CLASSIFICATION METHODS

Chapter 4 of the book “An Introduction to Statistical Learning” (James et al., 2017)

LINEAR DISCRIMINANT ANALYSIS (LDA) &

QUADRATIC DISCRIMINANT ANALYSIS (QDA)

Outline

- Overview of LDA
- Why not Logistic Regression?
- Estimating Bayes’ Classifier
- LDA Example with One Predictor (p=1)
- LDA Example with more than One Predictor (P>1)
- LDA on Default Data
- Overview of QDA
- Comparison between LDA and QDA

Linear Discriminant Analysis

- LDA undertakes the same task as Logistic Regression. It classifies data based on categorical variables
  - Making profit or not
  - Buy a product or not
  - Satisfied customer or not
  - Political party voting intention

Why Linear? Why Discriminant?

- LDA involves the determination of linear equation (just like linear regression) that will predict which group the case belongs to.

$$D = v_1 X_1 + v_2 X_2 + \ldots + v_i X_i + a$$

  - D: discriminant function
  - v: discriminant coefficient or weight for the variable
  - X: variable
  - a: constant

Purpose of LDA

- Choose the v's in a way to maximize the distance between the means of different categories
- Good predictors tend to have large v's (weight)
- We want to discriminate between the different categories
- Think of food recipe. Changing the proportions (weights) of the ingredients will change the characteristics of the finished cakes. Hopefully that will produce different types of cake!

Assumptions of LDA

- The observations are a random sample
- Each predictor variable is normally distributed

Why not Logistic Regression?

- Logistic regression is unstable when the classes are well separated
- In the case where n is small, and the distribution of predictors X is approximately normal, then LDA is more stable than Logistic Regression
- LDA is more popular when we have more than two response classes

Bayes' Classifier

- Bayes' classifier is the golden standard. Unfortunately, it is unattainable.

- So far, we have estimated it with two methods:
  - KNN classifier
  - Logistic Regression

Estimating Bayes' Classifier

- With Logistic Regression we modeled the probability of Y being from the k[th] class as

$$p(X) = \Pr(Y = k|X = x) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}}.$$

- However, Bayes' Theorem states

$$p(X) = \Pr(Y = k|X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}.$$

π_k : Probability of coming from class k (prior probability)

$f_k(x)$: Density function for X given that X is an observation from class k

Estimate π_k and $f_k(x)$

- We can estimate π_k and $f_k(x)$ to compute $p(X)$

- The most common model for $f_k(x)$ is the Normal Density

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} \exp\left(-\frac{1}{2\sigma_k^2}(x - \mu_k)^2\right)$$

- Using the density, we only need to estimate three quantities to compute $p(X)$

$$\mu_k \quad \sigma_k^2 \quad \pi_k$$

Use Training Data set for Estimation

- The mean μ_k could be estimated by the average of all training observations from the k[th] class.
- The variance σ_k^2 could be estimated as the weighted average of variances of all k classes.
- And, π_k is estimated as the proportion of the training observations that belong to the k[th] class.

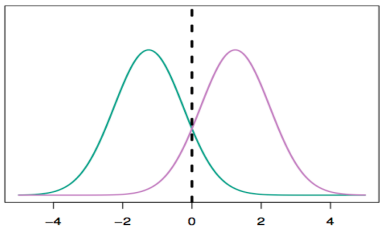
$$\hat{\mu}_k = \frac{1}{n_k} \sum_{i:y_i=k} x_i$$

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

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

A Simple Example with One Predictor (p =1)

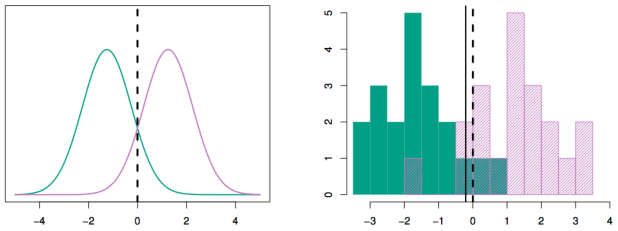
- Suppose we have only one predictor (p = 1)
- Two normal density function $f_1(x)$ and $f_2(x)$, represent two distinct classes
- The two density functions overlap, so there is some uncertainty about the class to which an observation with an unknown class belongs
- The dashed vertical line represents Bayes' decision boundary



Apply LDA

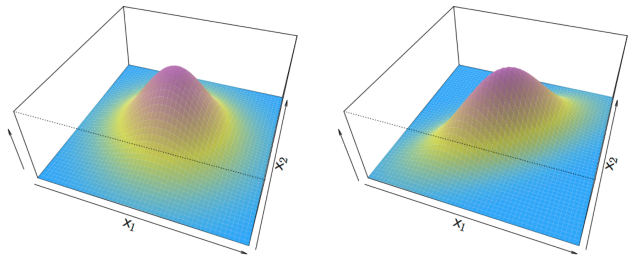
- LDA starts by assuming that each class has a normal distribution with a common variance
- The mean and the variance are estimated
- Finally, Bayes' theorem is used to compute p_k and the observation is assigned to the class with the maximum probability among all k probabilities

- 20 observations were drawn from each of the two classes
- The dashed vertical line is the Bayes' decision boundary
- The solid vertical line is the LDA decision boundary
  - Bayes' error rate: 10.6%
  - LDA error rate: 11.1%
- Thus, LDA is performing pretty well!

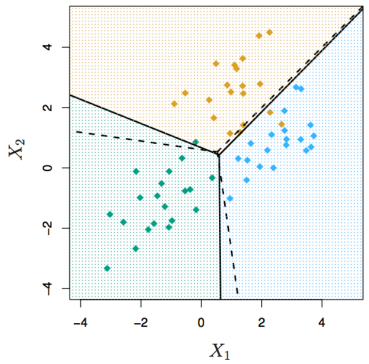


An Example When p > 1

- If X is multidimensional (p > 1), we use exactly the same approach except the density function $f(x)$ is modeled using the multivariate normal density



- We have two predictors (p =2)
- Three classes
- 20 observations were generated from each class
- The solid lines are Bayes' boundaries
- The dashed lines are LDA boundaries



Running LDA on Default Data

- LDA makes 252+ 23 mistakes on 10000 predictions (2.75% misclassification error rate)
- But LDA miss-predicts 252/333 = 75.5% of defaulters!
- Perhaps, we shouldn't use 0.5 as threshold for predicting default?

		True Default Status		
		No	Yes	Total
Predicted Default Status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

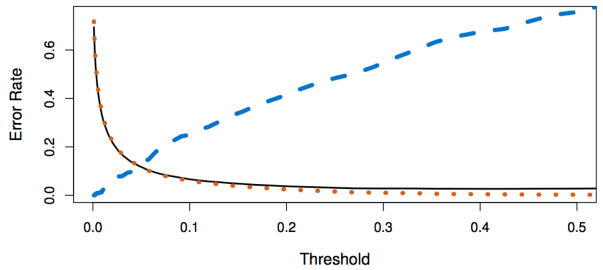
Use 0.2 as Threshold for Default

- Now the total number of mistakes is 235+138 = 373 (3.73% misclassification error rate)
- But we only miss-predicted 138/333 = 41.4% of defaulters
- We can examine the error rate with other thresholds

		True Default Status		
		No	Yes	Total
Predicted Default Status	No	9432	138	9570
	Yes	235	195	430
	Total	9667	333	10000

Default Threshold Values vs. Error Rates

- Black solid: overall error rate
- Blue dashed: Fraction of defaulters missed
- Orange dotted: non defaulters incorrectly classified



Quadratic Discriminant Analysis (QDA)

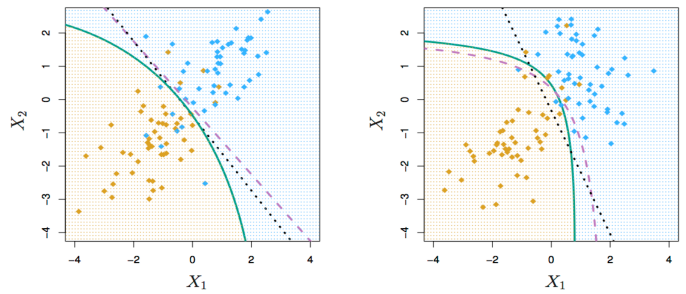
- LDA assumed that every class has the same variance/ covariance
- However, LDA may perform poorly if this assumption is far from true
- QDA works identically as LDA except that it estimates separate variances/ covariance for each class

Which is better? LDA or QDA?

- Since QDA allows for different variances among classes, the resulting boundaries become quadratic
- Which approach is better: LDA or QDA?
  - QDA will work best when the variances are very different between classes and we have enough observations to accurately estimate the variances
  - LDA will work best when the variances are similar among classes or we don't have enough data to accurately estimate the variances

Comparing LDA to QDA

- Black dotted: LDA boundary
- Purple dashed: Bayes' boundary
- Green solid: QDA boundary
- Left: variances of the classes are equal (LDA is better fit)
- Right: variances of the classes are not equal (QDA is better fit)



Comparison of Classification Methods

- KNN (Chapter 2)
- Logistic Regression (Chapter 4)
- LDA (Chapter 4)
- QDA (Chapter 4)

Logistic Regression vs. LDA

- Similarity: Both Logistic Regression and LDA produce linear boundaries
- Difference: LDA assumes that the observations are drawn from the normal distribution with common variance in each class, while logistic regression does not have this assumption. LDA would do better than Logistic Regression if the assumption of normality hold, otherwise logistic regression can outperform LDA

KNN vs. (LDA and Logistic Regression)

- KNN takes a completely different approach
- KNN is completely non-parametric: No assumptions are made about the shape of the decision boundary!
- Advantage of KNN: We can expect KNN to dominate both LDA and Logistic Regression when the decision boundary is highly non-linear
- Disadvantage of KNN: KNN does not tell us which predictors are important (no table of coefficients)

QDA vs. (LDA, Logistic Regression, and KNN)

- QDA is a compromise between non-parametric KNN method and the linear LDA and logistic regression
- If the true decision boundary is:
  - Linear: LDA and Logistic outperforms
  - Moderately Non-linear: QDA outperforms
  - More complicated: KNN is superior