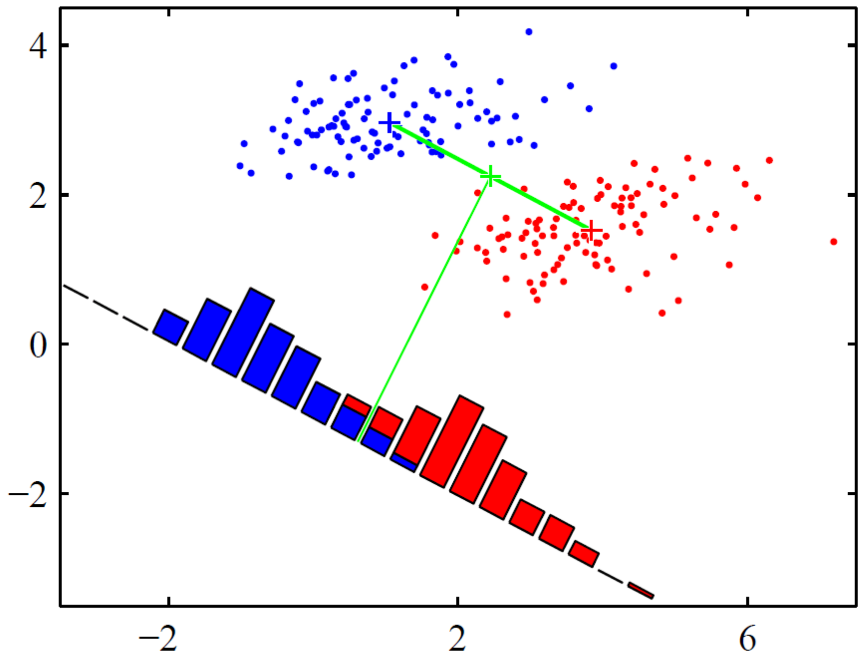
FIGURE 12.17
Second discriminant function separates group 2 from group 3



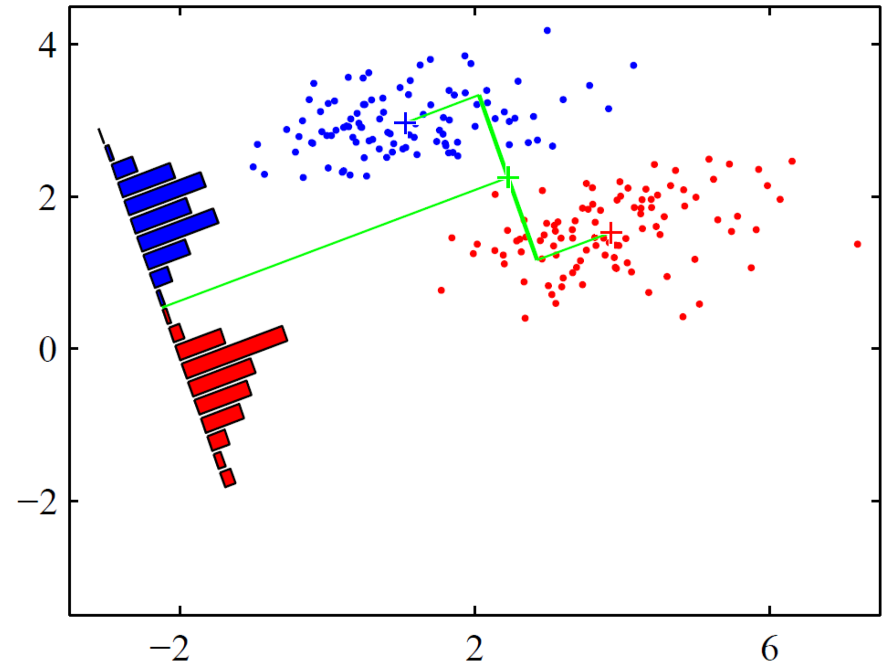
Fisher's discriminant Analysis (FDA)

- Adjust components of basis vector *w*

=> Select projection that maximizes the class separation



From PRML (Bishop, 2006)



From PRML (Bishop, 2006)

Fisher's discriminant Analysis (FDA)

K classes case: class $k(Y = k)$ with n_k observations

- Mean vector for each class k:

$$\mu_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

- Covariance matrix for each class k:

$$\Sigma_k = \frac{1}{n_k} \sum_{i:y_i=k} (x_i - \mu_k)(x_i - \mu_k)^T$$

Fisher's discriminant Analysis (FDA)

- Goal: To find the linear combination w to maximize the Fisher criterion

$$f = \frac{\text{Between} - \text{class sum of squares of the discriminant scores}}{\text{Within} - \text{class sum of squares of the discriminant scores}}$$

$$f(w) = \frac{w^T SS_B w}{w^T SS_W w}$$

$$SS_W = \sum_{k=1}^{K} \Sigma_k$$

$$SS_B = \sum_{k=1}^{K} (\mu_k - \bar{\mu})(\mu_k - \bar{\mu})^T$$

where $\bar{\mu} = \frac{1}{K} \sum_{k=1}^{K} \mu_k$

Fisher's discriminant Analysis (FDA)

$$w^* = \underset{w}{\operatorname{argmax}} \frac{w^T SS_B w}{w^T SS_W w}$$

We find w by setting $\frac{df}{dw} = 0$

$$\frac{df}{dw} = 0 \Leftrightarrow (w^T SS_W w) SS_B w - (w^T SS_B w) SS_W w = 0$$

$$\Leftrightarrow SS_B w - f SS_W w = 0$$

$$\Leftrightarrow SS_B w = f SS_W w$$

$$\Leftrightarrow {SS_W}^{-1} SS_B w = f w$$

This is an eigenvalue problem.

The projection vector is the eigenvector of ${SS_W}^{-1} SS_B$.

$$w \propto {SS_W}^{-1} SS_B$$

Linear Discriminant Analysis

Using Bayes' Theorem for Classification

Instead of estimating $P(Y|X)$, we will estimate:

- $P(X|Y)$: Given the response, what is the distribution of the inputs.
- $P(Y)$: How likely are each of the classes.

Then, we use Bayes rule to obtain the estimate:

$$P(Y = k|X = x) = \frac{P(X = x|Y = k)P(Y = k)}{P(X = x)}$$

$$= \frac{P(X = x|Y = k)P(Y = k)}{\sum_j P(X = x|Y = j)P(Y = j)}$$

Using Bayes' Theorem for Classification

Let

- $P(Y = k) = \pi_k$
- $P(X = x|Y = k) = f_k(x)$ follows a multivariate normal distribution:

$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}[(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)]}$$

- μ_k: Mean of the inputs for class k
- Σ: Covariance matrix common to all classes

Using Bayes' Theorem for Classification

By Bayes rule, the probability of class k, given the input x is:

$$P(Y = k|X = x) = \frac{f_k(x)\pi_k}{P(X = x)}$$

The denominator does not depend on the response k, so we can write it as a constant:

$$P(Y = k|X = x) = c_1 f_k(x)\pi_k$$

$$P(Y = k|X = x) = \frac{c_1 \pi_k}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}[(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)]}$$

Using Bayes' Theorem for Classification

Absorb everything that does not depend on k into a constant c_2 :

$$P(Y = k|X = x) = c_2 \pi_k e^{-\frac{1}{2}[(x-\mu_k)^T \Sigma^{-1} (x-\mu_k)]}$$

Take log of both sides:

$$\ln[P(Y = k|X = x)]$$
$$= \ln(c_2) + \ln(\pi_k) - \frac{1}{2}[(x - \mu_k)^T \Sigma^{-1} (x - \mu_k)]$$

So we want to find the maximum of this over k.

LDA has linear decision boundaries

Goal, maximize the following over k:

$$\ln(\pi_k) - \frac{1}{2}[(x-\mu_k)^T\Sigma^{-1}(x-\mu_k)]$$

$$= \ln(\pi_k) - \frac{1}{2}[x^T\Sigma^{-1}x + \mu_k{}^T\Sigma^{-1}\mu_k] + x^T\Sigma^{-1}\mu_k$$

$$= c_3 + \ln(\pi_k) - \frac{1}{2}\mu_k{}^T\Sigma^{-1}\mu_k + x^T\Sigma^{-1}\mu_k$$

Objective function (linear discriminant function):

$$\delta_k(x) = \ln(\pi_k) - \frac{1}{2}\mu_k{}^T\Sigma^{-1}\mu_k + x^T\Sigma^{-1}\mu_k$$

At an input x, we predict the response with the highest $\delta_k(x)$.

LDA has linear decision boundaries

Decision boundary

$$\delta_k(x) = \delta_l(x)$$

$$\ln(\pi_k) - \frac{1}{2}{\mu_k}^T\Sigma^{-1}\mu_k + x^T\Sigma^{-1}\mu_k$$
$$= \ln(\pi_l) - \frac{1}{2}{\mu_l}^T\Sigma^{-1}\mu_l + x^T\Sigma^{-1}\mu_l$$

This equation is a linear function of x

The locus of x by LDA is the set of all points x *perpendicular* to w, Fisher's discriminant function coefficients.

Parameter estimation

- Estimating π_k
  - proportion of the training observations that belong to the kth class

$$\hat{\pi}_k = \frac{n_k}{n}$$

- Estimating μ_k
  - average of training observations in the kth class

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

Parameter estimation

- Estimating Σ
  - weighted average of the sample covariance matrices for each of the k classes.

$$\hat{\Sigma} = \frac{1}{n-K}\sum_{k=1}^{K}\sum_{i:y_i=k}(x_i - \hat{\mu}_k)(x_i - \hat{\mu}_k)^T$$

LDA prediction

- For an input x, predict the class with the largest:

$$\hat{\delta}_k(x) = \ln(\hat{\pi}_k) - \frac{1}{2}\hat{\mu}_k{}^T\hat{\Sigma}^{-1}\hat{\mu}_k + x^T\hat{\Sigma}^{-1}\hat{\mu}_k$$

The decision boundaries

$$\ln(\hat{\pi}_k) - \frac{1}{2}\hat{\mu}_k{}^T\hat{\Sigma}^{-1}\hat{\mu}_k + x^T\hat{\Sigma}^{-1}\hat{\mu}_k$$
$$= \ln(\hat{\pi}_l) - \frac{1}{2}\hat{\mu}_l{}^T\hat{\Sigma}^{-1}\hat{\mu}_l + x^T\hat{\Sigma}^{-1}\hat{\mu}_l$$

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

FDA = LDA

- TWO classes case: class 1 ($Y = 1$) with n_1 observations and class 2 ($Y = 2$) with n_2 observations

In FDA,

$$f(w) = \frac{w^T SS_B w}{w^T SS_W w} = \frac{w^T(\mu_1 - \bar{\mu})(\mu_2 - \bar{\mu})^T w}{w^T(\Sigma_1 + \Sigma_2)w} = \frac{(w^T(\mu_2 - \mu_1))^2}{w^T(\Sigma_1 + \Sigma_2)w}$$

$$\frac{df}{dw} = 0 \Leftrightarrow (\mu_2 - \mu_1)^2 w = f(\Sigma_1 + \Sigma_2)w$$

$$w \propto (\Sigma_1 + \Sigma_2)^{-1}(\mu_2 - \mu_1)^2$$

If the equality of covariance matrices is assumed (as in LDA)

$$w \propto (2\Sigma)^{-1}(\mu_2 - \mu_1)^2 \propto \Sigma^{-1}(\mu_2 - \mu_1)^2$$

$$w^T x \propto (\Sigma^{-1}(\mu_2 - \mu_1)^2)^T x$$

FDA = LDA

In LDA,

$$\Sigma_1 = \Sigma_2 = \Sigma$$

$$\frac{c_1\pi_1}{(2\pi)^{p/2}|\Sigma|^{1/2}}\,e^{-\frac{1}{2}[(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)]} = \frac{c_1\pi_2}{(2\pi)^{p/2}|\Sigma|^{1/2}}\,e^{-\frac{1}{2}[(x-\mu_2)^T\Sigma^{-1}(x-\mu_2)]}$$

which is equivalent to

$$\ln(\pi_1) - \frac{1}{2}{\mu_1}^T\Sigma^{-1}\mu_1 + x^T\Sigma^{-1}\mu_1 = \ln(\pi_2) - \frac{1}{2}{\mu_2}^T\Sigma^{-1}\mu_2 + x^T\Sigma^{-1}\mu_2$$

up to a scaling factor $(\mu_2 - \mu_1)^T\Sigma^{-1}(\mu_2 - \mu_1)$ if $\pi_1 = \pi_2$

Quadratic Discriminant Analysis

Quadratic discriminant analysis (QDA)

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

- The assumption that the inputs of every class have the same covariance Σ can be quite restrictive:
- If the k are not assumed to be equal, then convenient cancellations in our derivations earlier do not occur.
- The quadratic pieces in 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.

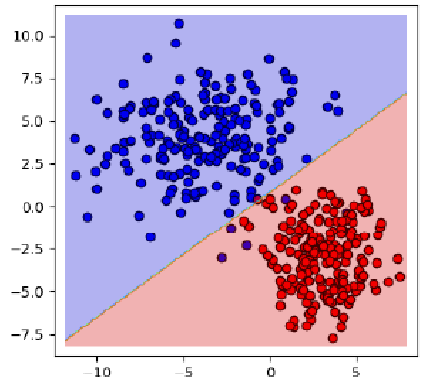
Quadratic discriminant analysis (QDA)

- In quadratic discriminant analysis we estimate a mean $\hat{\mu}_k$ and a covariance matrix $\hat{\Sigma}_k$ for each class separately.
- Given an input, it is easy to derive an objective function:

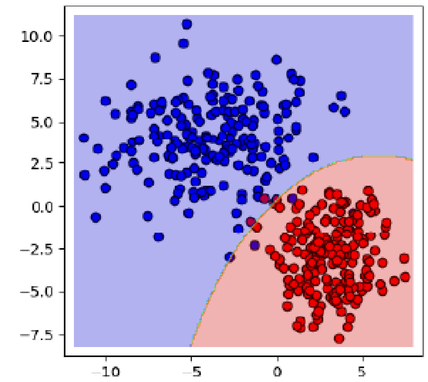
$$\delta_k(x) = \ln(\pi_k) - \frac{1}{2}\mu_k{}^T\Sigma^{-1}\mu_k + x^T\Sigma^{-1}\mu_k - \frac{1}{2}x^T\Sigma^{-1}x - \frac{1}{2}ln|\Sigma_k|$$

- This objective is now quadratic in x and so are the decision boundaries.

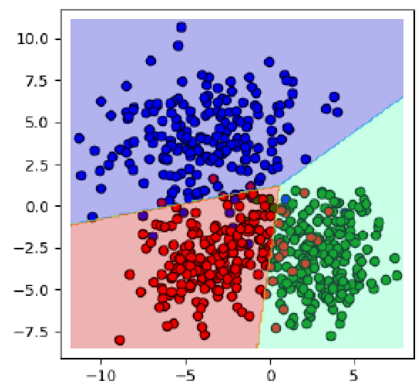
LDA VS. QDA



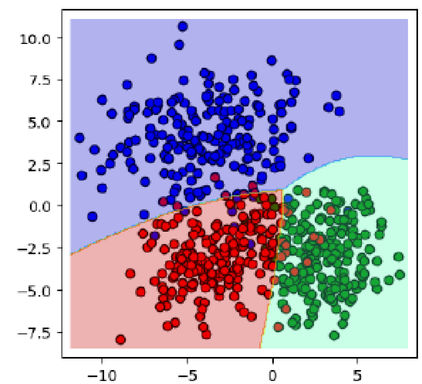
(a)



(b)



(e)



(f)

Box's Test for Equality of Covariance Matrices Across Groups

$$H_0\colon \Sigma_1 = \Sigma_2 = \cdots \Sigma_G = \Sigma$$

$$B = (1-c)\left\{\left[\sum_g (n_g - 1)\right] ln|C_w| - \sum_g \left[(n_g - 1)\ln\left|C_{w(g)}\right|\right]\right\}$$

where

$$c = \left[\sum_g \frac{1}{(n_g - 1)} - \frac{1}{\sum_g (n_g - 1)}\right]\left[\frac{2p^2 + 3p - 1}{6(p+1)(G-1)}\right]$$

Box’s Test for Equality of Covariance Matrices Across Groups

and where

$$p = number\ of\ independent\ vaiables$$
$$n_g = number\ of\ observations\ in\ group\ g$$
$$G = number\ of\ groups$$
$$n = \sum_g n_g = total\ sample\ size$$
$$C_{w(g)} = sample\ within - group\ covariance\ matrix\ for\ group\ g$$
$$C_w = sample\ within - group\ covariance\ matrix\ pooled\ across\ groups$$

Then

$$B \sim \chi^2\ (\frac{1}{2}p(p+1)(G-1))$$

Diagnostic testing

Confusion matrix:

		Predicted class		
		− or Null	+ or Non-null	Total
True class	− or Null	True Neg. (TN)	False Pos. (FP)	N
	+ or Non-null	False Neg. (FN)	True Pos. (TP)	P
	Total	N^*	P^*	

TABLE 4.6. *Possible results when applying a classifier or diagnostic test to a population.*

Name	Definition	Synonyms
False Pos. rate	FP/N	Type I error, 1−Specificity
True Pos. rate	TP/P	1−Type II error, power, sensitivity, recall
Pos. Pred. value	TP/P^*	Precision, 1−false discovery proportion
Neg. Pred. value	TN/N^*	

TABLE 4.7. *Important measures for classification and diagnostic testing, derived from quantities in Table 4.6.*

Diagnostic testing

- Precision:

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Recall:

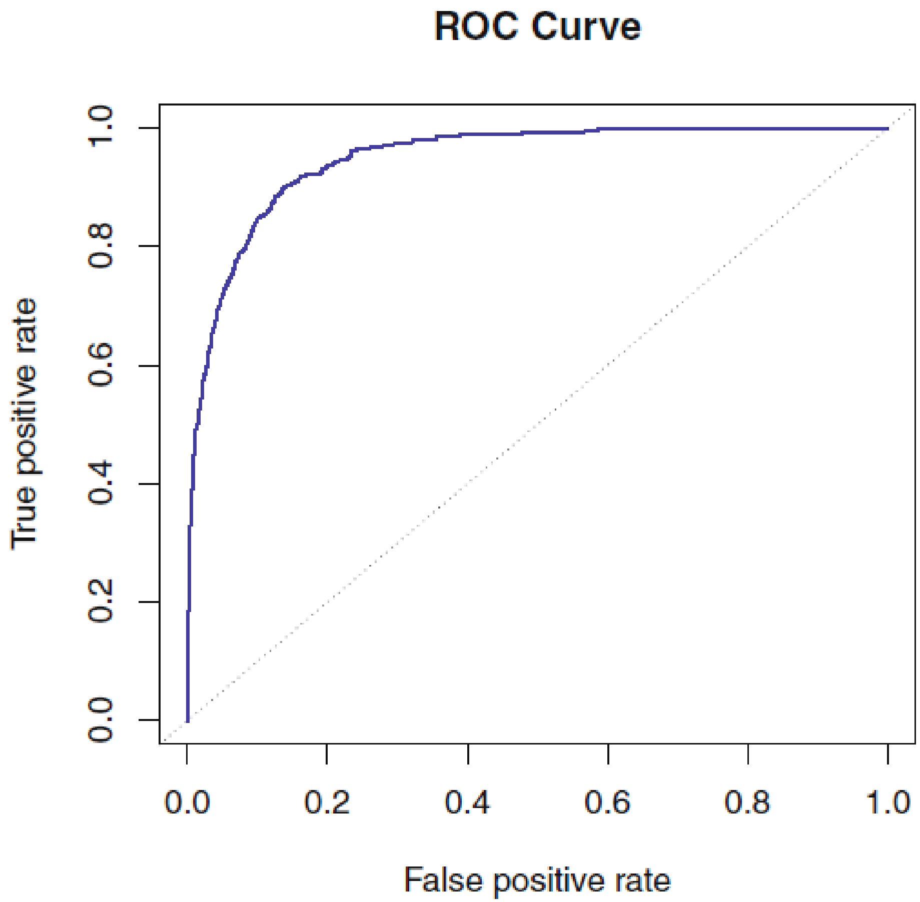
$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

- F1 score:

$$\text{Precision} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}$$

Diagnostic testing

- ROC (Receiver Operating Characteristics) Curve



Displays the performance of the method for any choice of threshold.
The area under the curve (AUC) measures the quality of the classifier:

- 0.5 is the AUC for a random classifier
- The closer AUC is to 1, the better

Sample Problem: Real Estate

- Real Estate data from a multiple listing service (MLS) for three communities in the San Francisco Bay Area: Los Altos, Menlo Park, and Palo Alto
- Samples: 9 homes in Los Altos, 13 in Menlo Park, and 13 in Palo Alto
- Three characteristics for each listing:
  1. Asking price for the property (in thousands of dollars)
  2. Number of bedrooms in the home
  3. Approximate square footage of the property (in thousands).

Sample Problem: Real Estate

Research Questions:

- Are the three communities significantly different with respect to the characteristics of the properties available for sale?
- If so, how do we describe the differences across communities?
- How many discriminant functions are necessary and how do we interpret them?

Sample Problem: Real Estate

Test of equality of covariance matrices for real estate data

TABLE 12.12 Test of equality of covariance matrices for real estate data

District	$\ln \lvert \mathbf{C}_W \rvert$
Los Altos	11.1762
Menlo Park	8.9920
Palo Alto	9.9406
Pooled	10.3657

$\chi^2 = 12.97$ with 12 *df* $p = 0.3713$

Sample Problem: Real Estate

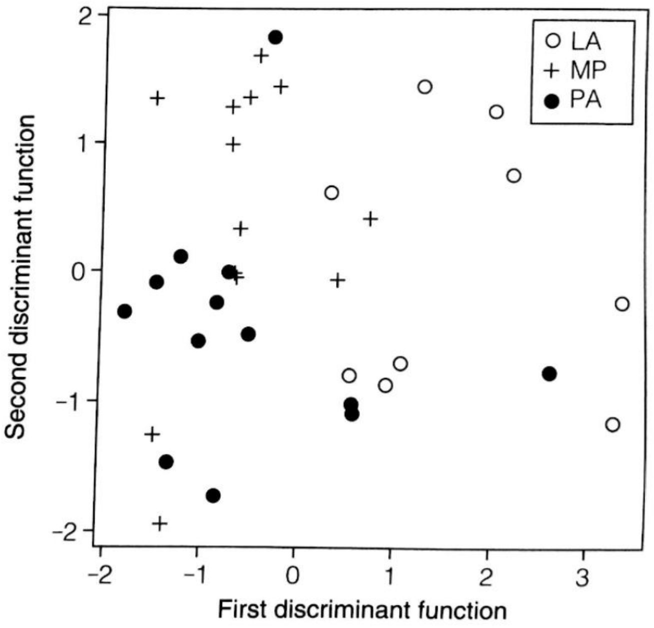
1. Fisher's discriminant Analysis Results

TABLE 12.13 Results of Fisher's discriminant analysis of real estate data

	Eigenvalues of $W^{-1}A$
	λ
1	1.0352
2	0.1552

	Standardized Discriminant Function Coefficients	
	k_1	k_2
Price	0.1164	−0.7570
Bedrooms	0.2363	1.300
Lot Size	1.2818	0.1252
	Correlations between Variables and Discriminant Functions	
	1	2
Price	0.6181	0.0399
Bedrooms	0.2660	0.8403
Lot Size	0.9746	−0.1585
	Group Means on Discriminant Functions	
Group	1	2
Los Altos	1.6517	0.0306
Menlo Park	−0.6258	0.4259
Palo Alto	−0.5177	−0.4471

FIGURE 12.18 Plot of real estate data in discriminant function space



Sample Problem: Real Estate

2. Linear discriminant Analysis (LDA) Results

TABLE 12.14 Results of Mahalanobis method: Goodness of fit and predictive validation

	Coefficients of Mahalanobis Distance Function by Group		
	Los Altos	Menlo Park	Palo Alto
Constant	−27.3139	−16.4950	−14.2933
Price	0.0034	−0.0008	0.0042
Bedrooms	6.5300	6.4993	5.0921
Lot Size	2.1363	1.2895	1.2981

	Classification Summary: Goodness of Fit			
	Number of Observations Classified into . . .			
From . . .	Los Altos	Menlo Park	Palo Alto	Total
Los Altos	7	1	1	9
Menlo Park	1	8	4	13
Palo Alto	1	4	8	13
Total	9	13	13	35

	Classification Summary: Jackknifed Validation			
	Number of Observations Classified into . . .			
From . . .	Los Altos	Menlo Park	Palo Alto	Total
Los Altos	4	2	3	9
Menlo Park	1	7	5	13
Palo Alto	1	5	7	13
Total	6	14	15	35