2. Linear Classification

Jesper Armouti-Hansen

University of Cologne

December 17, 2018

Course website

jeshan49.github.io/eemp2/

Today

- Lecture[1]:
  - Logistic Regression
  - Discriminant Analysis
- Tutorial:
  - Reproducing some results from the lecture

[1] Some of the figures in this presentation are taken from "An Introduction to Statistical Learning, with applications in R" (Springer, 2013) with permission from the authors: G. James, D. Witten, T. Hastie and R. Tibshirani

Recall: Classification

Our response Y is qualitative

- e.g. $\mathcal{C} = \{\text{spam, ham}\}$ or $\mathcal{C} = \{0, ..., 9\}$

We wish to build a classifier $C(X)$ that assigns a class label from $\mathcal{C}$ to a future unlabeled observation X

- Suppose $\mathcal{C}$ contains K elements numbered $1, ..., K$
- Let $p_k(x) = Pr(Y = k|X = x)$, $k = 1, ..., K$
- Suppose we knew the conditional probability of Y given X
- Then, the *Bayes optimal classifier* at x given by

$$C(x) = j \text{ if } p_j(x) = \max\{p_1(x), ..., p_K(x)\} \quad (1)$$

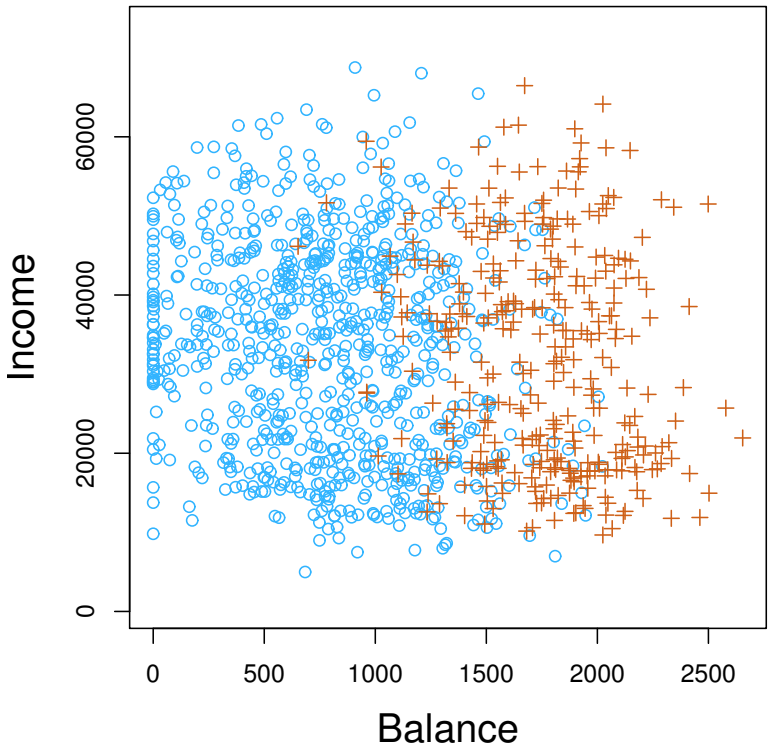
is optimal in the sense that it minimizes the expected one-zero loss

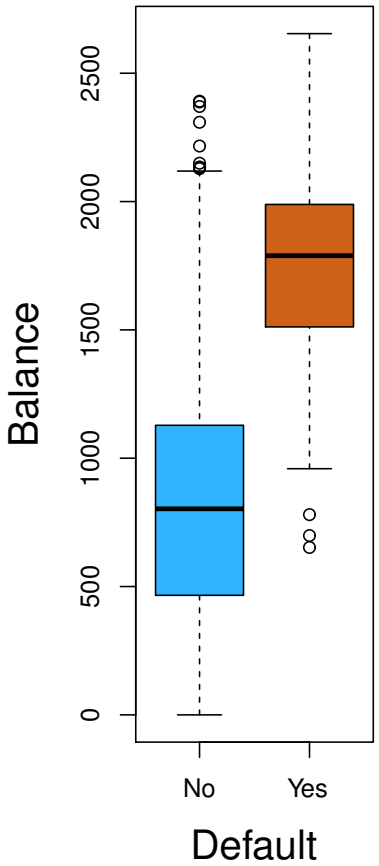
Linear Classification

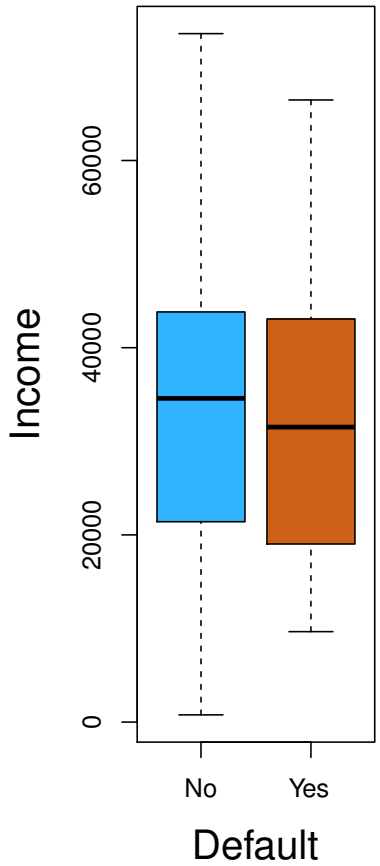
- What do we mean by linear classification?
- We have already seen one classification method: The kNN classifier
- kNN makes no assumption about the boundary that determines which of two classes an observation will be classified as
- We saw that in this case, we get an unstructured rough decision boundary
- In this lecture, we will look at methods that explicitly require the boundary to be linear

The Default data

- annual income and monthly credit card balance for 10,000 individuals







Linear Regression

- Suppose we want to use the balance to classify default. In this case, we could consider using linear regression
- we simply code Y

$$Y = \mathbb{I}(\text{Default} = \text{Yes}) \tag{2}$$

- then we classify to Yes if our estimate $\hat{Y}$ is larger than 0.5
- Since, in the population, we have $E[Y|X = x] = Pr(Y = 1|X = x)$, regression seems to be good for this task
- However, if the range of X is not limited, we will see probability estimates below zero and above one
- In the case of multiple unordered classes, there is no straightforward way of applying linear regression

Logistic Regression

- To avoid getting probability estimates outside $[0, 1]$, we model $p(X)$ using a function which lies in the interval for all values of X
- In the case of logistic regression, we use the logistic or sigmoid function

$$p(X) = \frac{exp(\beta_0 + \beta_1 X)}{1 + exp(\beta_0 + \beta_1 X)} = \frac{1}{exp(-(\beta_0 + \beta_1 X)) + 1} \quad (3)$$

- To fit (3) we use maximum likelihood
- From this expression, we can get to the odds

$$\frac{p(X)}{1 - p(X)} = exp(\beta_0 + \beta_1 X) \quad (4)$$

- The higher the odds, the more likely is the event

Linear vs. Logistic regression on Default data ($p = 1$)

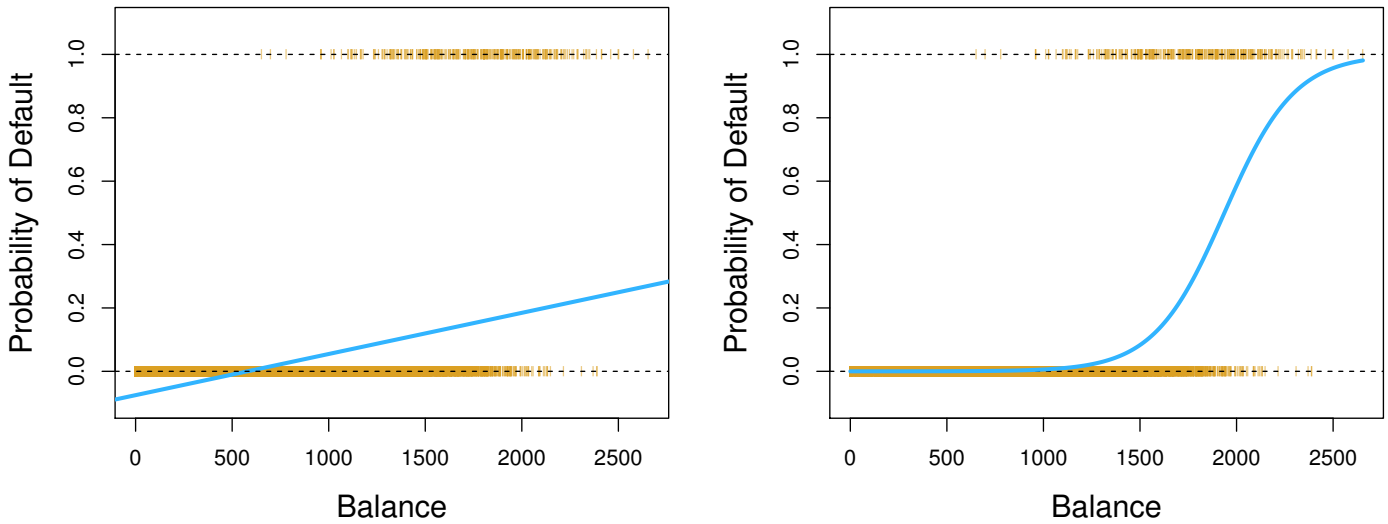


Figure: Left: Estimated probability of default using linear regression. Right: Same estimation using logistic regression (See ISLR p. 131)

- Log-transforming (4) gives us the log-odds or logit:

$$log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X \tag{5}$$

- since the logit is linear in X, increasing X by one unit changes the log-odds by β_1, or multiplies the odds by $exp(\beta_1)$
- However, since $p(X)$ is not a linear function of X, the amount that $p(X)$ changes due to a one unit increase in X will depend on the current value of X
- Note: our decision boundary is the set of points $\{x|p(x) = 0.5\}$
- This is equivalent to the set of points $\{x|\beta_0 + \beta_1 x = 0\}$. Thus, the decision boundary is linear in x

Estimating the Coefficients

- We estimate the coefficients using maximum likelihood

$$\ell(\beta_0, \beta_1) = \prod_{i|y_i=1} p(x_i) \prod_{i'|y_{i'}=0} (1 - p(x_{i'})) \tag{6}$$

- That is, given our specification of $p(X)$, we find the β_0, β_1 that yields the highest likelihood of the observed 0's and 1's in our training data

	Coefficients	Std. Error	Z-statistic	P-value
Intercept	-10.6513	0.3612	-29.5	$<$0.0001
balance	0.0055	0.0002	24.9	$<$0.0001

Thus,

$$\hat{p}(1000) = \frac{1}{exp(10.6513 - 0.0055 \times 1000) + 1} \approx 0.006 \tag{7}$$

Multiple Logistic Regression

- We can extend logistic regression to the case with $p > 1$ straightforwardly

$$p(X) = \frac{exp(\beta_0 + \sum_{i=1}^{p} \beta_i X_i)}{1 + exp(\beta_0 + \sum_{i=1}^{p} \beta_i X_i)} \tag{8}$$

- it follows immediately that the logit is linear in X

	Coefficients	Std. Error	Z-statistic	P-value
Intercept	-10.8690	0.4923	-22.08	$<$0.0001
balance	0.0057	0.0002	24.74	$<$0.0001
income	0.0030	0.0082	0.37	0.7115
student	-0.6468	0.2362	-2.74	0.0062

Multi-class Logistic Regression

- We can extend logistic regression to the case where $K > 2$
- In general, the model has the form

$$p_k(x) = \frac{exp(\beta_{0k} + \sum_{i=1}^{p} \beta_{ik} x_i)}{1 + \sum_{\ell=1}^{K-1} exp(\beta_{0\ell} + \sum_{i=1}^{p} \beta_{i\ell} x_i)}, k = 1, ..., K-1$$

$$p_K(x) = \frac{1}{1 + \sum_{\ell=1}^{K-1} exp(\beta_{0\ell} + \sum_{i=1}^{p} \beta_{i\ell} x_i)}$$

- It follows immediately that $\sum_{i=1}^{K} p_i(x) = 1$
- Multi-class logistic regression is known as multinomial regression

Discriminant Analysis

- In this alternative approach, we instead estimate
  1. $\hat{Pr}(X|Y)$: distribution of inputs given the output
  2. $\hat{Pr}(Y)$: distribution of classes
- Then, we use Bayes' theorem to obtain $\hat{Pr}(Y|X)$
- Finally, we then classify observations optimally
- Here, we will restrict ourselves in assuming that class conditional distributions of X are normal
- We will see that this leads to linear or quadratic decision boundaries

Using Bayes' Theorem for Classification

- Bayes' theorem states:

$$Pr(Y = k|X = x) = \frac{Pr(X = x|Y = k)Pr(Y = k)}{Pr(X = x)} \tag{9}$$

- Let
  - $\pi_k = Pr(Y = k)$ be the prior probability for class k
  - $f_k(x) = Pr(X = x|Y = k)$ be the density function of X for an observation from class k
- Then we can rewrite (9) as

$$Pr(Y = k|X = x) = p_k(x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \tag{10}$$

- Thus, we can plug in π_k and $f_k(X)$ into (10) to get $p_k(X)$
- Since we have a random sample, estimating π_k is not difficult
- However, estimating $f_k(X)$ is more challenging
  - unless we assume some simple forms for these densities
  - If we can find a way to estimate $f_k(X)$, we can develop model that approximates the Bayes classifier

Linear Discriminant Analysis (LDA)

- For now, assume that $p = 1$
- We assume that the class conditional input density is normal

$$f_k(x) = \frac{1}{\sqrt{2\pi}\sigma_k} exp\left(-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma_k}\right)^2\right) \tag{11}$$

- For the LDA, we assume a shared variance term for all classes, i.e. $\sigma_1^2 = \cdots = \sigma_K^2 = \sigma^2$
- inserting this back into to (10) gives us

$$p_k(x) = \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2\right)}{\sum_{l=1}^{K} \pi_l \frac{1}{\sqrt{2\pi}\sigma} exp\left(-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2\right)} \tag{12}$$

- The Bayes classifier assigns an observation $X = x$ to the class for which (12) is the highest
- Taking the log and removing constants yields the discriminant function:

$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + log(\pi_k) \tag{13}$$

- The decision boundary between two classes k and l is given by the set of points such that

$$\{x|\delta_k(x) = \delta_l(x)\}$$

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

$$\iff x\underbrace{\frac{\mu_k - \mu_l}{\sigma^2}}_{c_1} + \underbrace{\frac{\mu_l^2 - \mu_k^2}{2\sigma^2} + log(\pi_k) - log(\pi_l)}_{c_0} = 0$$

- This is a linear equation in x

- Naturally, we need to estimate the parameters with our training data

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

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

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

$$= \sum_{k=1}^{K} \frac{n_k - 1}{n-K} \hat{\sigma}_k^2$$

- Note: $\hat{\sigma}_k^2 = \frac{1}{n_k - 1} \sum_{i|y_i=k} (x_i - \hat{\mu}_k)^2$ is the usual formula for the estimated variance in class k

Example (LDA, $p = 1$)

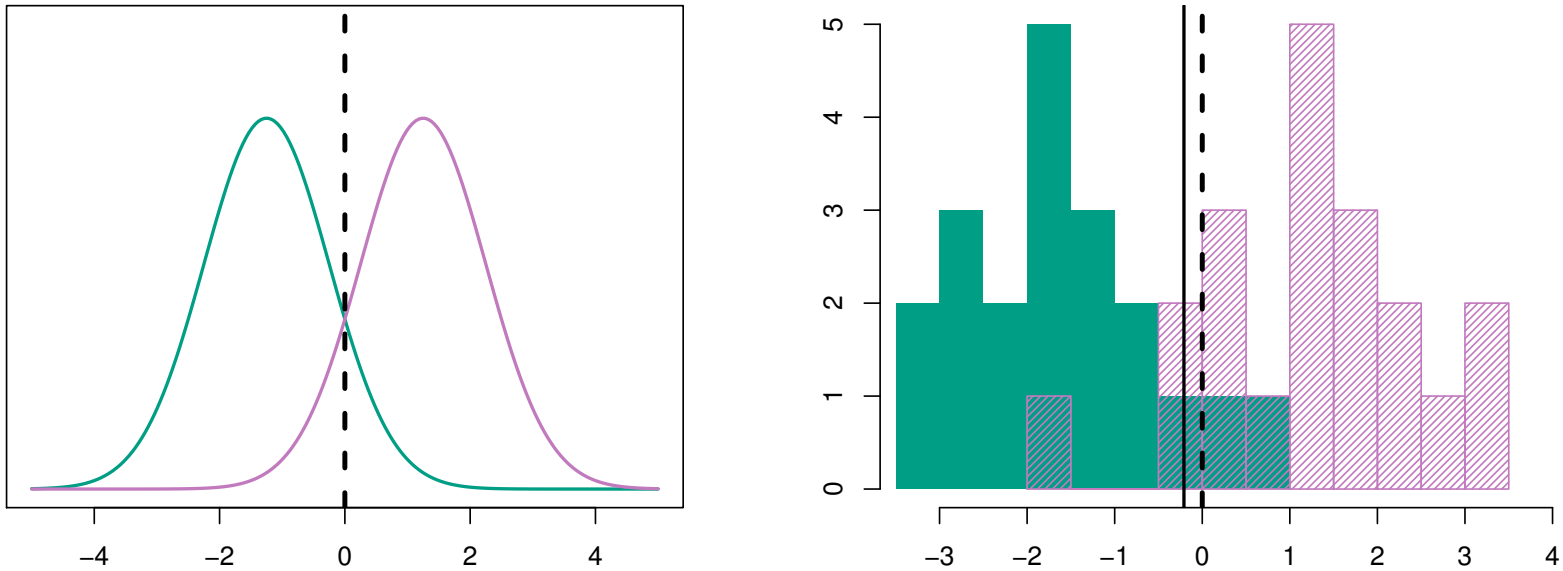


Figure: Left: Two normal density functions. Right: 20 observations from each class as histograms. Bayes decision boundary (dashed) and LDA decision boundary (solid) (See ISLR p. 140)

LDA for $p > 1$

- LDA can be extended simply to higher dimensions:

$$f_k(x) = \frac{1}{(2\pi)^{\frac{p}{2}}|\mathbf{\Sigma}|^{\frac{1}{2}}} exp\left(-\frac{1}{2}(x-\mu_k)^T\mathbf{\Sigma}^{-1}(x-\mu_k)\right) \quad (14)$$

- Where
  - μ_k is a class-specific mean vector
  - $\mathbf{\Sigma}$ is a covariance matrix that is common to all K classes
- Performing a little algebra gives

$$\delta_k(x) = x^T\mathbf{\Sigma}^{-1}\mu_k - \frac{1}{2}\mu_k^T\mathbf{\Sigma}^{-1}\mu_k + log(\pi_k) \quad (15)$$

- revealing that the decision boundaries are linear

Example (LDA, $p = 2$)

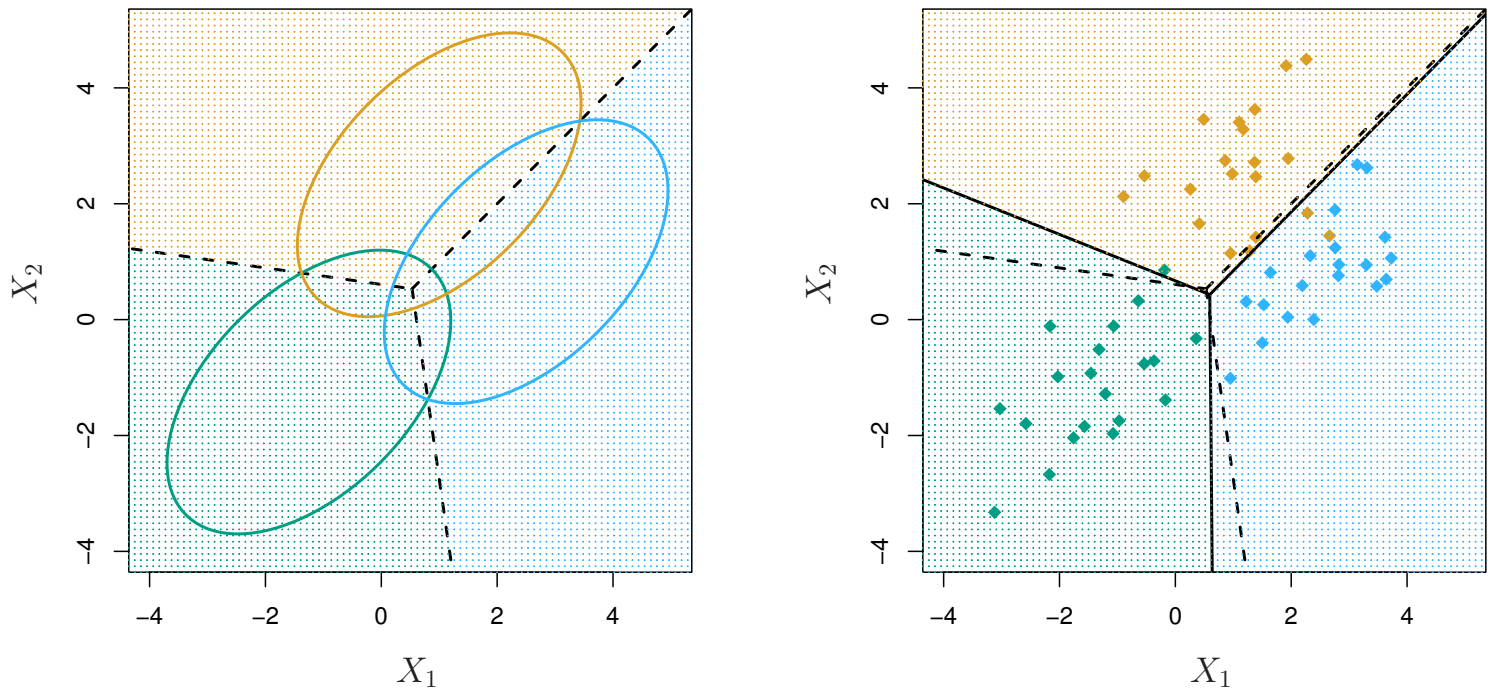


Figure: Multivariate normal distribution with class specific mean and a common covariance matrix. Left: Elipses that contain 95% of the probability for each of the three classes. The dashed lines are the Bayes decision boundaries. Right: 20 observations drawn from each class, and the corresponding LDA decision boundaries are indicated with solid black lines (See ISLR p. 143)

Turning $\delta_k(x)$ into $p_k(x)$

- We can turn our estimates $\delta_k(x)$ into estimates of class probabilities

$$\hat{p}_k(x) = \frac{exp(\hat{\delta}_k(x))}{\sum_{l=1}^{K} exp(\hat{\delta}_l(x))} \tag{16}$$

- Thus, classifying to the largest $\hat{\delta}_k(x)$ is equivalent to classify according to the largest $\hat{p}_k(x)$.

Applying LDA to the Default data

- If we apply LDA to the Default data, we get:

		True default status		
		No	Yes	Total
Predicted default status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

- Thus, we have $(23 + 252)/10000 = 2.75\%$ misclassification rate.
- This looks like a small error rate, however:
  - recall this is our training error – it may be too optimistic
  - if we were to classify according to the prior of default, we would have a misclassification rate of $333/10000 = 3.33\%$

- If we apply LDA to the Default data, we get:

		True default status		
		No	Yes	Total
Predicted default status	No	9644	252	9896
	Yes	23	81	104
	Total	9667	333	10000

- Of the individuals who do not default, we misclassify with a rate of $23/9667 = 0.2\%$
- Of the individuals who do default, we misclassify with a rate of $252/333 = 75.7\%$
- Thus, the total misclassification rate may not provide us with enough information to evaluate our method

Types of Errors

- **False positive rate (FPR):** fraction of negative examples classified as positive
  - LDA on Default: 0.2%
- **True positive rate (TPR):** fraction of positive examples classified as positive
  - LDA on Default: $81/333 = 24.3\%$
- **False negative rate (FNR):** fraction of positive examples classified as negative
  - LDA on Default: 75.7%
- **True negative rate (TNR):** fraction of negative examples classified as negative
  - LDA on Default: $9644/9667 = 99.8\%$
- If we vary the threshold of classification, we can increase TPR at the cost of increasing FPR

Receiver Operating Characteristics (ROC)

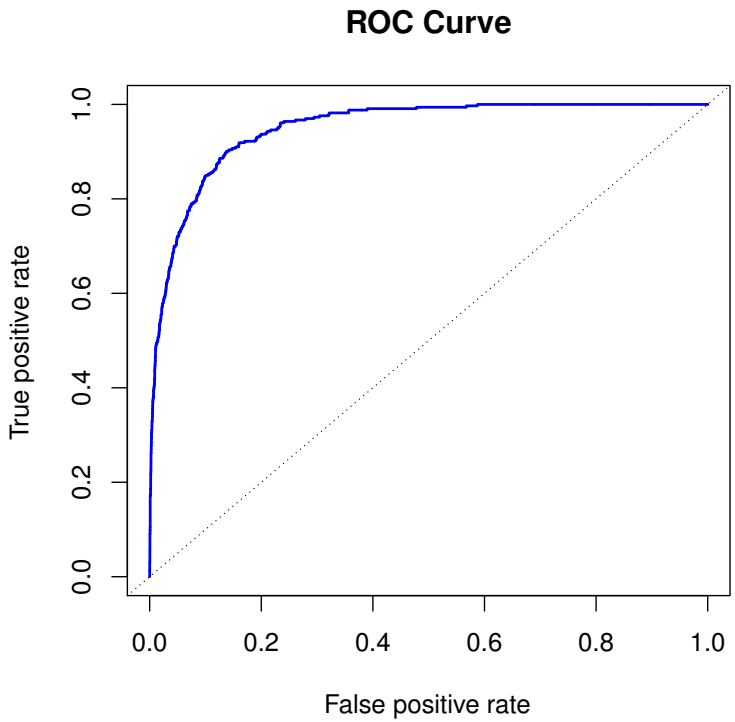


Figure: ROC curve for the LDA on the Default data (blue, solid). area-under-curve (AUC) = 0.95 (See ISLR p. 148)

- Our current true positive error rate: $81/333 = 24.3\%$

Quadratic Discriminant Analysis (QDA)

- As LDA, QDA assumes that observations from each class follow a normal distribution
- As LDA, QDA plugs estimates for the parameters into Bayes' theorem
- Unlike LDA, QDA assumes each class has its own covariance matrix
  - That is, observations from class k is of the form $X \sim \mathcal{N}(\mu_k, \mathbf{\Sigma}_k)$
- Following this estimation, we classify an observation to the class with the highest conditional probability

- Letting $f_k(x)$ have this form and plugging this into Bayes' theorem, it can be shown that the discriminant function of class k is given by

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

- Note: x appears as a quadratic function
- Thus, the decision boundary between class k and l is now non-linear

Example (QDA)

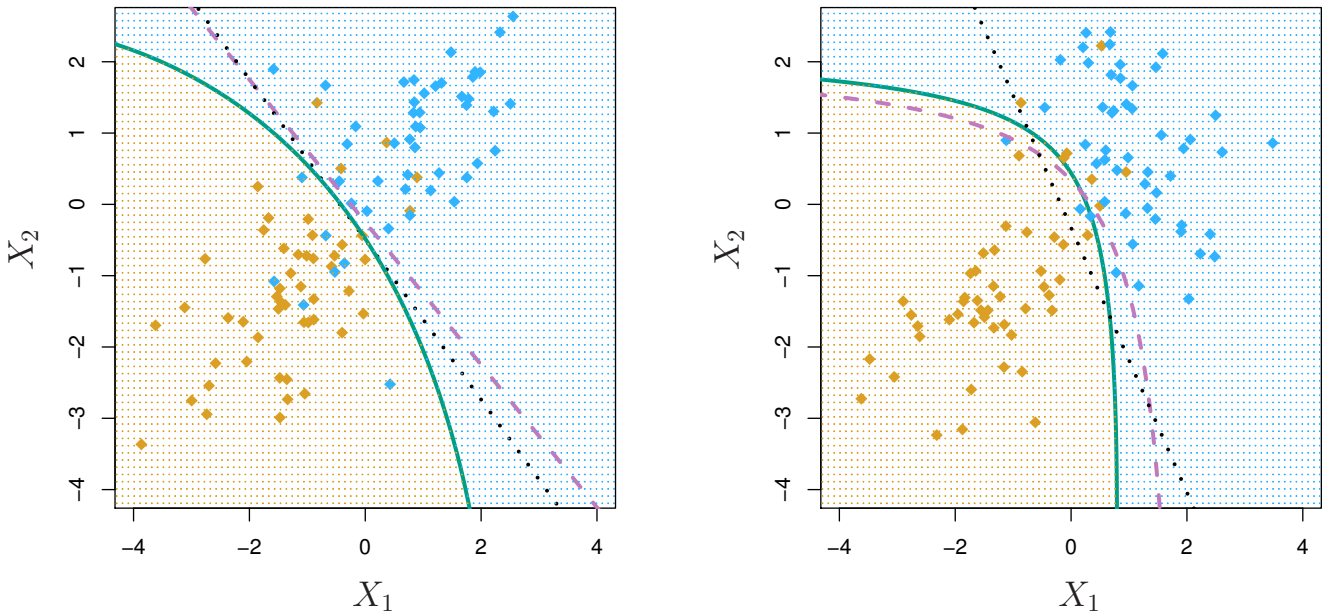


Figure: Left: The Bayes (purple, dashed), LDA (black, dotted), and QDA (green, solid) decision boundaries for a two class problem with $\mathbf{\Sigma}_1 = \mathbf{\Sigma}_2$. Right: Same as left, except that $\mathbf{\Sigma}_1 \neq \mathbf{\Sigma}_2$. (See ISLR p. 150)

LDA vs. QDA

- For LDA, it can be shown that we only need to estimate $(K-1) \times (p+1)$ parameters, since we only need the differences $\delta_k(x) - \delta_l(x)$
- For QDA, we will need to estimate $(K-1) \times (p(p+3)/2+1)$ parameters
- LDA will in general tend to have higher bias than QDA
- On the other hand, QDA will tend to have higher variance than LDA
- Thus, we are back at the bias-variance trade-off
- LDA and QDA are two simple tools that, historically, have a good track recorded on various classification tasks

Other forms of discriminant analysis

- Naive Bayes assumes that the inputs are independent:

$$f_l(x) = \prod_{k=1}^{p} f_{lk}(x_k) \tag{17}$$

  - this can be appropriate when p is large, making density estimation unattractive
- General non-parametric density estimates
- More flexible mixtures of normal distributions allowing for non-linear decision boundaries

Comparison

- let $p = 1$ and $K = 2$, then the LDA log-odds are given by

$$log\left(\frac{p_1}{1-p_1}\right) = c_0 + c_1 x \tag{18}$$

  - where c_0, c_1 are functions of $\mu_1, \mu_2, \pi_1, \pi_2$, and σ^2
- On the other hand, the logistic regression log-odds are given by

$$log\left(\frac{p_1}{1-p_1}\right) = \beta_0 + \beta_1 x \tag{19}$$

- Thus, they both produce linear decision boundaries, but are estimated differently
- LDA can provide improvement over logistic regression when its assumptions holds approximately
- On the other hand, logistic regression can provide an improvement when LDA's assumptions are clearly violated

- Recall that kNN classification is a completely non-parametric approach
  - we make no assumption about the decision boundary
  - hence, this method may dominate both LDA and logistic regression if the decision boundary is highly non-linear
- QDA can be seen as a compromise between kNN and LDA, logistic regression
- it is less flexible than kNN, but may do better in presence of limited training observations
  - This is due to the fact that it makes assumptions about the decision boundary

References

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning (Vol. 112). **Chapter 4**

Friedman, J., Hastie, T., & Tibshirani, R. (2001). The elements of statistical learning (Vol. 1, No. 10). **Chapters 4**