Classification Methods

Dr. Lan Du

Faculty of Information Technology, Monash University, Australia

FIT5149 week 4

Outline

Classification

- Qualitative (categorical) variables take values in an unordered set $\mathcal{C}$, for example,
  - email: spam and ham
  - breast tumor diagnosis: benign, malignant
  - sentiment: positive, negative, or neutral
- Classification: given a feature vector X and a categorical response $Y \in \mathcal{C}$, the classification task is to build a function $f(X)$ that takes input the feature vector X and predicts its value for Y. For example,
  - To determine whether or not a breast tumor is benign on the basis of features computed from a digitised image of a fine needle aspirate (FNA) of a breast mass.
  - To predict the sentiment of a product reviews on the basis of the review content (i.e., sequences of words).
  - To predict whether or not an individual will default on his or her credit card payment, on the basis of gender, education, age, history of past payment, etc.
- Methods: classify an observation based on the predicted probability of each of the categories of a qualitative variable.

Outline

Example: Credit Card Default Data

- Predict credible or not credible clients based on how likely the customers are to default.
- Possible predictors X:
  - Annual income
  - Monthly credit card balance
- The response variable default Y is categorical: Yes and NO.
- Questions:
  - How to check the relationship between Y and X?
  - How to build a model to predict default (Y) for any give value of balance (X_1) and income (X_2)?

Example: Credit Card Default Data

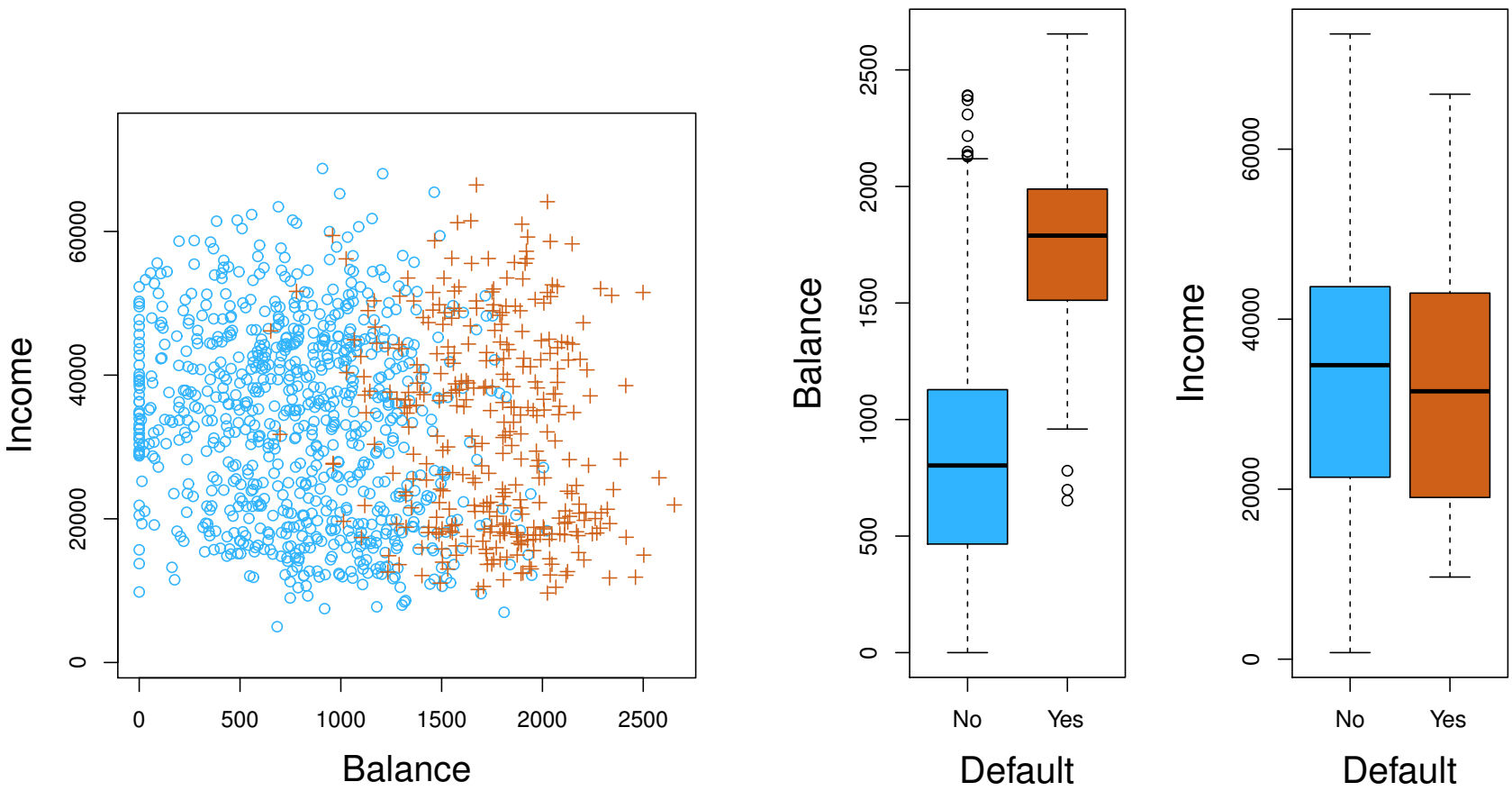


Figure: The Default dataset in ISL.

Can We Use Linear Regression?

- Surppose that we use the dummy variable approach to code the response variable:

$$\text{default} \quad Y = \begin{cases} 0 & \text{If No} \\ 1 & \text{If Yes} \end{cases}$$

- Can we fit a linear regression of Y on X and predict default if $\hat{Y} > 0.5$?

Can We Use Linear Regression?

- Surppose that we use the dummy variable approach to code the response variable:

$$\text{default} \quad Y = \begin{cases} 0 & \text{If No} \\ 1 & \text{If Yes} \end{cases}$$

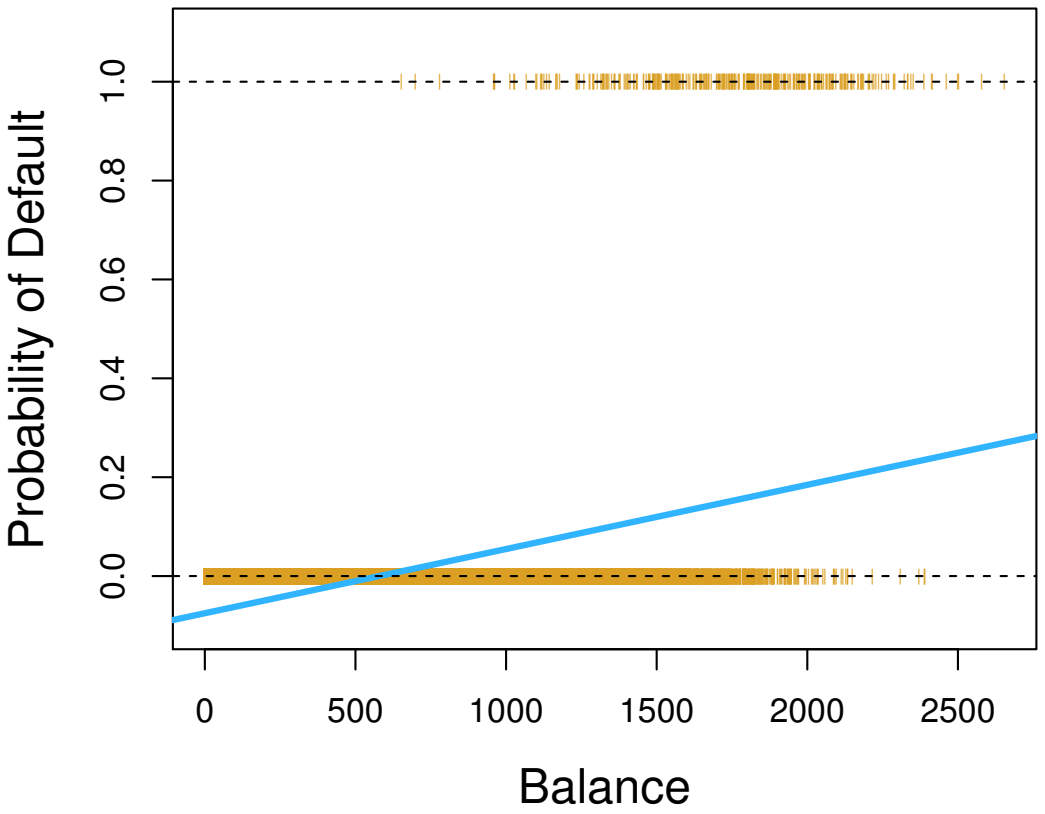
- Can we fit a linear regression of Y on X and predict default if $\hat{Y} > 0.5$?

$$default = \beta_0 + \beta_1 \times balance$$

  - In this case of a binary outcome, linear regression does a good job as a classifier, and is equivalent to linear discriminant analysis which we discuss later.

$$\mathbb{E}(default \mid balance) = p(default = yes \mid balance)$$

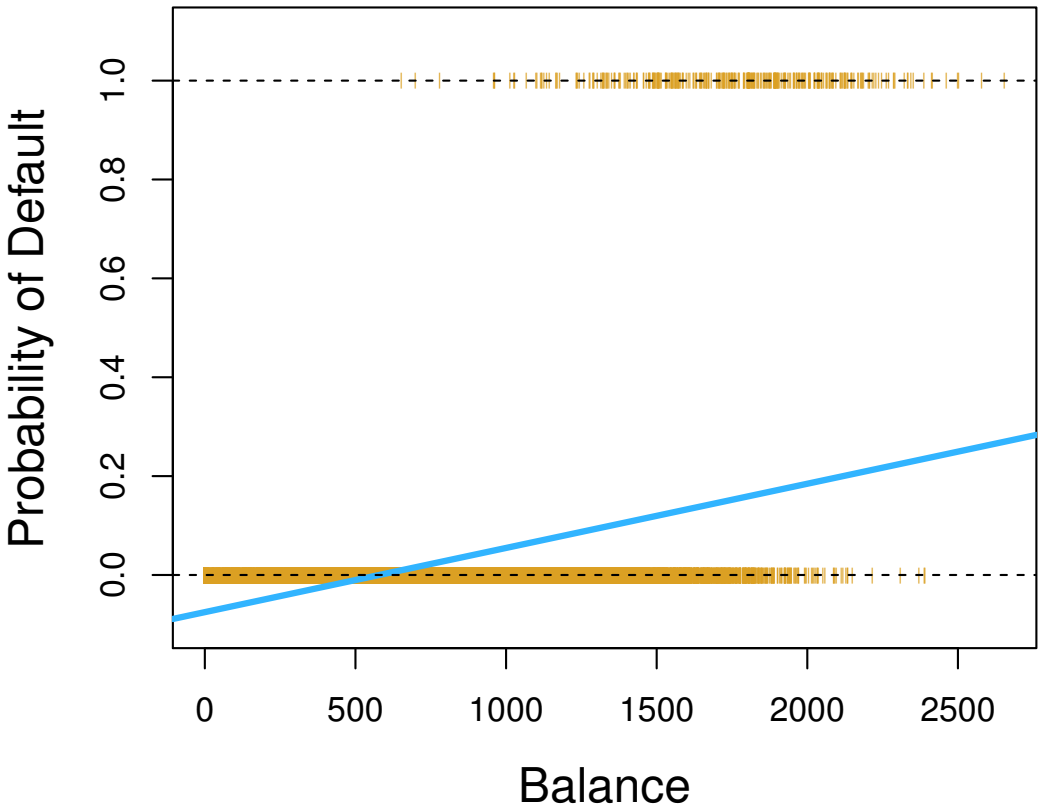
Why Not Linear Regression?



Problems:

- Linear regression might produce probabilities less than zero or bigger than one.

Why Not Linear Regression?



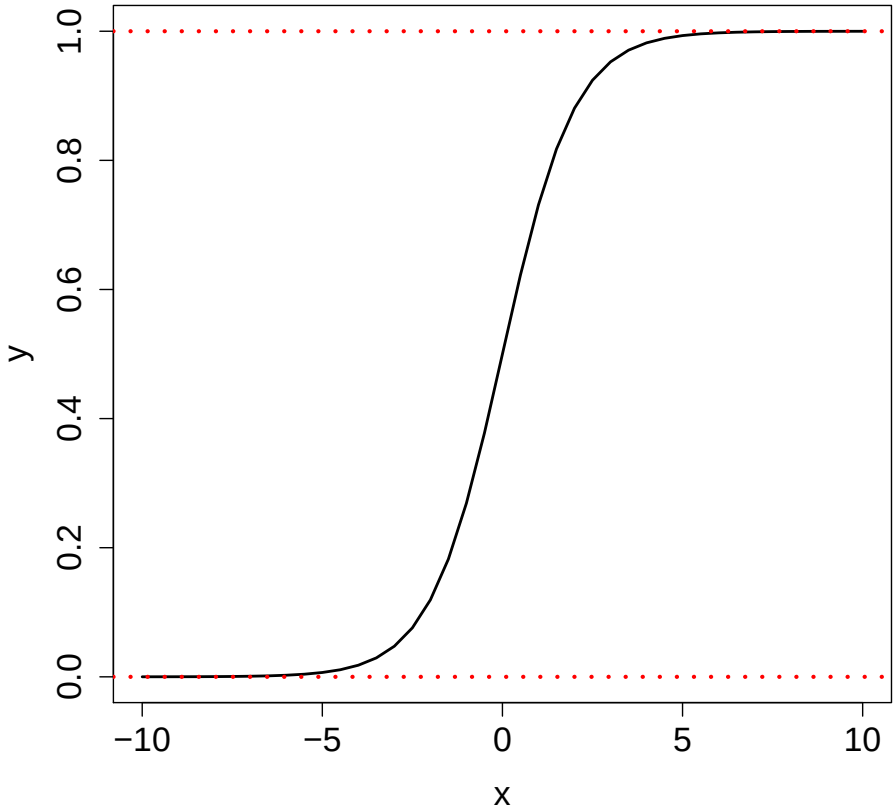
Problems:

- Linear regression might produce probabilities less than zero or bigger than one.
- For a response variable with three possible values,

$$\text{sentiment } Y = \begin{cases} 1 & \text{positive} \\ 2 & \text{negative} \\ 3 & \text{neutral} \end{cases}$$

This coding implies an ordering on the outcomes.

Solution: Logistic Function



- Logitic function:

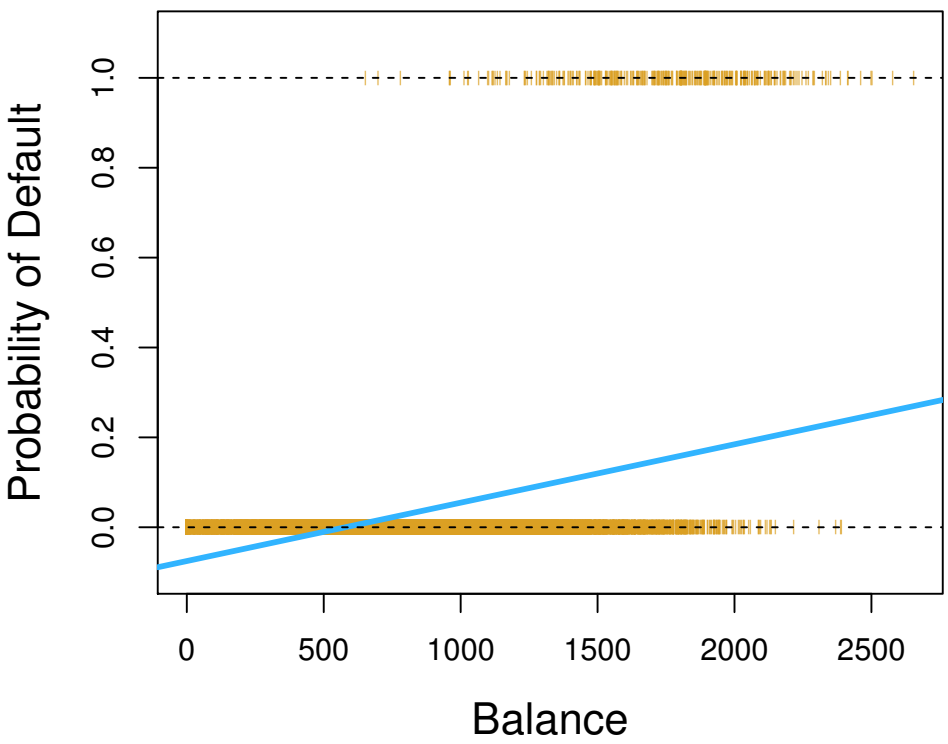
$$f(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}$$

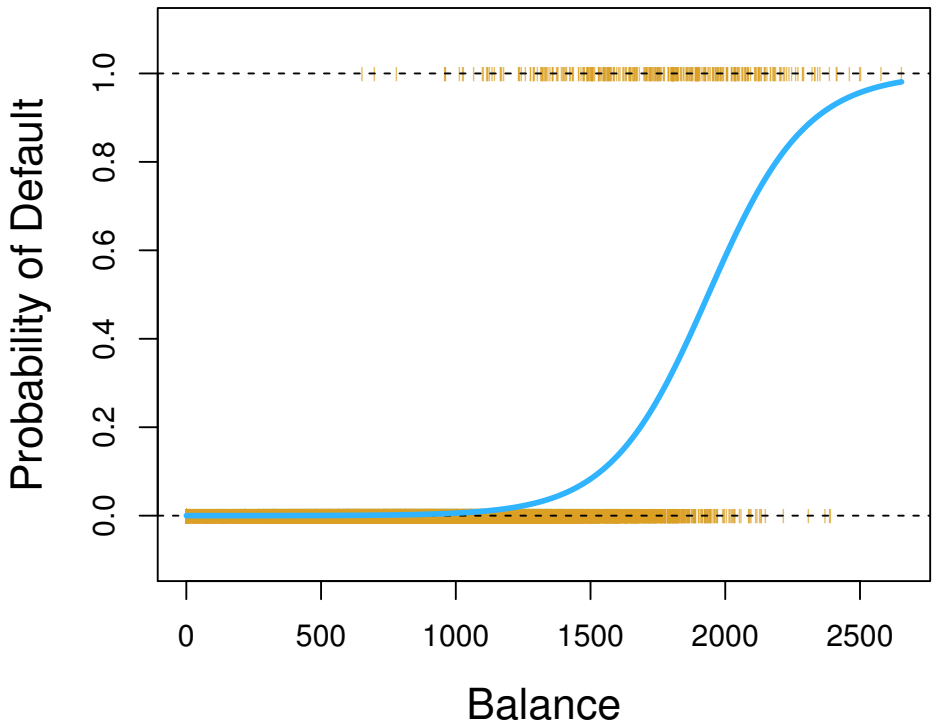
- Other similar functions:
  - hyperbolic tangent function

Logistic Regression

- Logistic regression on the Default data set uses the following form

$$p(\textit{default} = \textit{Yes} \mid \textit{balance}) = \frac{e^{\beta_0+\beta_1\times\textit{balance}}}{1+e^{\beta_0+\beta_1\times\textit{balance}}}$$





Logistic regression ensures that our estimate for $p(\textit{default} = \textit{Yes})$ is between 0 and 1

Logistic Regression

- Logistic regression on the Default data set uses the following form

$$p(\textit{default} = \textit{Yes} \mid \textit{balance}) = \frac{e^{\beta_0 + \beta_1 \times \textit{balance}}}{1 + e^{\beta_0 + \beta_1 \times \textit{balance}}}$$

- After a bit of manipulation, we have the odds:

$$\frac{p(\textit{default} = \textit{Yes} \mid \textit{balance})}{1 - p(\textit{default} = \textit{Yes} \mid \textit{balance})} = e^{\beta_0 + \beta_1 \times \textit{balance}}$$

Logistic Regression

- Logistic regression on the Default data set uses the following form

$$p(default = Yes \mid balance) = \frac{e^{\beta_0 + \beta_1 \times balance}}{1 + e^{\beta_0 + \beta_1 \times balance}}$$

- After a bit of manipulation, we have the odds:

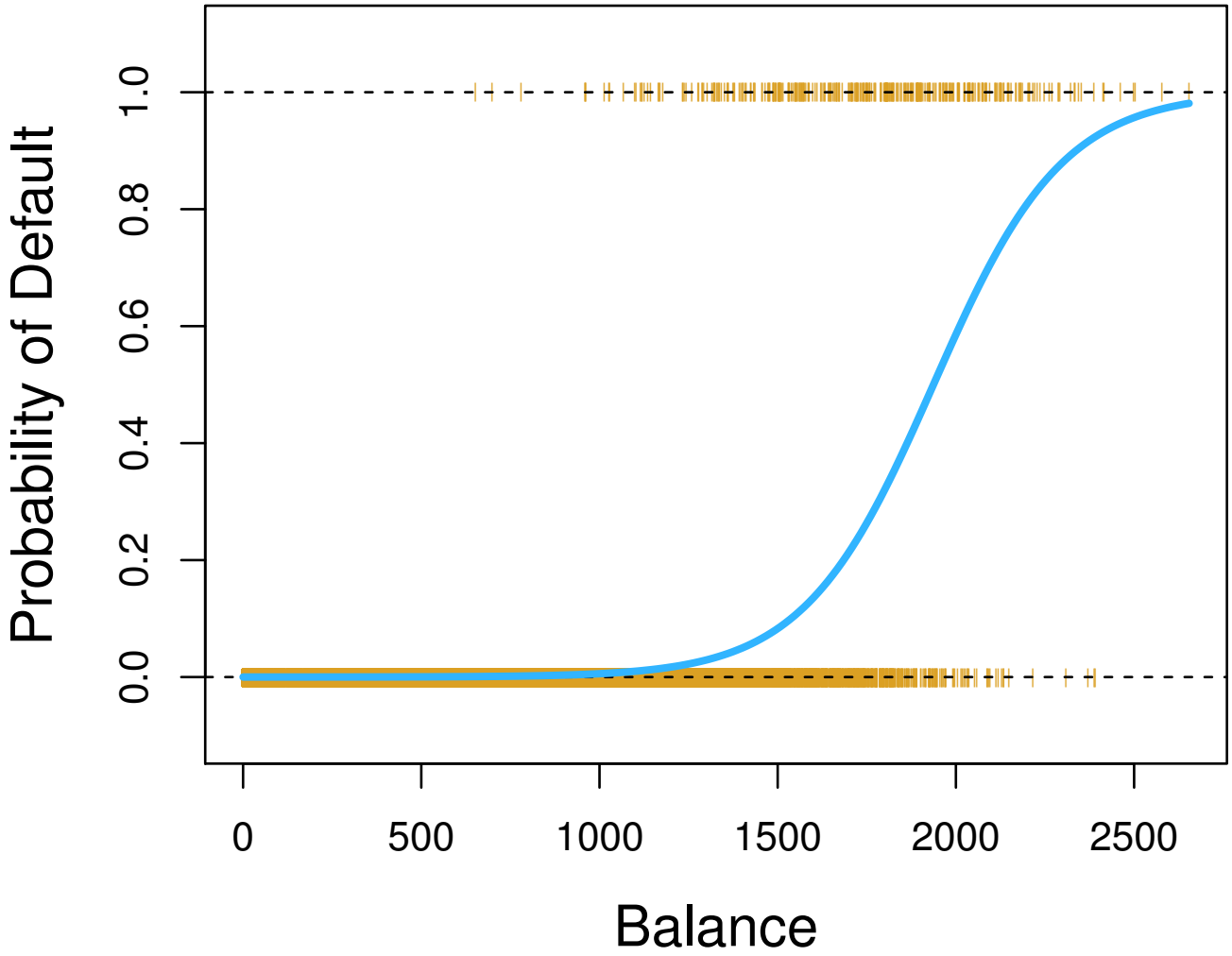
$$\frac{p(default = Yes \mid balance)}{1 - p(default = Yes \mid balance)} = e^{\beta_0 + \beta_1 \times balance}$$

- The logit transformation gives the following logit link function or log odds:

$$log\left(\frac{p(default = Yes \mid balance)}{1 - p(default = Yes \mid balance)}\right) = \beta_0 + \beta_1 \times balance$$

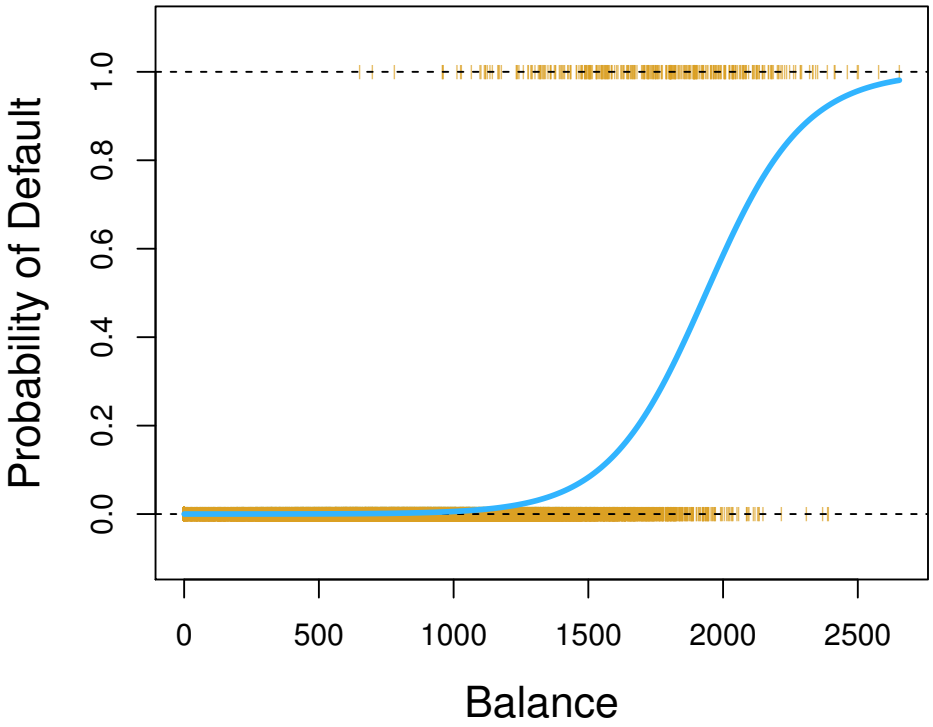
Interpreting the Coefficients

- In a simple linear regression model, β_1 is the slope of the regression line.
- In a logistic regression model, what is β_1?



Interpreting the Coefficients

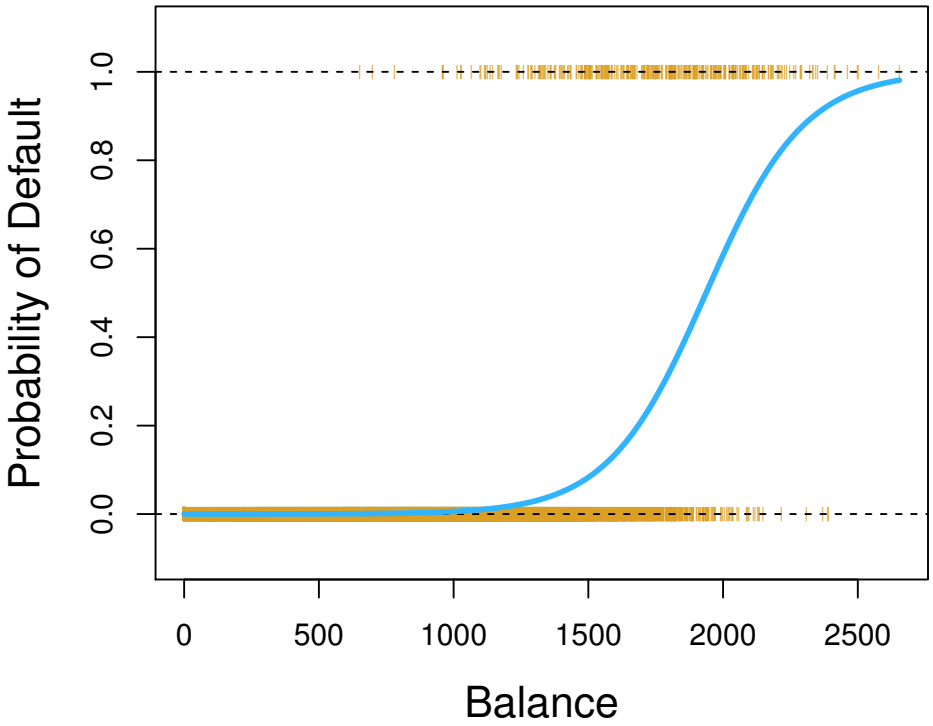
- In a simple linear regression model, β_1 is the slope of the regression line.
- In a logistic regression model, what is β_1?



- β_1 does not correspond to the change in $p(default = Yes \mid balance)$, why?

Interpreting the Coefficients

- In a simple linear regression model, β_1 is the slope of the regression line.
- In a logistic regression model, what is β_1?



- if $\beta_1 > 0$, increasing balance will increase $p(default = Yes \mid balance)$.
- if $\beta_1 < 0$, increasing balance will decrease $p(default = Yes \mid balance)$.

Estimating the Regressing Coefficients

- Use maximum likelihood to estimate the coefficients

$$\mathcal{L}(\beta_0, \beta_1) = \prod_{i:\mathit{default}_i = \mathit{Yes}} p(\mathit{default}_i \mid \mathit{balance}_i) \prod_{i:\mathit{default}_i = \mathit{No}} (1 - p(\mathit{default}_i \mid \mathit{balance}_i))$$

- This likelihood gives the probability of the observed zeros and ones in the data. We pick β_0 and β_1 to maximise the likelihood of the observed data.

$$\underset{\beta_0, \beta_1}{\text{argmax}}\, \mathcal{L}(\beta_0, \beta_1)$$

- Seek for β_0 and β_1 such that the predicted probability $\hat{p}(\mathit{default}_i \mid \mathit{balance}_i)$ for each $\mathit{balance}_i$ corresponds to as closely as possible to its observed $\mathit{default}_i$.

Assessing the Accuracy of the Coefficients

Similar to linear regression, we still want to perform a hypothesis test

$$H_0 : \beta_1 = 0$$
$$H_a : \beta_1 \neq 0$$

Instead of t-statistic, we use the z-statistic associate with $\hat{\beta}_1$; $\hat{\beta}_1 / SE(\hat{\beta}_1)$

- The logistic regression model that predicts the probability of default using balance

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

- The logistic regression model that predicts the probability of default using student

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

Make Predictions

The logistic regression model that predicts the probability of default using balance

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

- What is the estimated probability of default for someone with a balance of \$1,000?

$$\hat{p}(default = yes|balance = \$1,000) = \frac{e^{-10.6513+0.0055\times1000}}{1+e^{-10.6513+0.0055\times1000}} = 0.00576$$

Make Predictions

The logistic regression model that predicts the probability of default using student

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

- What is the estimated probability of default for a student?

$$\hat{p}(default = yes|student = true) = \frac{e^{-3.5041+0.4049\times 1}}{1 + e^{-3.5041+0.4049\times 1}} = 0.0431$$

Multiple Logistic Regression

- Considering predicting a binary response using multiple predictors, we can generalise the one-variable logistic regression as folllows:

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

and the *logit* function is now

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

- For the Default data, the estimated coefficients of the multiple logistic regression model are

	Coefficient	Std. error	Z-statistics	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

Multiple Logistic Regression

The negative coefficient: students are less likely to default than non-students.

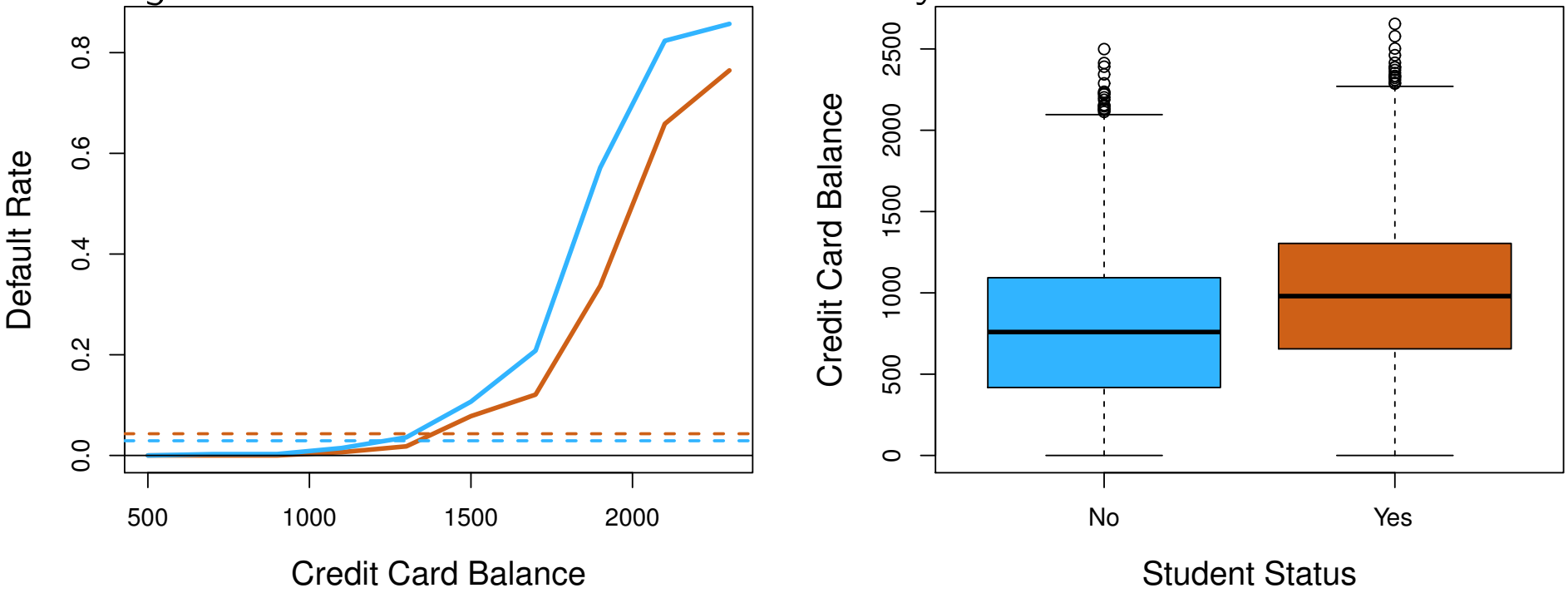


Figure: : students (orange), non-student (blue)

- 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.

Logistic regression with more than two classes

- Classify a response variables that has more than two classes: (one version used in the R package glmnet)

$$p(Y = k \mid X) = \frac{e^{\beta_{k,0}+\beta_{k,1}X_1+\cdots+\beta_{k,p}X_p}}{\sum_{l=1}^{K} e^{\beta_{l,0}+\beta_{l,1}X_1+\cdots+\beta_{l,p}X_p}}$$

  where a linear function is associated with each classes.

- Multiclass logistic regression is also referred to as multinomial regression, or Maximum Entropy Model (MaxEnt).

Logistic regression with ordinal responses

- Since the response variable is categorised and ordered,

$$c(x) = \ln\left(\frac{P(Y \leq j \mid x)}{P(Y > j \mid x)}\right)$$

and

$$\ln\left(\frac{\sum p(Y \leq j \mid x)}{1 - \sum p(Y \leq j \mid x)}\right) = \alpha_j + \boldsymbol{\beta}\mathbf{x}$$

- Assumption
  - the proportional odds assumption or the parallel regression assumption: the relationship between each value pair of the response variable is the same

Example: MaxEnt and Ordinal logistic regression

Outline

Discriminant Analysis

- Discriminant analysis belongs to the branch of classification methods called generative modelling, where we try to estimate the within class density of X given the class label. Combined with the prior probability (unconditioned probability) of classes, the posterior probability of Y can be obtained by the Bayes formula.

$$\begin{aligned} p(Y = k \mid X = x) &= \frac{p(X = x, Y = k)}{p(X = x)} \\ &= \frac{p(X = x \mid Y = k)p(Y = k)}{\sum_l^K p(X = x \mid Y = l)p(Y = l)} \\ &= \frac{\pi_k f_k(x)}{\sum_l^K \pi_l f_l(x)} \end{aligned}$$

where

- $f_k(x) = p(X = x \mid Y = k)$ is the density for X in class k.
- $\pi_k = p(Y = k)$ is the prior probability for class k.

Discriminant Analysis

MONASH University

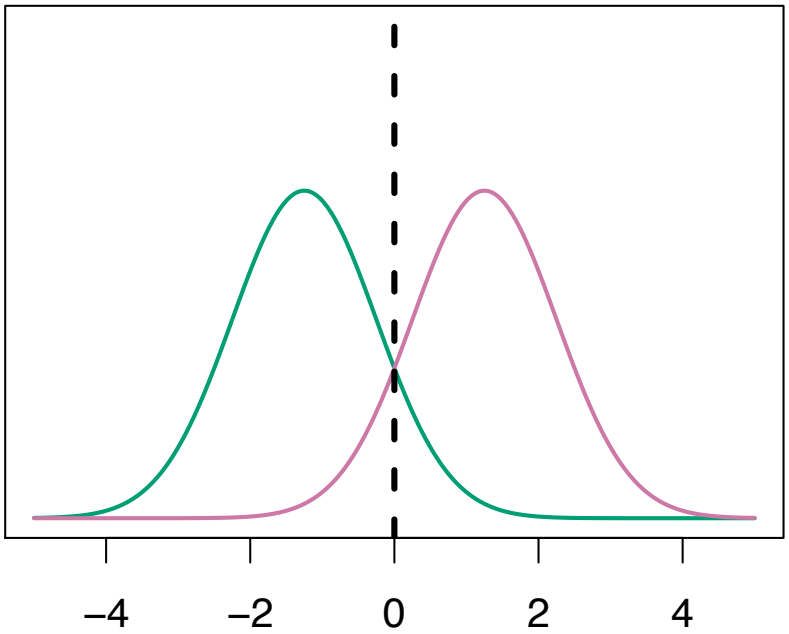
- Discriminant analysis belongs to the branch of classification methods called generative modelling, where we try to estimate the within class density of X given the class label. Combined with the prior probability (unconditioned probability) of classes, the posterior probability of Y can be obtained by the Bayes formula.

$$\begin{aligned} p(Y = k \mid X = x) &= \frac{p(X = x, Y = k)}{p(X = x)} \\ &= \frac{p(X = x \mid Y = k)p(Y = k)}{\sum_l^K p(X = x \mid Y = l)p(Y = l)} \\ &= \frac{\pi_k f_k(x)}{\sum_l^K \pi_l f_l(x)} \end{aligned}$$

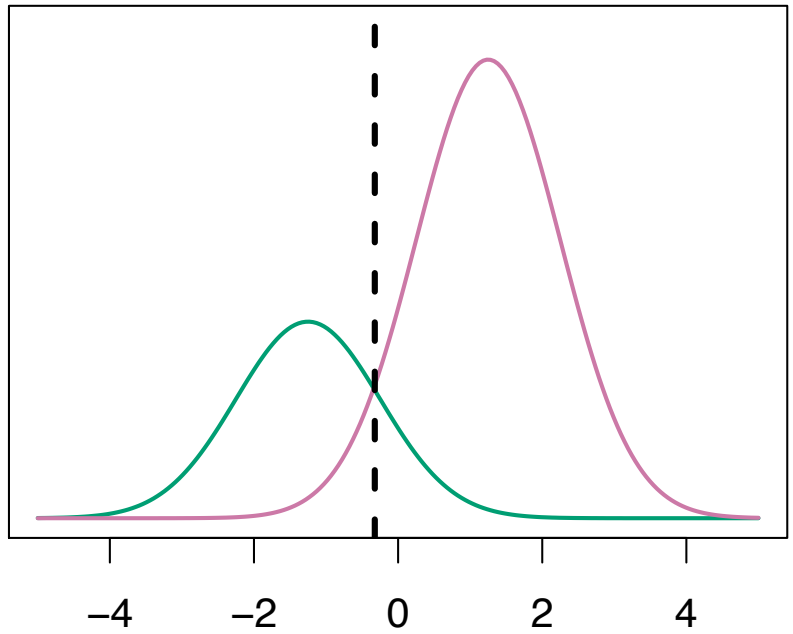
- Assume every density within each class is a Gaussian distribution:
  - Linear Discriminate Analysis (LDA)
    - Gaussian distributions for different classes share the same covariance structure.
  - Quadratic Discriminant Analysis (QDA):
    - No such a constraint on the covariance structure.

Classify to the Highest Density

π_1=.5, π_2=.5



π_1=.3, π_2=.7



- We classify a new point according to which density is highest

Linear Discriminant Analysis for $p = 1$

- In the one-dimensional setting, the Gaussian density has the form

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

  Here μ_k is the mean, and σ_k the variance in class k. We will assume that all the $\sigma_k = \sigma$ are the same.

Linear Discriminant Analysis for $p = 1$

- In the one-dimensional setting, the Gaussian density has the form

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

  Here μ_k is the mean, and σ_k the variance in class k. We will assume that all the $\sigma_k = \sigma$ are the same.

- Plugging this into Bayes formula, we get a rather complex expression for

$$\begin{aligned} p_k(x) &= p(Y = k \mid X = x) \\ &= \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_l^K \pi_l \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2}} \end{aligned}$$

Linear Discriminant Analysis for $p = 1$

- Plugging this into Bayes formula, we get a rather complex expression for

$$\begin{aligned} p_k(x) &= p(Y = k \mid X = x) \\ &= \frac{\pi_k \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_k}{\sigma}\right)^2}}{\sum_l^K \pi_l \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_l}{\sigma}\right)^2}} \end{aligned}$$

- To classify at the value $X = x$, we need to see which of the $p_k(x)$ is largest. Taking logs, and discarding terms that do not depend on k, we see that this is equivalent to assigning x to the class with the largest discriminant score:

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

Note that $\delta_k(x)$ is a linear function of x.

Linear Discriminant Analysis for $p = 1$

- To classify at the value $X = x$, we need to see which of the $p_k(x)$ is largest. Taking logs, and discarding terms that do not depend on k, we see that this is equivalent to assigning x to the class with the largest discriminant score:

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

Note that $\delta_k(x)$ is a linear function of x.
If there are $K = 2$ classes and $\pi_1 = \pi_2$,then one can see that the decision boundary is at

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

(The mathier students will recognize that the boundary is given by setting $\delta_1(x) = \delta_2(x)$.)

A Simple Example with $p = 1$

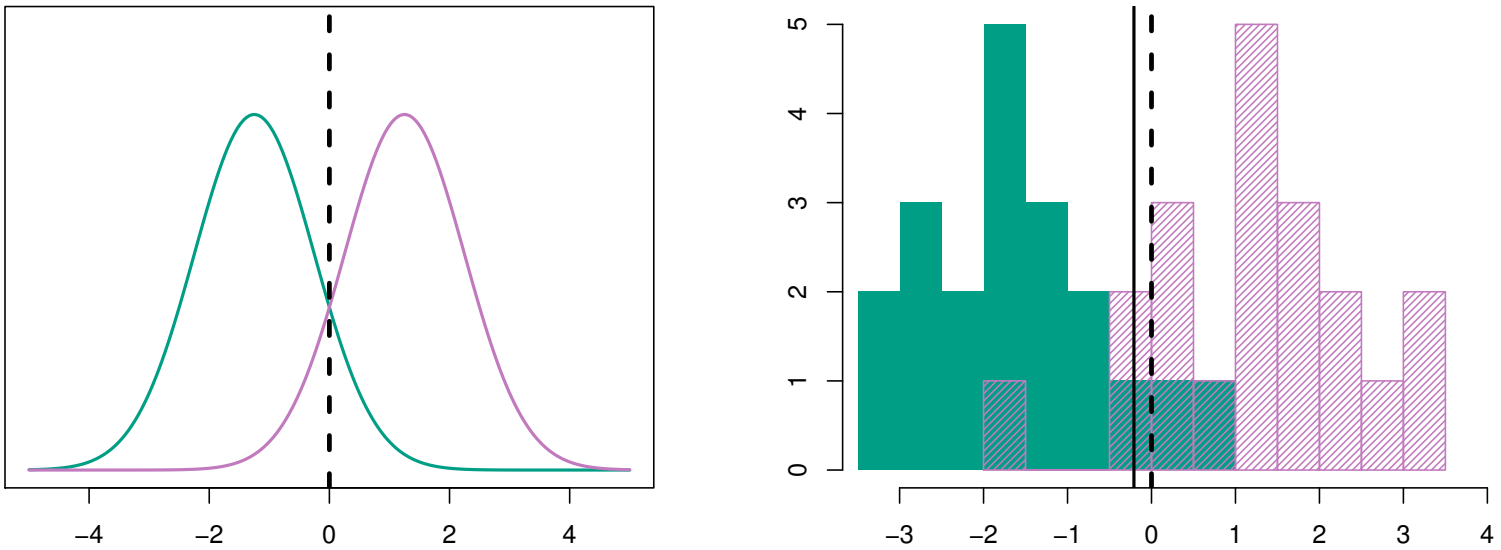


Figure: Example with $\mu_1 = -1.25$, $\mu_2 = 1.25$, $\pi_1 = \pi_2 = 0.5$,and $\sigma^2 = 1$.

- 20 observations were drawn from each of the two classes
- Error rates
  - Bayes' error rate: 10.6%
  - LDA error rate: 11.1%

A Simple Example with $p = 1$

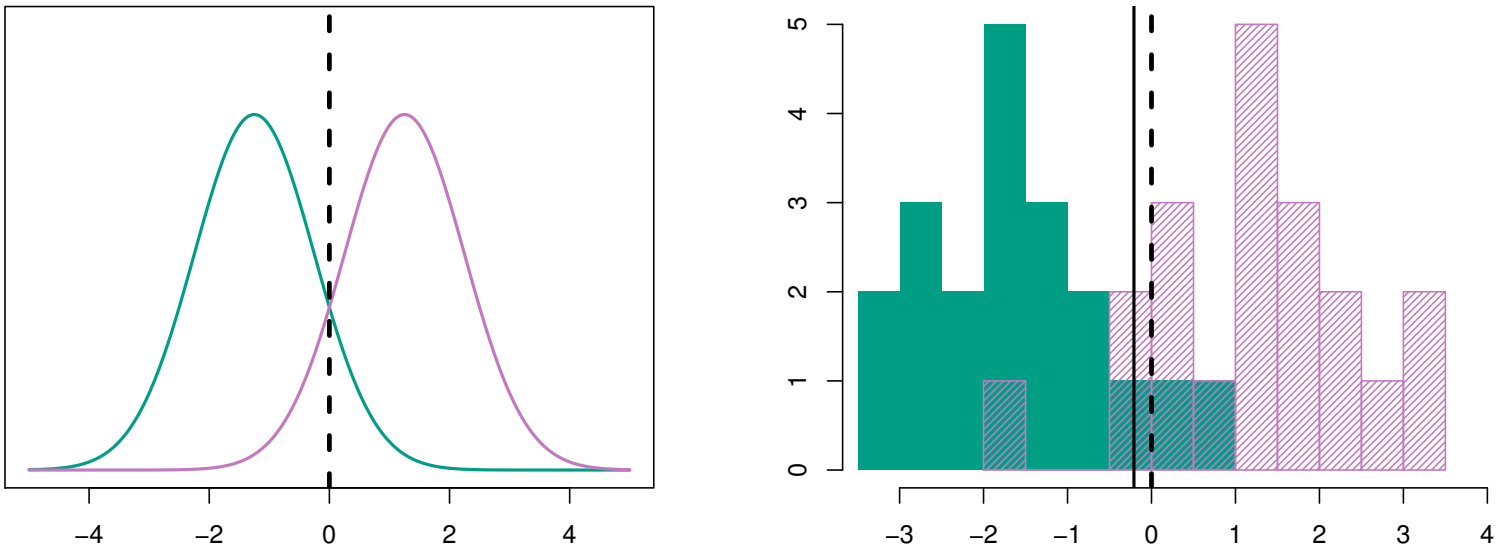
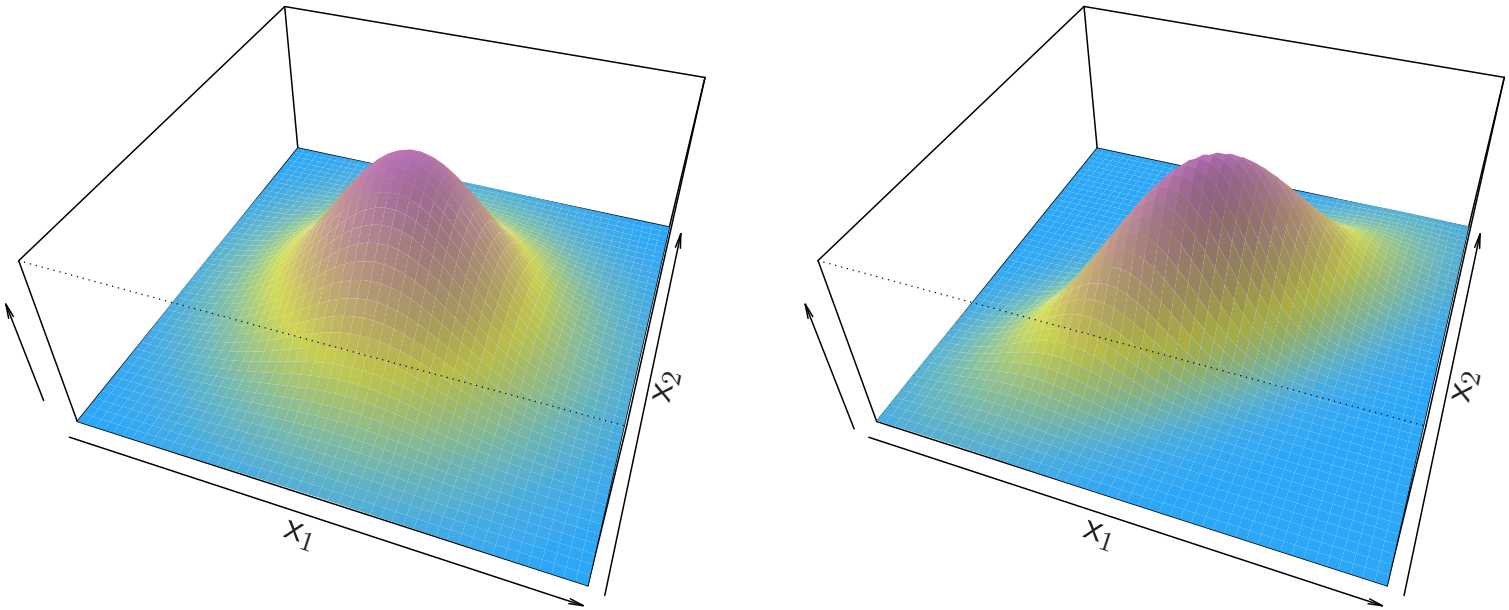


Figure: Example with $\mu_1 = -1.25$, $\mu_2 = 1.25$, $\pi_1 = \pi_2 = 0.5$,and $\sigma^2 = 1$.

- 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 the rule.

Linear Discriminate Analysis when $p > 1$



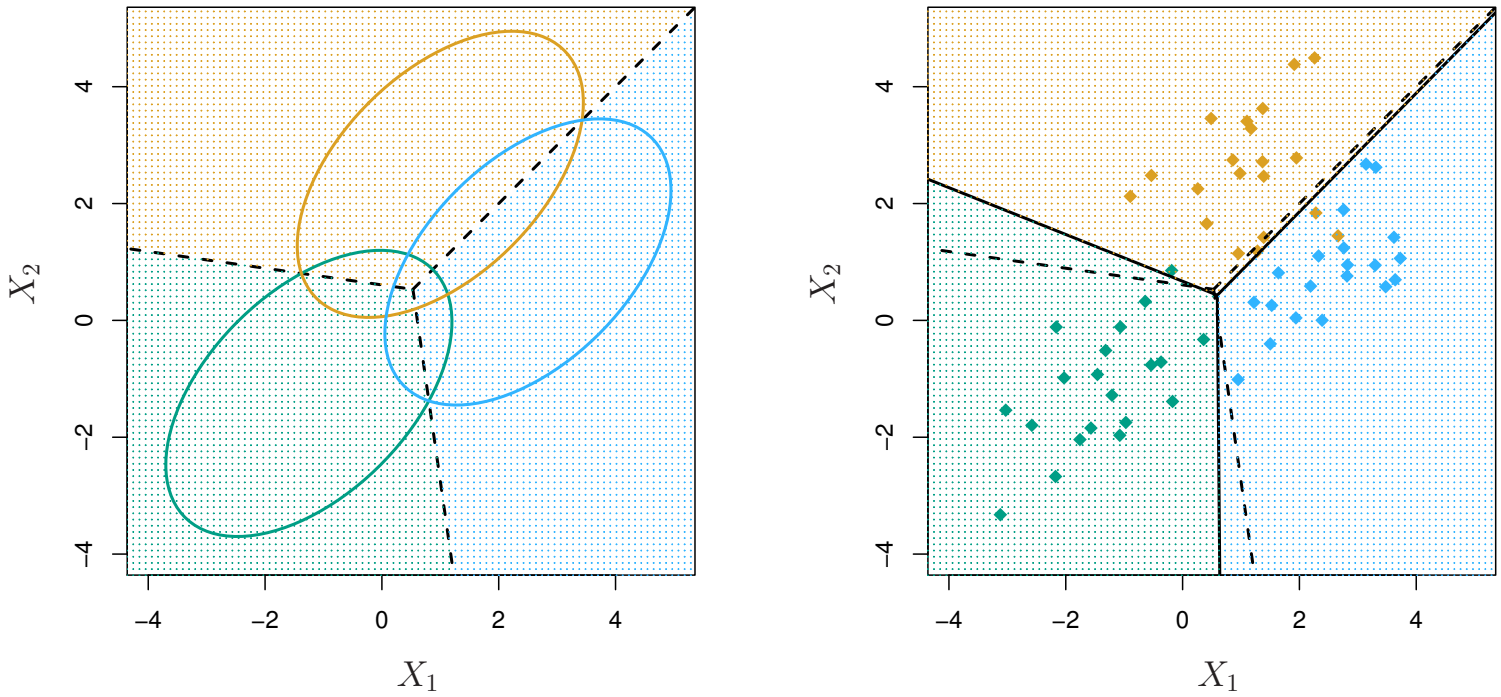
- The multivariate Gaussian density:

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

- The Discriminant function (a linear function):

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

Illustration: $p = 2$ and $K = 3$



- 20 observations were generated from each class
- Ellipses contain 95% of the probability for each of the three classes.
- The solid lines are Bayes decision boundaries
- The dashed lines are LDA decision boundaries

Example: LDA on Credit Data

		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

Figure: With classification threshold = 0.5, we receive $\frac{23+252}{10000}$ errors — a 2.75% misclassification rate!

Some caveats:

- This is training error, and we may be overfitting. Not a big concern here since $n = 10000$ and $p = 4$!
- If we classified to the prior — always to class No in this case — we would make 333/10000 = 3.33% errors.
- Of the true No's, we make 23/9667 = 0.2% errors; of the true Yes's, we make 252/333 = 75.7% errors.

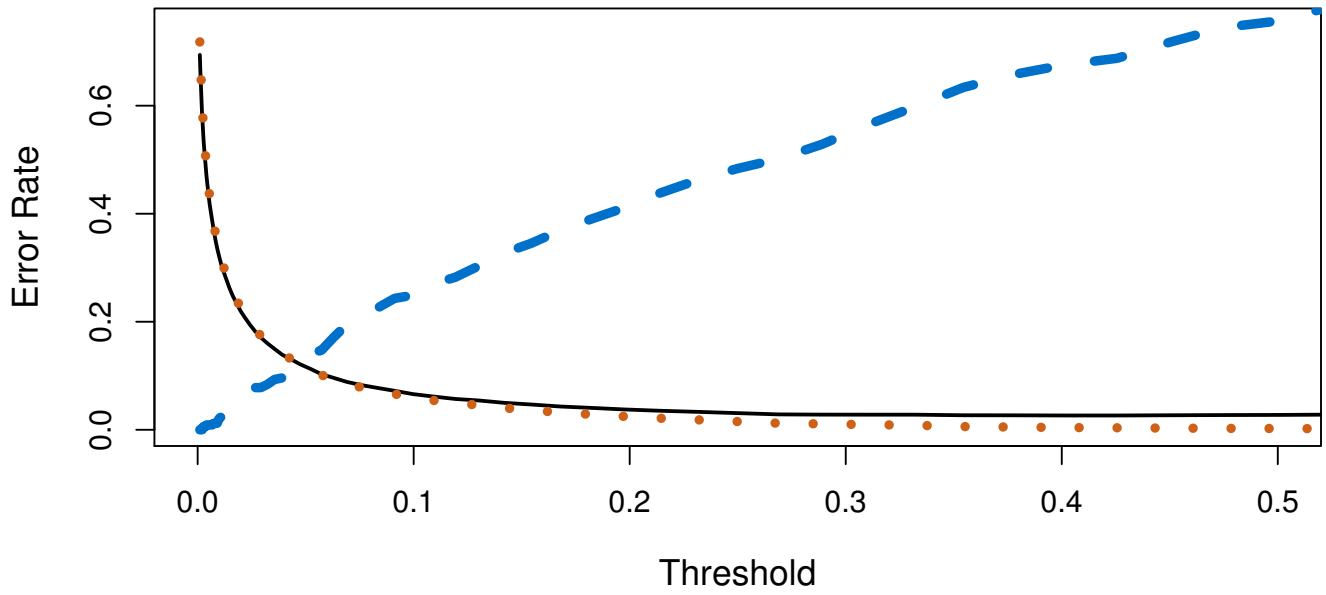
Example: LDA on Credit Data

		True default status		
		No	Yes	Total
Predicted default status	No	9,432	138	9,570
	Yes	235	195	430
	Total	9,667	333	10,000

Figure: With classification threshold = 0.2, we receive $\frac{235+138}{10000}$ errors — a 3.73% misclassification rate!

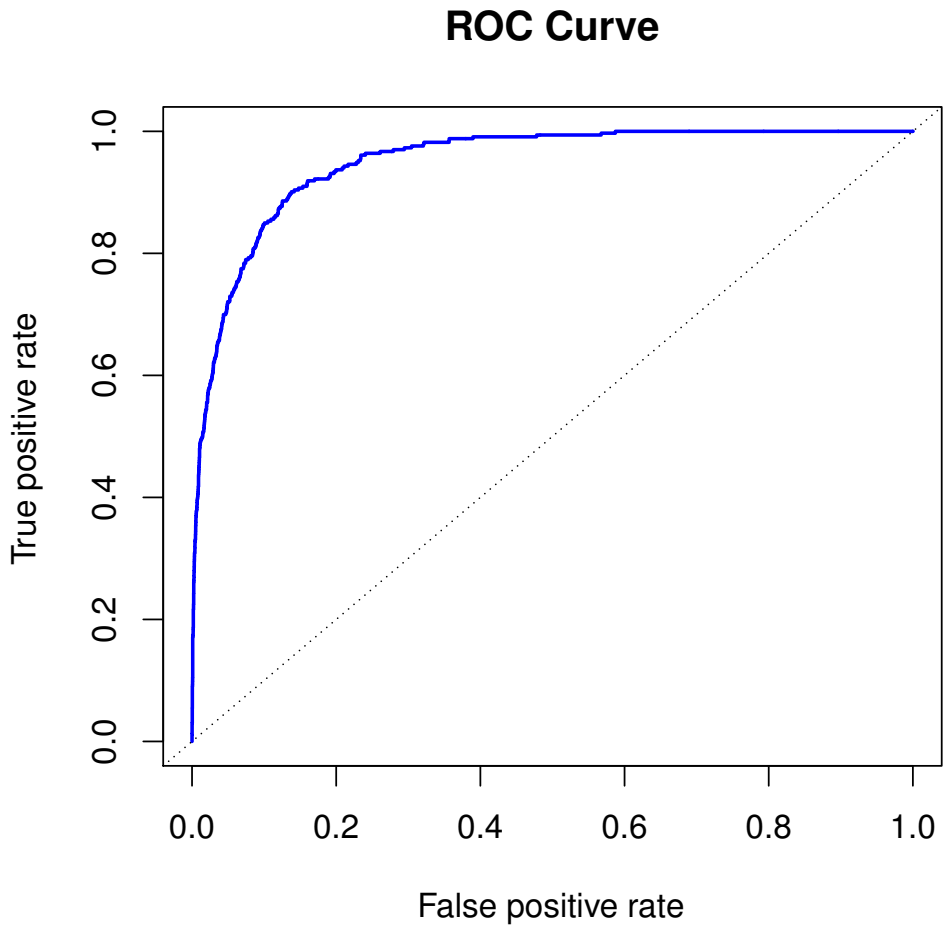
- LDA miss-predicted 138/333 = 41.4% of defaulters.
- Trade-off between overall error rate and the sensitivity (the percentage of true defaulters identified).

Varying the Classification Threshold



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

Receiver Operating Characteristics (ROC) Curve



- False positive rate: The fraction of negative samples that are mis-classified.
- True positive rate: The fraction of positive samples that are correctly classified.

Why Discriminate Analysis?

MONASH University

- 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 popular when we have more than two response classes

Possible Applications of LDA

- Bankruptcy prediction: In bankruptcy prediction based on accounting ratios and other financial variables, linear discriminant analysis was applied to systematically explain which firms entered bankruptcy vs. survived.
- Marketing: determine the factors which distinguish different types of customers and/or products on the basis of surveys or other forms of collected data.
- Biomedical studies: assess the severity state of a patient and prognosis of disease outcome. Examples:
  - Use results of clinical and laboratory analyses to built discriminant functions to classify disease in a future patient into mild, moderate or severe form.

Example: LDA on Fisher's Iris data

Outline

Quadratic Discriminant Analysis

MONASH University

- The posterior probability:

$$p(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_l^K \pi_l f_l(x)}$$

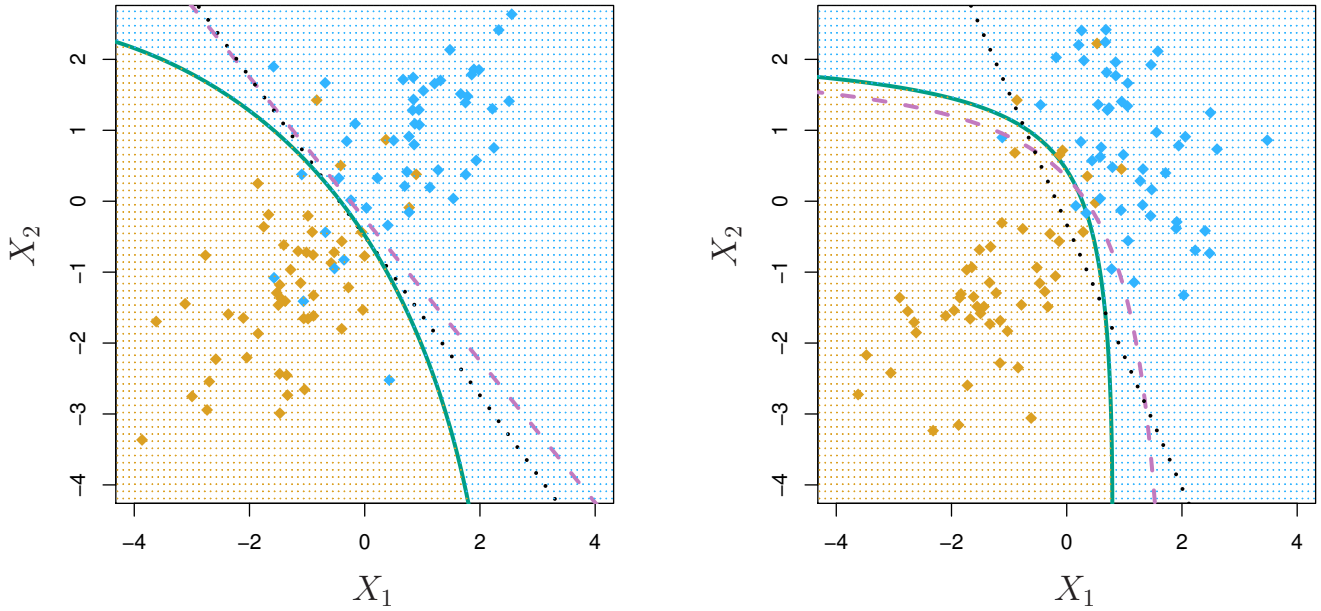
  where $f_k(x)$ are Gaussian densities.
- LDA: the same covariance matrix Σ in each class.

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

- QDA: different covariance matrix Σ_k in each class.

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

Quadratic Discriminant Analysis



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

Logistic Regression v.s. LDA

For a two-class problem, one can show that for LDA

$$log\left(\frac{p_1(x)}{1-p_1(x)}\right) = log\left(\frac{p_1(x)}{p_2(x)}\right) = c_0 + c_1x_1 + \ldots, + c_px_p$$

- Similarity: Both logistic regression and LDA produce linear boundaries.
- Difference in how the parameters are estimated
  - Logistic regression uses the conditional likelihood based on $p(Y|X)$, known as discriminative learning
  - LDA uses the full likelihood based on $p(X, Y)$ (known as generative learning).
- Note: LDA would do better than Logistic Regression if the assumption of normality hold, otherwise logistic regression can outperform LDA.

Summary

- Logistic regression is very popular for classification, especially when $K = 2$.
- LDA is useful when n is small, or the classes are well separated, and Gaussian assumptions are reasonable. Also when $K > 2$.
- Hit: if the decision boundary is
  - Linear: LDA and Logistic outperforms
  - Moderately Non-linear: QDA outperforms.
  - More complicated: KNN is superior.
- See Section 4.5 for some comparisons of logistic regression, LDA and KNN.

Reference

- Reading material:
  - "Classification", Chapter 4 of "Introduction to Statistical Learning", 6th edition
- Some figures in this presentation were 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
- Some of the slides are reproduced based on the slides from T. Hastie and R. Tibshirani