Classification II

Part 2: Linear and Quadratic Discriminant Analysis (LDA and QDA)

- Overview of Linear Discriminant Analysis (LDA)
- Why not Logistic Regression?
- Estimating Bayes' Classifier
- LDA on the `Default` Data
- Overview of Quadratic Discriminant Analysis (QDA)
- Comparison between LDA, QDA, KNN and Logistic regression

Why Linear and Discriminant in the name?

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

$$D = v_1X_1 + v_2X_2 + \cdots v_iX_i + a$$

D: Discriminant function
v_i: Discriminant coefficients or weights for the corresponding variable
X_i: Variables
a: Constant

Idea behind LDA

- Choose the v's in a way to maximize the distance between the (projected) means of different categories

- Good predictors tend to have large v's (weight)

- We want to discriminate between the different categories

Assumptions of LDA

- The observations are a random sample
- Each predictor variable is normally distributed
- Additional assumption about the covariance of the variables (equality of the covariance across classes)

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

LDA - estimating the Bayes classifier

- With Logistic Regression we modeled the probability of Y being from the k^{th} class (out of K possible) as

$$P(X) = \Pr\,(Y = k|X = x) = \frac{e^{\beta_{0,k}+\beta_{1,k}x}}{1 + \ e^{\beta_{0,k}+\beta_{1,k}x}}$$

- However, Bayes' Theorem states

$$p_k(x) \equiv \Pr(Y = k|X = x) = \frac{\Pr\,(X = x|Y = k)\,\Pr\,(\mathrm{Y} = \mathrm{k})}{\Pr\,(X = x)} = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$$

where

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

$f_k(x)$: Density function for X given that Y is an observation from class k (i.e. it is the conditional density $\Pr\,(X = x|Y = k)$)

Estimating π_k and $f_k(x)$

- We can estimate π_k, $f_k(x)$ to compute the posterior probabilities $p_k(X)$
- The most common model for $f_k(x)$ is the Normal Density

$$f_k(x) = \frac{1}{\sigma_k\sqrt{2\pi}}\, e^{-\frac{(x-\mu_k)^2}{2\sigma_k^2}}$$

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

$$\mu_k \qquad \sigma_k \qquad \pi_k$$

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 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.

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

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

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

- When common variance $\sigma_k = \sigma, k = 1, \dots, K$ for all classes is assumed, the above Bayes's formula reduces to the rule "assign the observation to the class for which

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log(\pi_k)$$

is maximum"

- If $p = 1$, $K = 2$ and $\pi_1 = \pi_2$, then the Bayes classifier assigns an observation to class 1 if

$$2x(\mu_1 - \mu_2) > \mu_1^2 - \mu_2^2$$

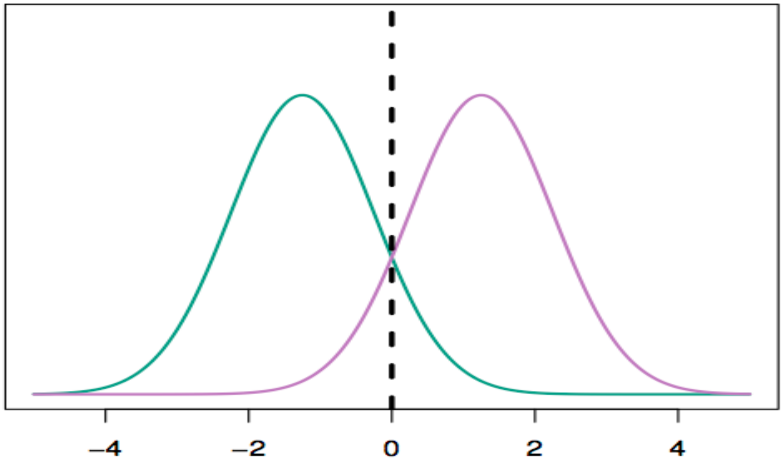
and to class 2 otherwise.

- In this case, the Bayes decision boundary corresponds to the point where

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

A Simple Example with One Predictor ($p = 1$) illustrating above rule

- 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



Summary of LDA

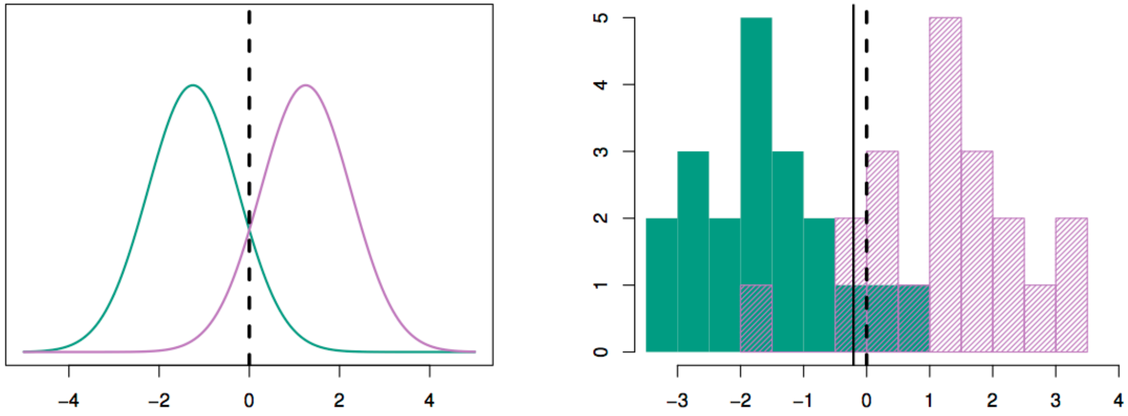
- LDA starts by assuming that each class has a normal distribution with a common variance

- 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

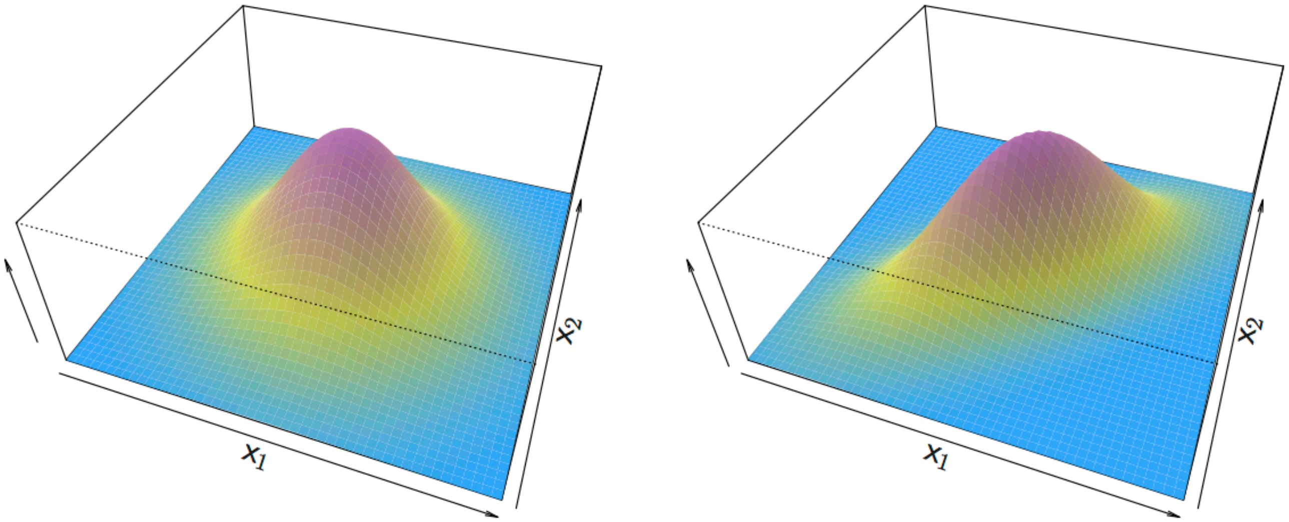
Example with $p = 1$

- 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!



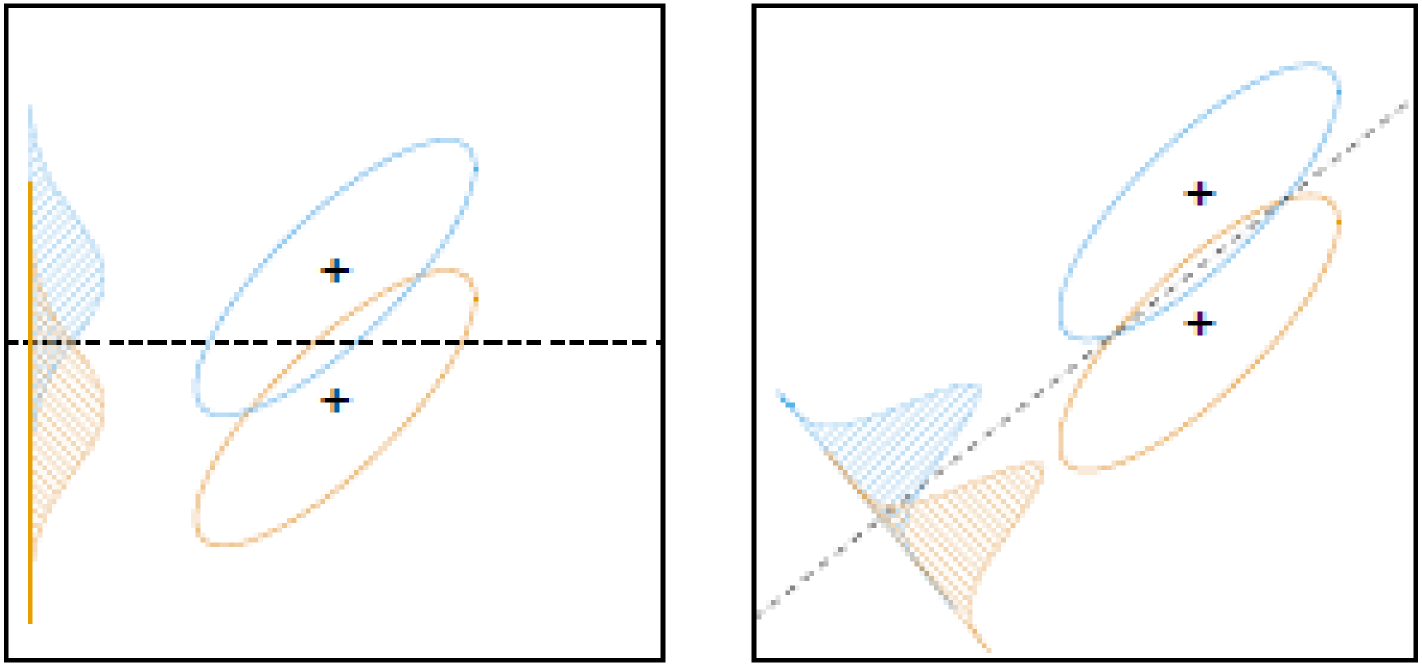
Example with $p > 1$

- 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



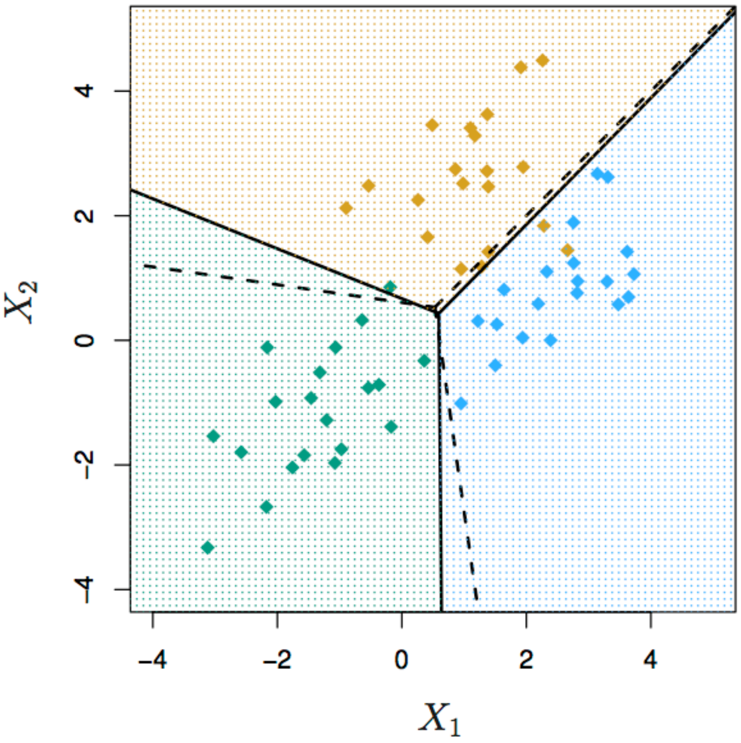
For the distribution on the left X_1 & X_2 are uncorrelated.
On the right, the correlation is 0.7

- LDA reduces the data to 1-dim using the discriminant function
- The graphs below illustrate this. On the left is a non-optimal direction of the vector of the coefficients in the linear combination (the 1-dim distributions overlap more) and on the right we have the vector defining the linear discriminant function:



Example. We have two predictors X_1 and X_2 ($p = 2$) and three classes

- 20 observations were generated from each class
- The solid lines are Bayes' boundaries
- The dashed lines are LDA boundaries



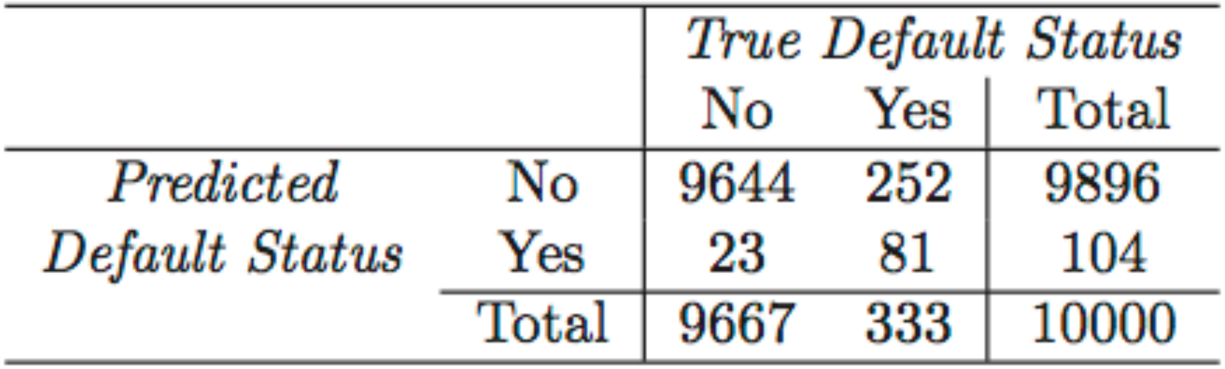
Running LDA on the “Default” Data

- We would like to be able to predict customers that are likely to default

- X variables are

    - Monthly credit card `balance` (X_1)
    - Annual `income` (X_2)
    - The Y variable (Default) is categorical: `Yes` or `No`

- LDA makes $252 + 23$ mistakes on $10{,}000$ predictions (2.75% misclassification error rate)

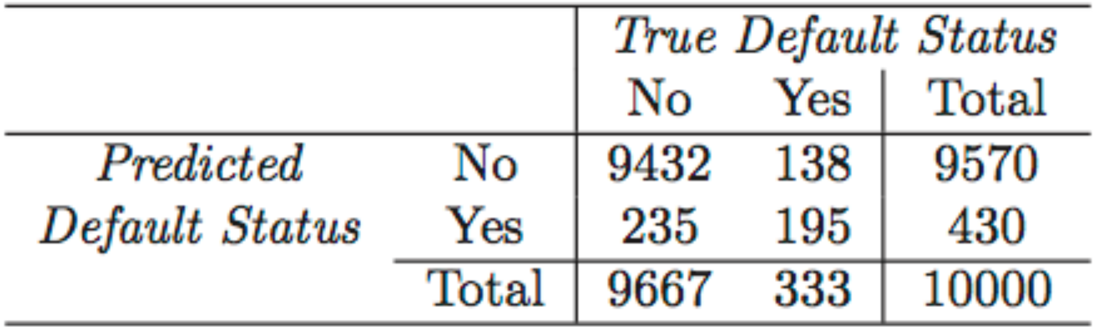


		True Default Status		
		No	Yes	Total
Predicted	No	9644	252	9896
Default Status	Yes	23	81	104
	Total	9667	333	10000

- But LDA miss-predicts $252/333 = 75.5\%$ of defaulters!
- Remember that LDA "approximates" the Bayes classifier which assigns the observation to the class with the greatest posterior probability $p_k(x)$. In the 2 class case we predict Default if $\Pr(default = Yes|\, X = x) > 0.5$
- Perhaps, we shouldn't use 0.5 as threshold for predicting default?

Let’s use a 0.2 threshold

- Now the total number of mistakes is $235 + 138 = 373$ (3.73% misclassification error rate)



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

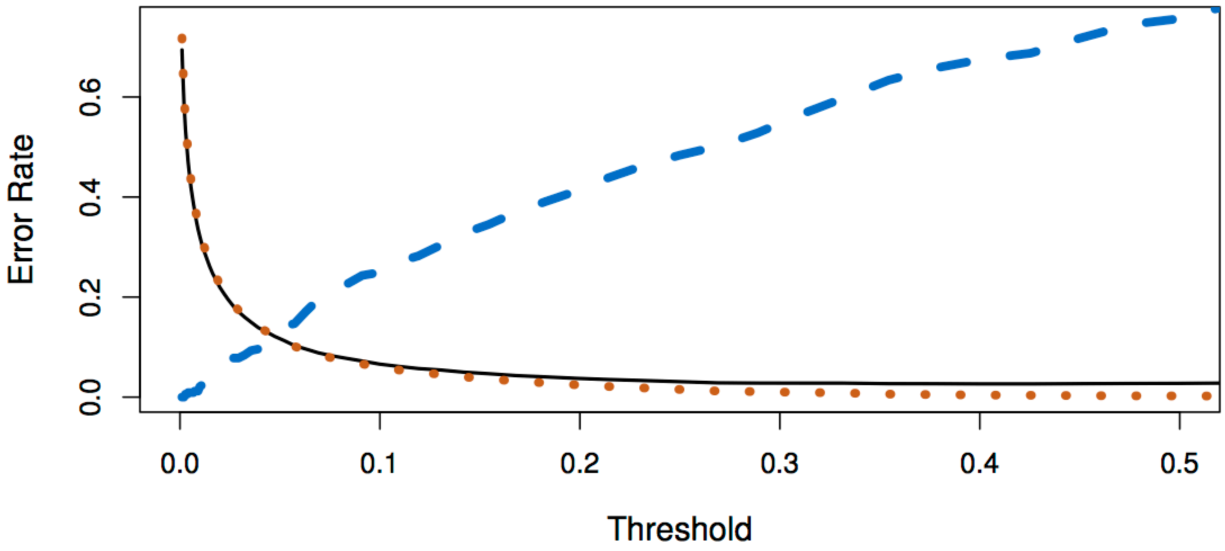
- But we only miss-predicted $138/333 = 41.4\%$ of defaulters

- In this case (small increase in overall error but detecting much more of the defaulters) is preferable

- We can examine the error rate with other thresholds

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
- But LDA may perform poorly if this assumption is far from being true
- QDA works identically as LDA except that it estimates separate variances/covariances for each class
- It assumes an observation from the class k is of the form $X \sim N(\mu_k, \mathbf{\Sigma}_k)$ where $\mathbf{\Sigma}_k$ is a covariance matrix for class k. An observation $X = x$ is assigned to the class for which

$$\begin{aligned}\delta_k(x) &= -\frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}_k^{-1}(x-\mu_k) - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log(\pi_k) \\ &= -\frac{1}{2}x^T\mathbf{\Sigma}_k^{-1}x + x^T\mathbf{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\mu_k^T\mathbf{\Sigma}_k^{-1}\mu_k - \frac{1}{2}\log|\mathbf{\Sigma}_k| + \log(\pi_k)\end{aligned}$$

is largest. This is a quadratic function of x thus the name QDA

Which is better? LDA or QDA?

- Since QDA allows for different variances among classes, the resulting boundaries become quadratic

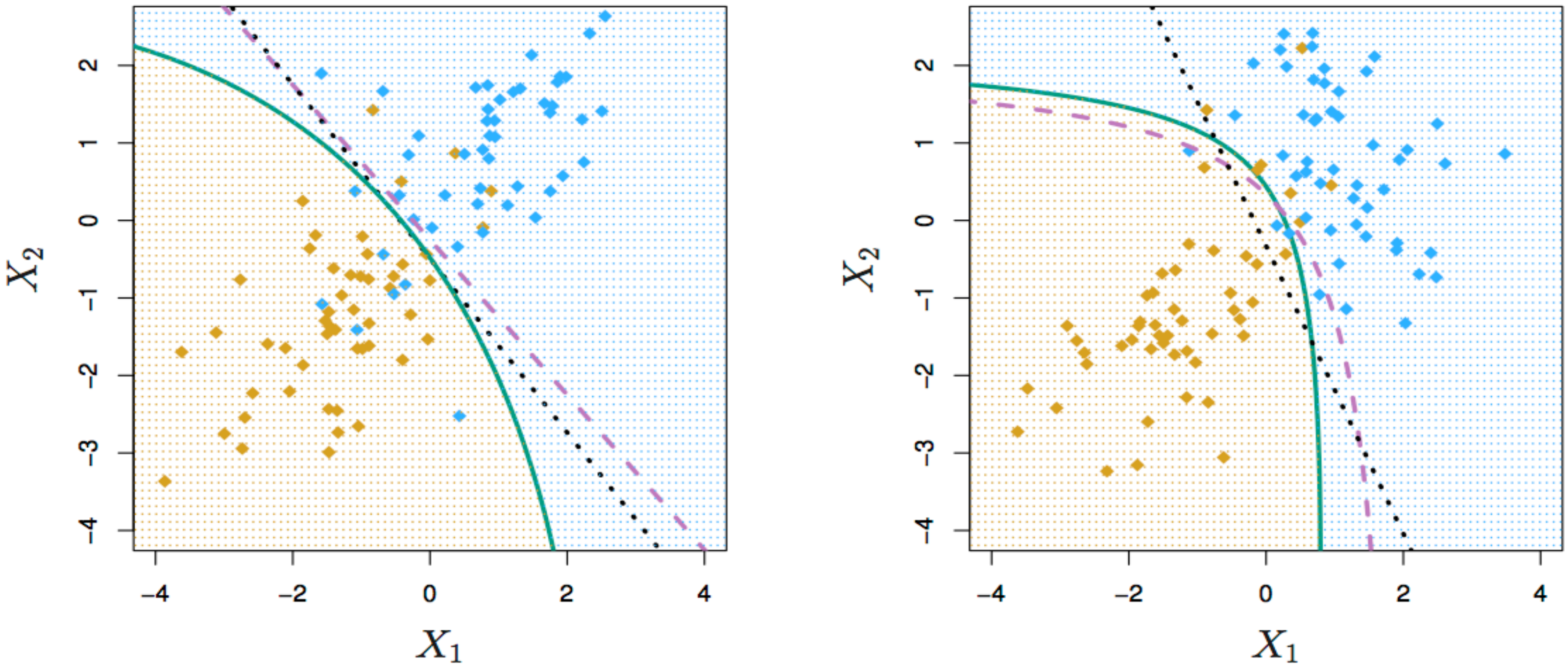
- Which approach is better?

  - 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

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 holds, 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
- Disadvantages of KNN: KNN does not tell us which predictors are important (no coefficients); KNN can only “learn” the best number of neighbors on train data. (And perhaps which distance to use in forming neighborhoods)

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 outperform
  - **Moderately Non-linear**: QDA outperforms
  - **More complicated**: KNN is superior

Python code for the `Smarket` data

Reading:

ESLII: Chapter 4.1-4.4
ISLR: Chapter 4