Lecture 7: Classification (Textbook 4.1-4.4)

Classification

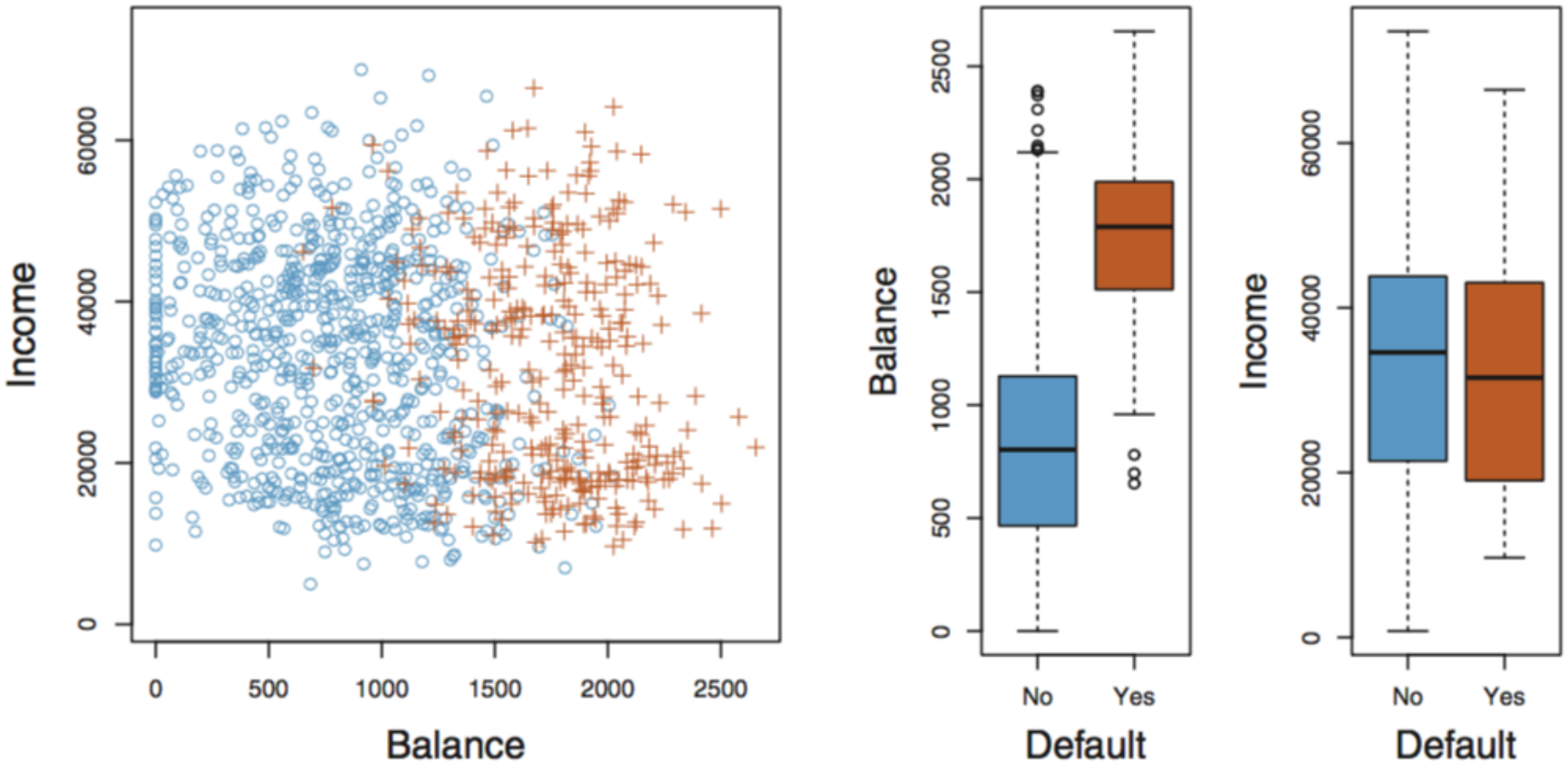
- Qualitative variables take values in an unordered set C, such as:

  eye color$\in$ {brown,blue,green},
  email$\in$ {spam, ham}.

- Our goal: Given a feature vector X and a qualitative response Y taking values in the set C, we aim to build a function $C(X)$ that uses the feature vector X to predict Y; i.e. $C(X) \in C$.

- In this chapter we discuss three of the most widely-used classifiers: **logistic regression**, **linear discriminant analysis**, and **K-nearest neighbors**.

Default Data

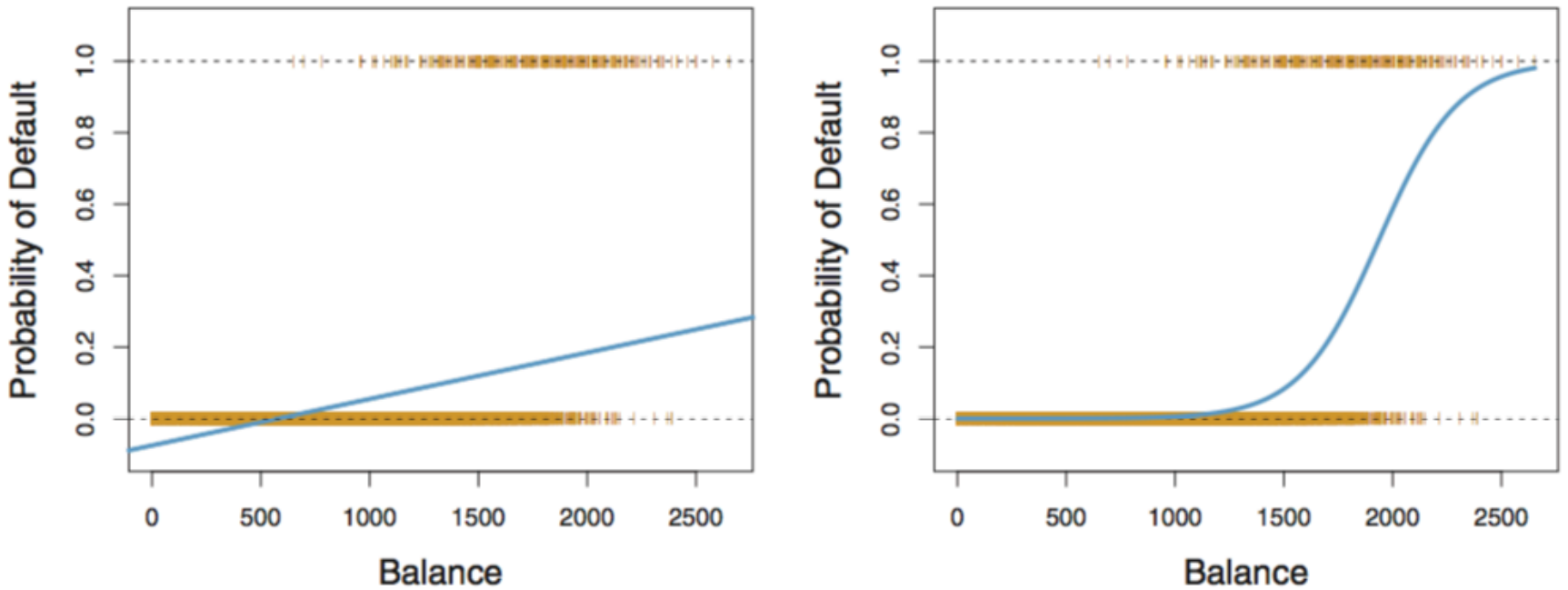


It shows the annual incomes and monthly credit card balances of a number of individuals. The individuals who defaulted on their credit card payments are shown in orange, and those who did not are shown in blue.

Why Not Linear Regression?

- In this case of a binary outcome, linear discriminant analysis, which is related but different from linear regression, does a good job as a classifier.
- The least squared method can estimate $E(Y|X = x) = Pr(Y = 1|X = x)$, we might think that regression is perfect for this task.
- However, linear regression might produce probabilities less than zero or bigger than one. Logistic regression is more appropriate.

Linear versus Logistic Regression



Left: Estimated probability of default using linear regression. Some estimated probabilities are negative! The orange points represents the 0/1 values coded for default (No or Yes). Right: Predicted probabilities of default using logistic regression. All probabilities lie between 0 and 1.

Logistic Regression

- How to model the relationship between $p(X) = Pr(Y = 1|X)$ and X?
- A linear regression may estimate $p(X) < 0$ or $p(X) > 1$.
- Logistic regression model $p(X)$ by the logistic function,

$$p(X) = \frac{e^{\beta_0+\beta_1 X}}{1 + e^{\beta_0+\beta_1 X}},$$

  It is easy to see that no matter what values β_0, β_1, or X take, $p(X)$ will have values between 0 and 1.

- Note that $p(X)$ is Not a linear function X or β.

Logistic Regression

- A bit of rearrangement gives

$$\underbrace{\frac{p(X)}{1-p(X)}}_{odds} = e^{\beta_0+\beta_1 X}, \quad \underbrace{\log\left[\frac{p(X)}{1-p(X)}\right]}_{log-odds} = \beta_0 + \beta_1 X.$$

  The odds takes value between 0 and $+\infty$, and the log odds takes value between $-\infty$ and $+\infty$.

- β_1 represents the change of log odds by increasing X by one unit, since

$$\beta_1 = \log\left[\frac{p(X+1)}{1-p(X+1)}\right] - \log\left[\frac{p(X)}{1-p(X)}\right]$$

Maximum Likelihood

Given training data $(x_1, y_1), ..., (x_n, y_n)$, we use **maximum likelihood** to estimate the parameters.

The maximum likelihood principle is that we seek the estimates of parameters such that the fitted probability corresponds as closely as possible to the individual's observed outcome.

The **likelihood function** of the observed data is

$$\ell(\beta_0, \beta_1) = \prod_{i:y_i=1} p(x_i) \prod_{i:y_i=0} (1 - p(x_i)).$$

We pick β_0 and β_1 to maximize the likelihood of the observed data.

Most statistical packages can fit logistic regression models by maximum likelihood. In R we use the glm function.

Inference

Consider again the Default data (fitted by maximum likelihood):

	Coefficient	Std. error	Z-statistic	P-value
`Intercept`	−10.6513	0.3612	−29.5	<0.0001
`balance`	0.0055	0.0002	24.9	<0.0001

Z-statistic is similar to t-statistic in regression, and is defined as

$$\hat{\beta}_1 / SE(\hat{\beta}_1).$$

It produces p-value for testing the null hypothesis $H_0 : \beta_1 = 0$. A large (absolute) value of the z-statistic or small p-value indicates evidence against H_0.

Making Predictions

Consider the Default data with student as predictor. What is our estimated probability of default for a student? To fit the model, we create a dummy variable that takes on a value of 1 for students and 0 for non-students.

	Coefficient	Std. Error	Z-statistic	P-value
`Intercept`	-3.5041	0.0707	-49.55	< 0.0001
`student[Yes]`	0.4049	0.1150	3.52	0.0004

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431,$$

$$\widehat{\Pr}(\texttt{default=Yes}|\texttt{student=No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292.$$

Multiple Logistic Regression

Consider the Default data using balance, income, and student status as predictors.

$$\log\left(\frac{p(X)}{1-p(X)}\right) = \beta_0 + \beta_1 X_1 + \cdots + \beta_p X_p$$

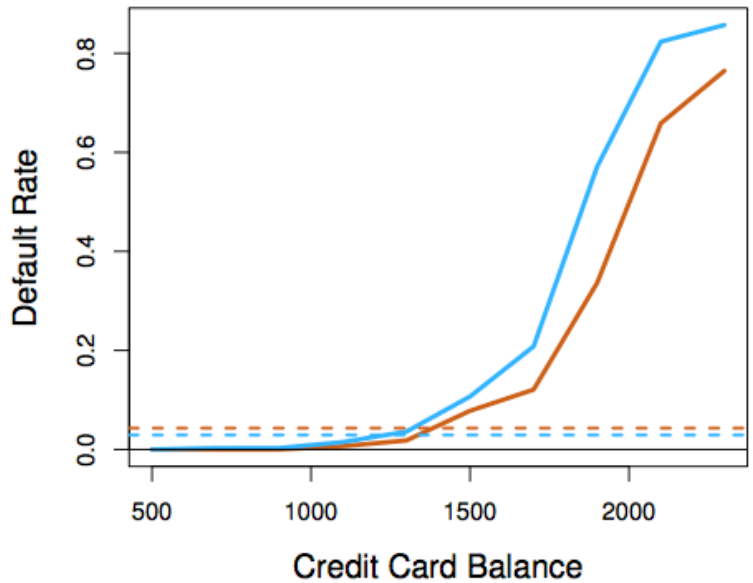
$$p(X) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}$$

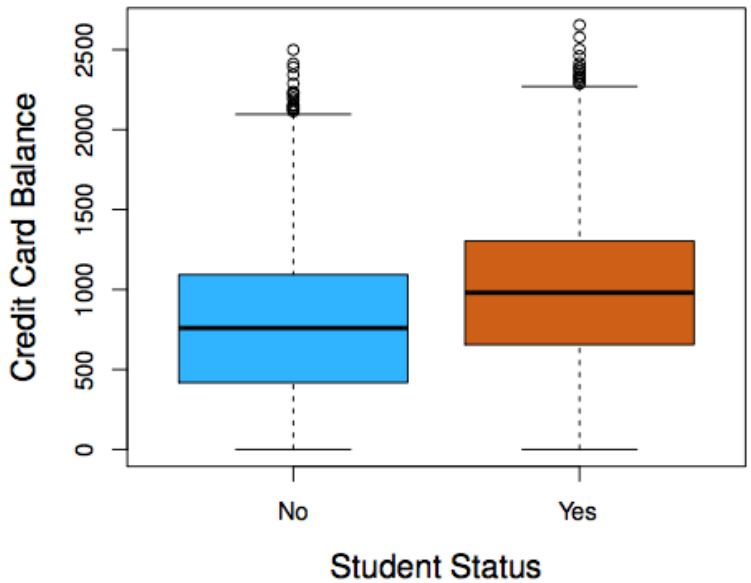
	Coefficient	Std. Error	Z-statistic	P-value
`Intercept`	-10.8690	0.4923	-22.08	< 0.0001
`balance`	0.0057	0.0002	24.74	< 0.0001
`income`	0.0030	0.0082	0.37	0.7115
`student[Yes]`	-0.6468	0.2362	-2.74	0.0062

Why is coefficient for student negative, while it was positive before?

Confounding

The results obtained using one predictor may be quite different from those obtained using multiple predictors, especially when there is correlation among the predictors. The phenomenon is known as **confounding**.





- Students tend to have higher balances than non-students, so their marginal default rate is higher than for non-students.
- But for each level of balance, students default less than non-students.
- Multiple logistic regression can tease this out.

Linear Discriminant Analysis

- Logistic regression involves directly modeling $P(Y = k|X = x)$.
- Here, **linear discriminant analysis** is to model the distribution of X in each of the classes separately, and then use Bayes' theorem to flip things around and obtain $P(Y = k|X = x)$. The Bayes' theorem is

$$P(Y = k|X = x) = P(X = x|Y = k)P(Y = k)/P(X = x).$$

- We usually assume the distribution of X in each of the classes to be normal distributions.

Why not Logistic Regression

- When the classes are well-separated, the parameter estimates for the logistic regression model are surprisingly unstable. Linear discriminant analysis does not suffer from this problem.
- If n is small and the distribution of the predictors X is approximately normal in each of the classes, the linear discriminant model is again more stable than the logistic regression model.
- Linear discriminant analysis is more convenient when we have more than two response classes.

Using Bayes' Theorem for Classification

- Recall that the Bayes' theorem is

$$P(Y = k|X = x) = P(X = x|Y = k)P(Y = k)/P(X = x).$$

- We can slightly rewrite it as

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

  π_k is the **prior** probability that a randomly chosen observation comes from the kth class, i.e. $P(Y = k)$.

  $f_k(X) = P(X = x|Y = k)$ denotes the **density function** of X for an observation that comes from the kth class.

  $p_k(x) = P(Y = k|X = x)$ is called **posterior** probability. It is the probability that the observation belongs to the kth class, given the predictor value for that observation.

- In the ideal case, we classify a new point according to which posterior probability is highest.

Linear Discriminant Analysis for $p = 1$

- We assume that $f_k(x)$ is normal, which takes the form

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

  where μ_k and σ_k^2 are the mean and variance parameters for the kth class. We assume all $\sigma_k = \sigma$ are the same.

- Plugging this into Bayes' formula, we get $p_k(x) = P(Y = k|X = x)$ as

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

Linear Discriminant Analysis for $p = 1$

- To classify at the value $X = x$, we need to see which k has the largest $p_k(x)$.
- Taking logs, and discarding terms that do not depend on k, the Bayes classifier is to assign x to the class with the largest

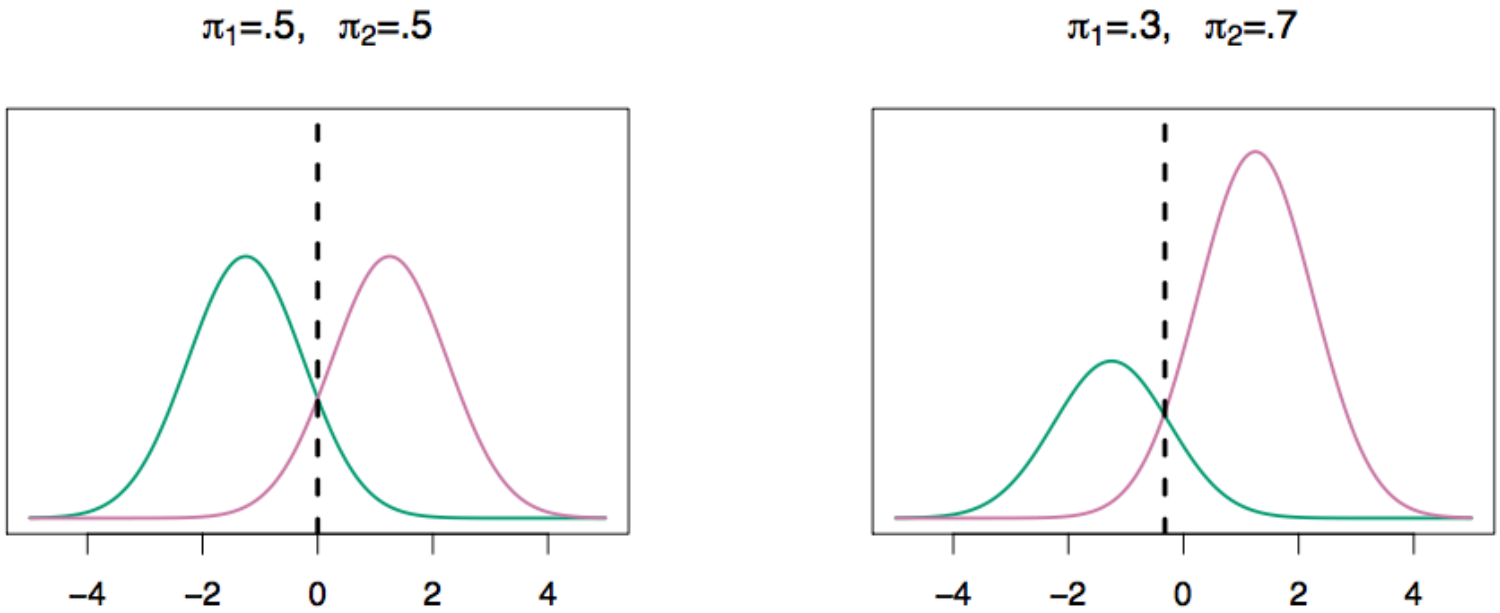
$$\delta_k(x) = x\frac{\mu_k}{\sigma^2} - \frac{\mu_k^2}{2\sigma^2} + \log \pi_k.$$

  Note that $\delta_k(x)$ is a linear function of x. That is why it is called linear discriminant analysis (LDA).

- If $K = 2$ and $\pi_1 = \pi_2$, then the Bayes decision boundary corresponds to

$$x = \frac{\mu_1 + \mu_2}{2}.$$

Example



There are $p_1(x)$ and $p_2(x)$ for classes 1 and 2 (green and red). The dashed vertical line represents the Bayes decision boundary. The examples have $\mu_1 = -1.5$, $\mu_2 = 1.5$, and $\sigma = 1$, and different values of π_1 and π_2.

Typically we don't know these parameters; we just have the training data. In that case we simply estimate the parameters and plug them into $\delta_k(x)$.

Discriminant functions

- Given training data, we estimate μ_k, and $\sigma = 1$ by

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

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

  where n_k is the number of training observations in the kth class. We also estimate π_k by

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

- Plugging the estimates into $\delta_k(x)$, we get

$$\hat{\delta}_k(x) = x\frac{\hat{\mu}_k}{\hat{\sigma}^2} - \frac{\hat{\mu}_k^2}{2\hat{\sigma}^2} + \log \hat{\pi}_k,$$

  which is called **discriminant function**.

- LDA just assigns x to the class with the largest $\hat{\delta}_k(x)$.

Linear Discriminant Analysis for $p > 1$

- We now extend the LDA classifier to the case of multiple predictors $X = (X_1, ..., X_p)$.
- Recall that the posterior probability has the form

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

- Now, we assume $X|Y = k$ follows a multivariate normal distribution $N(\mu_k, \Sigma)$,

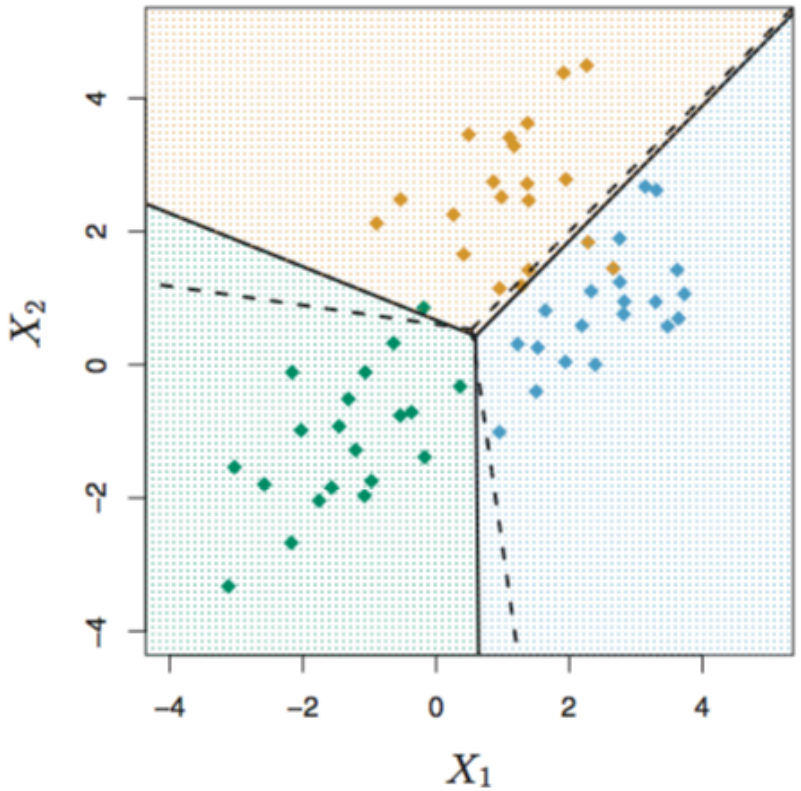
$$f_k(x) = \frac{1}{(2\pi)^{p/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T \Sigma^{-1}(x-\mu_k)}.$$

- Similarly, we assign x to the class with the largest

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

- The Bayes decision boundaries are the set of x for which $\delta_k(x) = \delta_l(x)$ for $k \neq l$. Again, the boundaries are collection of straight lines, since $\delta_k(x)$ is linear in x.

Example



There are three classes (orange, green and blue) with two predictors X_1 and X_2. Dashed lines are the Bayes decision boundaries. Solid lines are their estimates based on the LDA.

LDA on the Default Data

Our goal is predict whether or not an individual will default on the basis of credit card balance and student status.

		True default status		
		No	Yes	Total
Predicted default status	No	9,644	252	9,896
	Yes	23	81	104
	Total	9,667	333	10,000

The training error rate is $(23 + 252)/10000 = 2.75\%$. For a credit card company that is trying to identify high-risk individuals, an error rate of $252/333 = 75.7\%$ among individuals who default is unacceptable.