Statistical Learning for Data Science

Lecture 13

唐晓颖

电子与电气工程系

南方科技大学

April 12, 2023

Classification

- Qualitative (categorical) responses taking values in an unordered set

  Example:

  $$\text{eye color} \in \{\text{brown, green, blue}\}$$

  $$\text{email} \in \{\text{spam, email}\}$$

- Classification

  Predicting a qualitative response for an observation; it involves assigning the observation to a category, or class.

  Often, we are more interested in estimating the *probabilities* that the observation belongs to each category in the set. For example, it is more valuable to have an estimate of the probability that an insurance claim is fraudulent, than a classification fraudulent or not.

Classification

Three classifiers:

1. Logistic regression

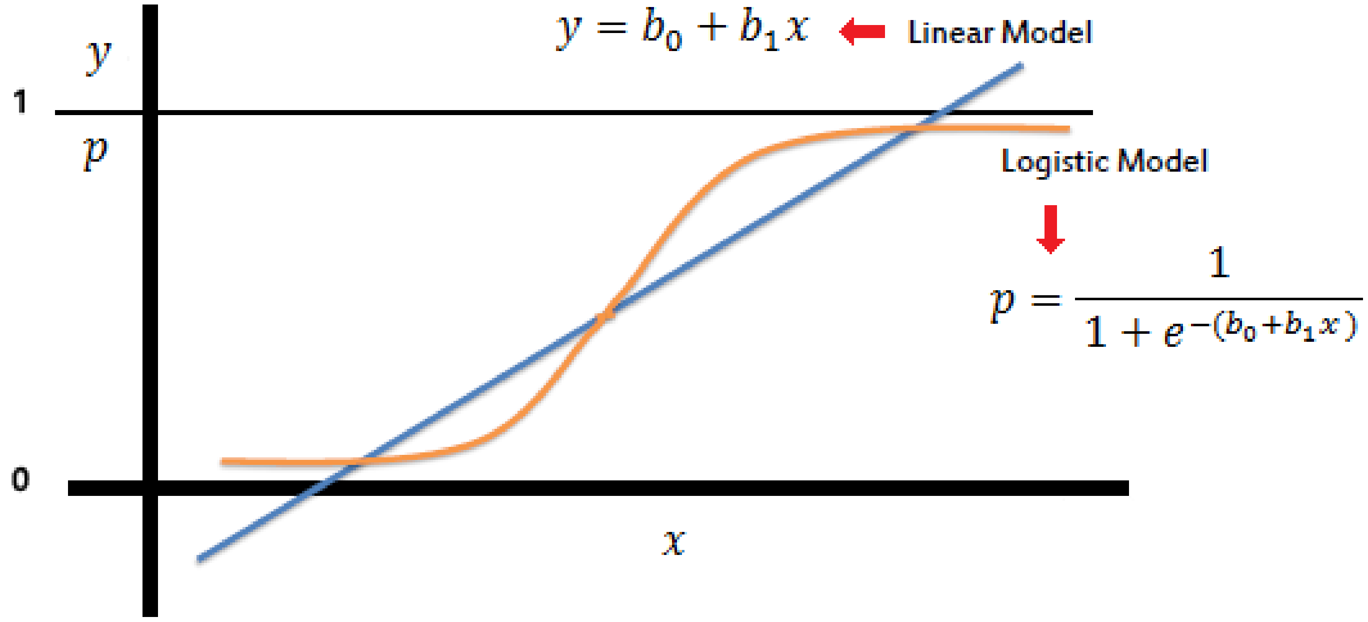
2. Linear discriminant analysis

3. K-nearest neighbors

Classification

Three classifiers:

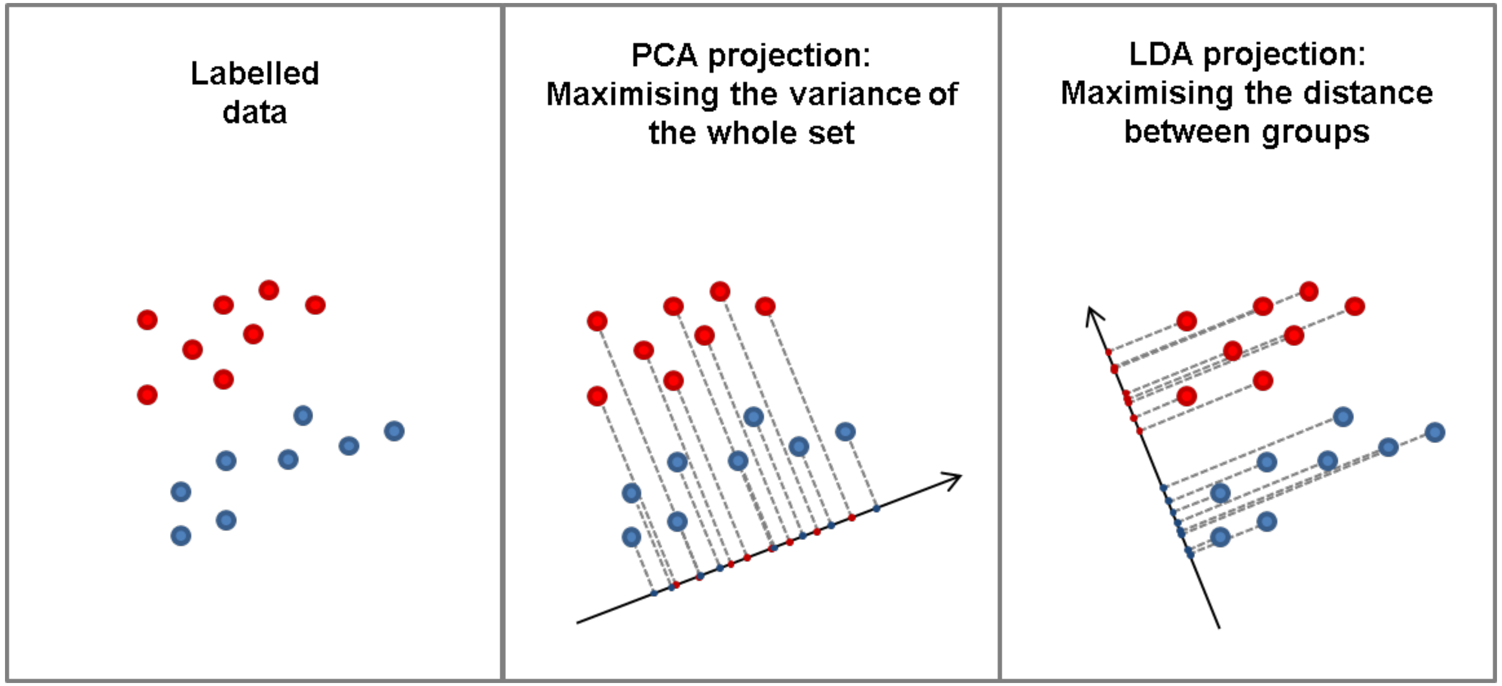
1. Logistic regression



Classification

Three classifiers:

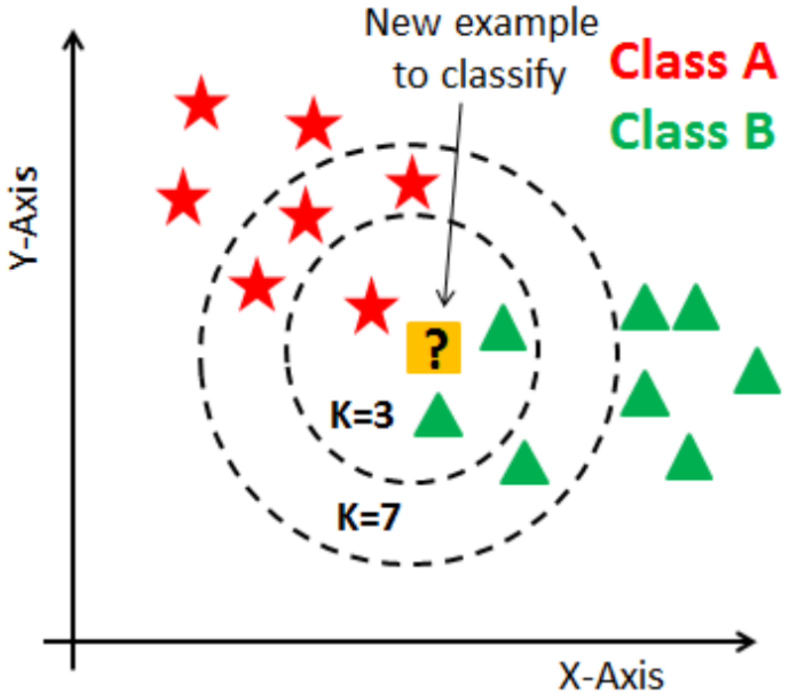
2. Linear discriminant analysis


Labelled data
PCA projection: Maximising the variance of the whole set
LDA projection: Maximising the distance between groups

Classification

Three classifiers:

3. K-nearest neighbors



An Overview of Classification

- Similar to regression, in classification, we have a set of training data $(x_1, y_1), \ldots, (x_n, y_n)$ that we can use to build a classifier.

- We want our classifier to perform well not only on the training data, but also on the test observations that were not used to train the classifier.

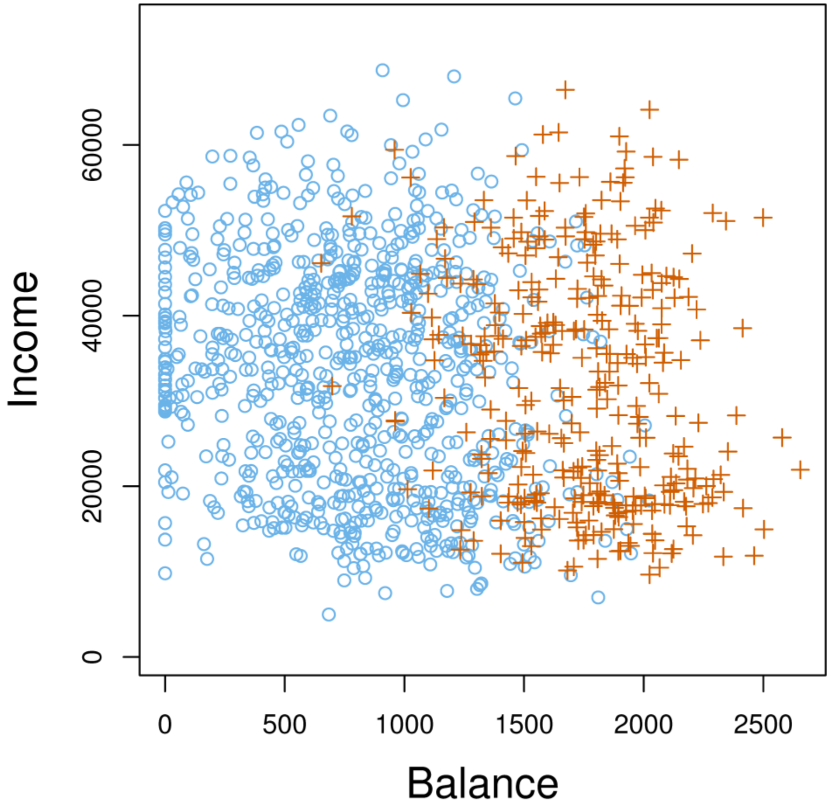
- Goal of classification: train a model that can accurately predict the class of new, unseen data based on its features

- Applications of classification: including image and speech recognition, spam filtering, credit scoring, and fraud detection, among others.

An Overview of Classification

Example: the default data set

Goal: we are interested in predicting whether an individual will default on his or her credit card payment, on the basis of annual income and monthly credit card balance.

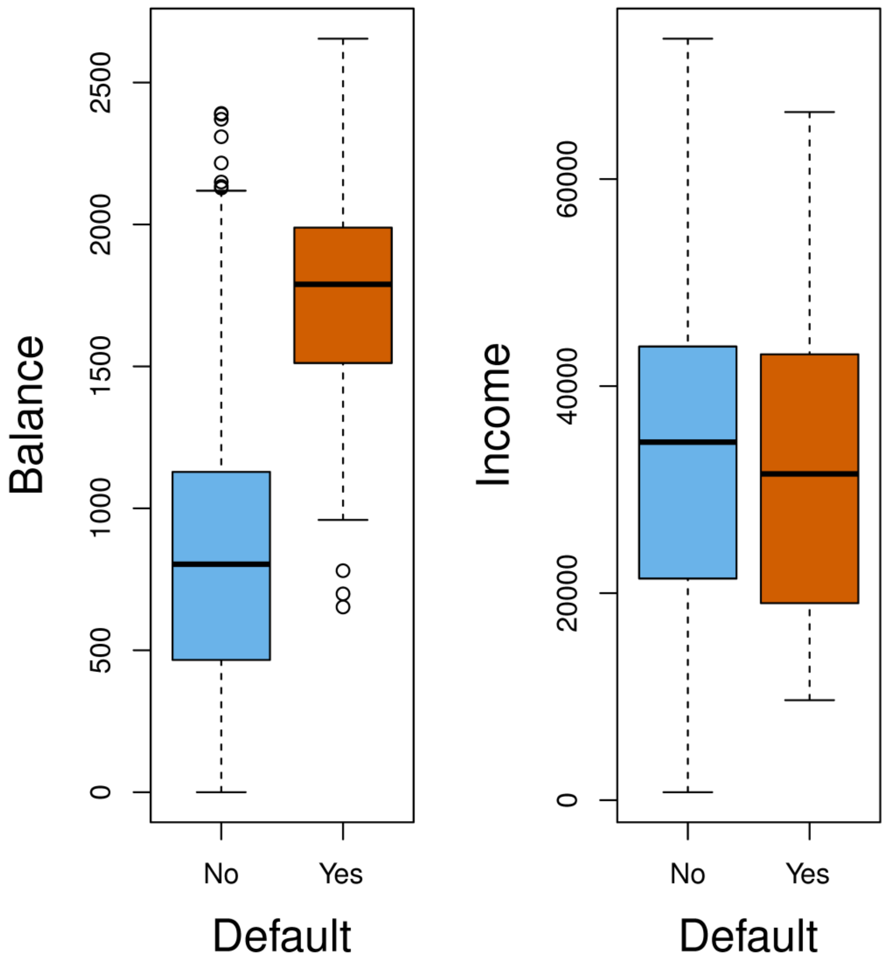


--orange: default
--blue: non-default

Pattern observed: individuals who defaulted tended to have higher credit card balances than those who did not.

An Overview of Classification

Example: the default data set



- We will learn how to build a model to predict default (Y) for any given value of balance (X_1) and income (X_2).

- There is a very pronounced relationship between balance and default in this example. In most real applications, the relationship between the predictor and the response will not be nearly so strong.

Can We Use Linear Regression

Example: the default data set

We code

$$Y = \begin{cases} 0, & \text{if no} \\ 1, & \text{if yes} \end{cases}$$

Can we simply perform a linear regression of Y on X and classify as Yes if $\hat{Y} > 0.5$?

- 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 will discuss later.

- Since in the population, $E(Y \mid X = x) = \Pr(Y = 1 \mid X = x)$ we might think that regression is perfect for this task.

Why we have $E(Y \mid X = x) = \Pr(Y = 1 \mid X = x)$?

Can We Use Linear Regression

Example: the default data set

We code

$$Y = \begin{cases} 0, & \text{if no} \\ 1, & \text{if yes} \end{cases}$$

Can we simply perform a linear regression of Y on X and classify as Yes if $\hat{Y} > 0.5$?

- However, if we use linear regression, some of our estimates might be outside the [0,1] interval, making them hard to interpret as probabilities. *Logistic regression* is more appropriate.

Can We Use Linear Regression

Now suppose we have a response variable with three possible values. A patient presents at the emergency room, and we must classify them according to their symptoms.

$$Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if drug overdose} \\ 3 & \text{if epileptic seizure} \end{cases}$$

This coding implies an ordering on the outcomes, putting drug overdose in between stroke and epileptic seizure, and insisting that the difference between stroke and drug overdose is the same as the difference between drug overdose and epileptic seizure. In practice, there is no particular reason that this needs to be the case.

Linear regression is not appropriate here. *Multiple Logistic Regression* and *Discriminant Analysis* are more appropriate.

Logistic Regression

Rather than modeling the response Y directly, logistic regression models the *probability* that Y belongs to a particular category.

Goal of logistic regression: predict a binary outcome, such as yes or no, true or false, or 0 or 1.

Key idea behind logistic regression: use a logistic function to model the probability of the binary outcome, given the input features.

The logistic function is an S-shaped curve that maps any input value to a probability value between 0 and 1.

Logistic Regression

Rather than modeling the response Y directly, logistic regression models the *probability* that Y belongs to a particular category.

Example: the default data set

Let $p(\text{balance}) = \Pr(\text{default} = \text{Yes} \mid \text{balance})$

To make predictions

$$\text{default} = \begin{cases} \text{Yes}, & p(\text{balance}) > 0.5 \\ \text{No}, & p(\text{balance}) \le 0.5 \end{cases}$$

To make conservative predictions in individuals who are at risk for default

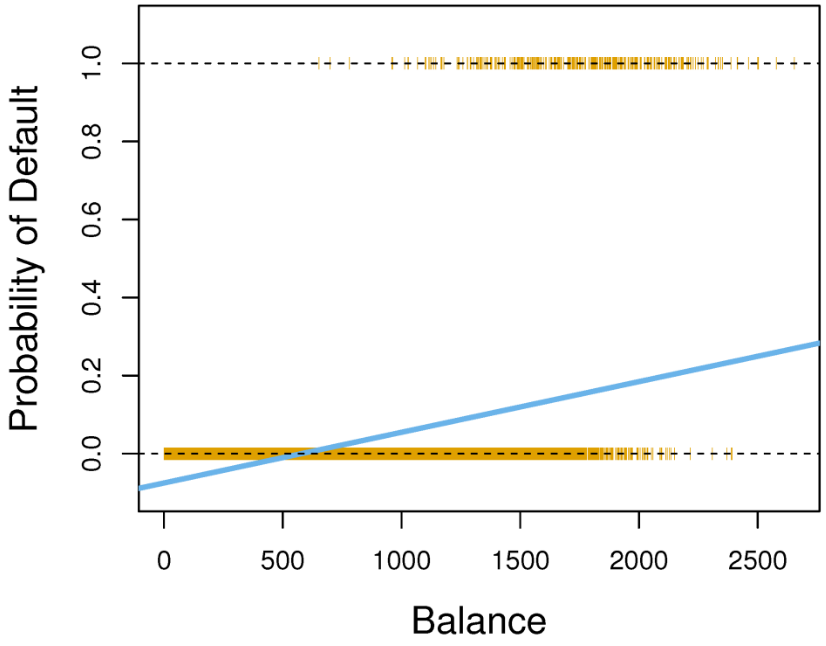
$$\text{default} = \begin{cases} \text{Yes}, & p(\text{balance}) > 0.1 \\ \text{No}, & p(\text{balance}) \le 0.1 \end{cases}$$

Logistic Regression

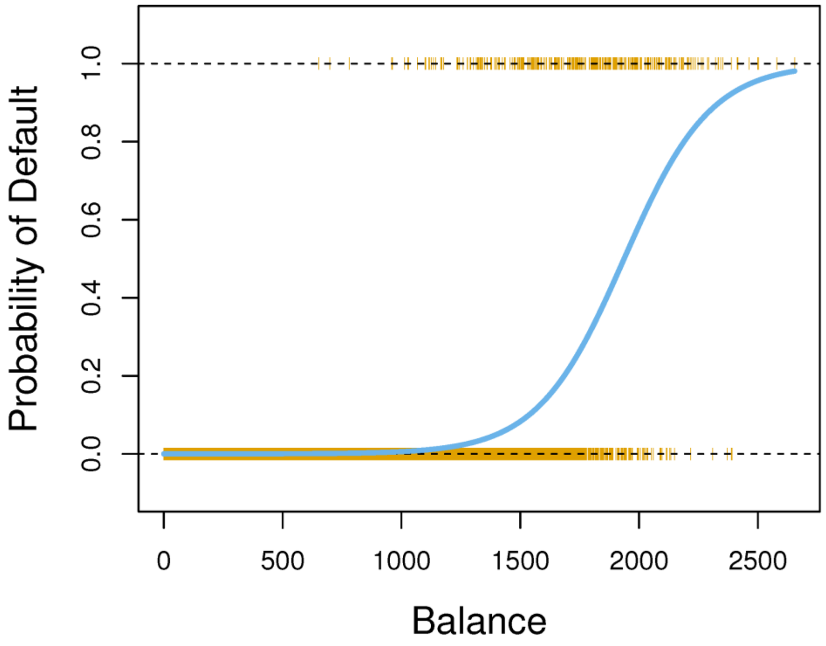
- The logistic model

How should we model the relationship between $p(X) = \Pr(Y = 1 | X)$ and X ?

$$p(X) = \beta_0 + \beta_1 X$$



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



Logistic Regression

- The logistic model

The *odds*

$$\frac{p(X)}{1-p(X)} = e^{\beta_0+\beta_1 X} \in (0,+\infty)$$

- Values of the odds close to 0 and positive infinity indicate very low and very high probabilities of default, respectively.

- Odds are traditionally used instead of probabilities in horse-racing, since they relate more naturally to the correct betting strategy.

Logistic Regression

- The logistic model

The *log-odds* or *logit*

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

- The logistic regression model has a logit that is linear in X.
- In a logistic regression model, increasing X by a one unit changes the log odds by β_1, or equivalently it multiplies the odds by e^{β_1}.
- The amount that $p(X)$ changes due to a one-unit change in X will depend on the current value of X.

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

Logistic Regression

- Estimating the regression coefficients

Maximum likelihood estimation

A statistical method used to estimate the parameters of a probability distribution by maximizing the likelihood function.

The basic idea of MLE: find the values of the model parameters that make the observed data most likely to occur.

Steps:

1. Specify the probability distribution
2. Write the likelihood function
3. Maximize the likelihood function
4. Check if the model fits

Logistic Regression

- Estimating the regression coefficients

Maximum likelihood estimation

Let $p(x) = \Pr(Y = 1 \mid X = x)$, then $1 - p(x) = \Pr(Y = 0 \mid X = x)$

The (conditional) likelihood function is:

$$
\begin{aligned}
l(\beta_0, \beta_1) &= \prod_{i=1}^{n} \Pr(Y = y_i \mid X = x_i) \\
&= \prod_{i:y_i=1} p(x_i) \prod_{i':y_{i'}=0} (1 - p(x_{i'})) \\
&= \prod_{i=1}^{n} p(x_i)^{y_i} (1 - p(x_i))^{1-y_i}
\end{aligned}
$$

The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are chosen to maximize this likelihood function.

Logistic Regression

- Estimating the regression coefficients

Maximum likelihood estimation

The log-likelihood function is:

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

$$\begin{aligned}\log l(\beta_0,\beta_1)&=\sum_{i=1}^{n}\left[y_i\log p(x_i)+(1-y_i)\log(1-p(x_i))\right]\\&=\sum_{i=1}^{n}\log(1-p(x_i))+\sum_{i=1}^{n}y_i\log\frac{p(x_i)}{1-p(x_i)}\\&=\sum_{i=1}^{n}-\log(1+e^{\beta_0+\beta_1 x_i})+\sum_{i=1}^{n}y_i(\beta_0+\beta_1 x_i)\end{aligned}$$

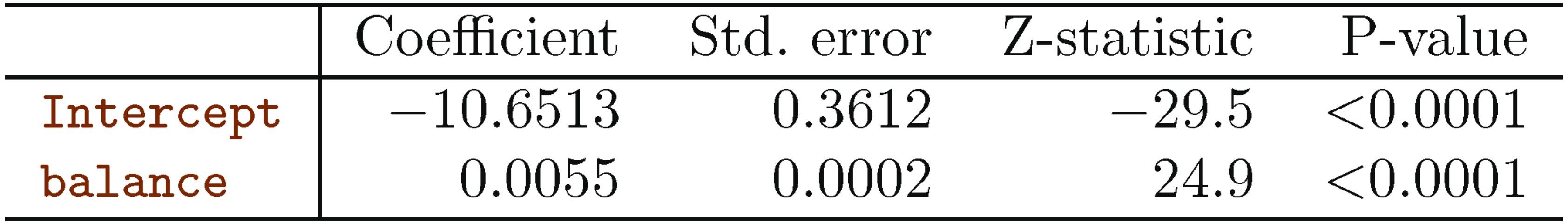
The estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are obtained via numerical optimization methods.

Logistic Regression

- Estimating the regression coefficients

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

Example: the default data set



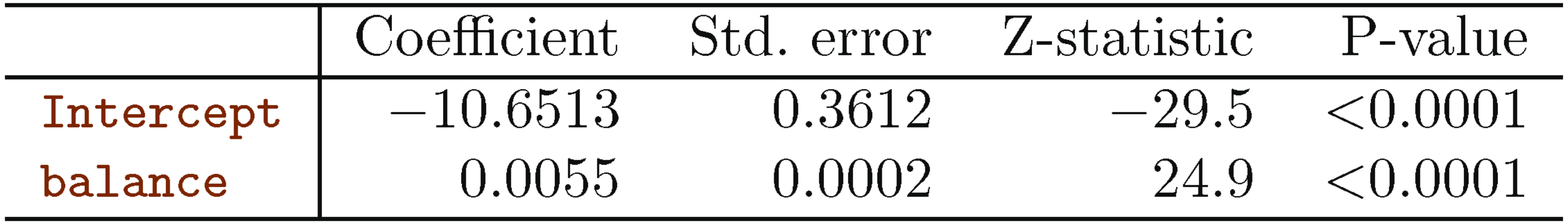
	Coefficient	Std. error	Z-statistic	P-value
`Intercept`	-10.6513	0.3612	-29.5	<0.0001
`balance`	0.0055	0.0002	24.9	<0.0001

A one-unit increase in balance is associated with an increase in the log-odds of default by 0.0055 units.

Logistic Regression

- Estimating the regression coefficients

Example: the default data set



	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

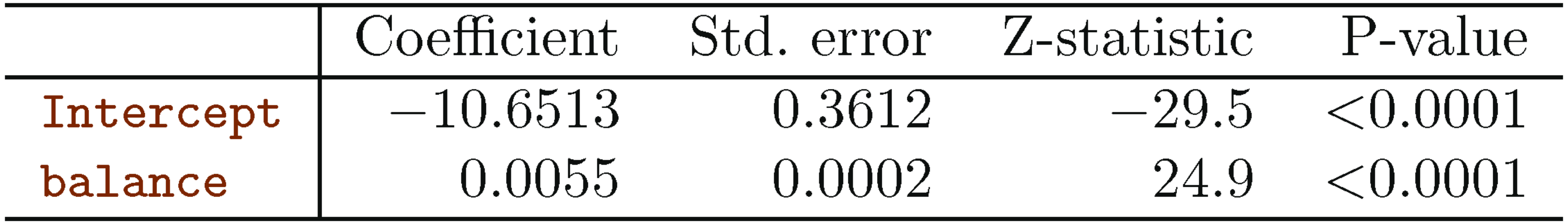
Many aspects of the logistic regression output are similar to the linear regression output:

- We can measure the accuracy of the coefficient estimates by computing their standard errors.

Logistic Regression

- Estimating the regression coefficients

Example: the default data set



	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

Many aspects of the logistic regression output are similar to the linear regression output:

- The z-statistic plays the same role as the t-statistic; $z(\beta_1) = \hat{\beta}_1 / SE(\hat{\beta}_1)$, a large (absolute) value of the z-statistic indicates evidence against the null hypothesis $H_0 : \beta_1 = 0$

Logistic Regression

- Making predictions

Example: the default data set

Once the coefficients have been estimated, it is a simple matter to compute the probability of default for any given credit card balance.

The default probability for an individual with a balance of $1,000 is

$$\hat{p}(X)=\frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}}=\frac{e^{-10.6513+0.0055\times 1000}}{1+e^{-10.6513+0.0055\times 1000}}=0.00576$$

The default probability for an individual with a balance of $2,000 is

$$\hat{p}(X)=\frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}}=\frac{e^{-10.6513+0.0055\times 2000}}{1+e^{-10.6513+0.0055\times 2000}}=0.586$$

Logistic Regression

- Making predictions

We can use qualitative predictors with the logistic regression model using the dummy variable approach.

Example: the default data set

Goal: we are interested in predicting whether an individual will default on his or her credit card payment, on the basis of his or her student status.

We create a dummy variable for student

$$X = \begin{cases} 0, & \text{if non-student} \\ 1, & \text{if student} \end{cases}$$

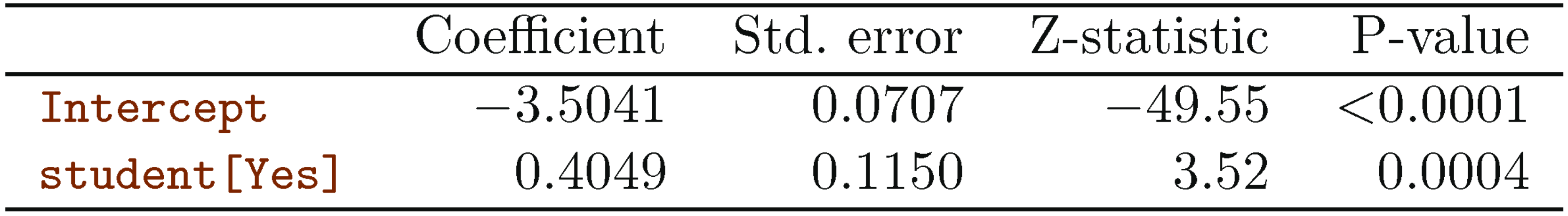
Logistic Regression

- Making predictions

We can use qualitative predictors with the logistic regression model using the dummy variable approach.

Example: the default data set

Goal: we are interested in predicting whether an individual will default on his or her credit card payment, on the basis of his or her student status.



	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

Logistic Regression

- Making predictions

We can use qualitative predictors with the logistic regression model using the dummy variable approach.

Example: the default data set

Goal: we are interested in predicting whether an individual will default on his or her credit card payment, on the basis of his or her student status.

$$\Pr(\text{default} = \text{Yes} \mid \text{student} = \text{Yes}) = \frac{e^{-3.5041+0.4049\times 1}}{1+e^{-3.5041+0.4049\times 1}} = 0.0431$$

$$\Pr(\text{default} = \text{Yes} \mid \text{student} = \text{No}) = \frac{e^{-3.5041+0.4049\times 0}}{1+e^{-3.5041+0.4049\times 0}} = 0.0292$$

Students tend to have higher default probabilities than non-students

Logistic Regression

- Multiple logistic regression (MLE)

Predicting a binary response using multiple predictors.

The predictors can be either continuous or categorical

MLE estimates the probability of a binary outcome based on the values of the independent variables.

The MLE model can be written as:

$$\log\left(\frac{p(X)}{1-p(X)}\right)=\beta_0+\beta_1 X_1+\beta_2 X_2+\ldots+\beta_p X_p,\quad X=(X_1,\ldots,X_p)$$

Logistic Regression

- Multiple logistic regression

Predicting a binary response using multiple predictors.

Similarly

$$\log\left(\frac{p(X)}{1-p(X)}\right)=\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p,\quad X=(X_1,\ldots,X_p)$$

$$p(X)=\frac{e^{\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p}}{1+e^{\beta_0+\beta_1X_1+\beta_2X_2+\ldots+\beta_pX_p}}$$

Again, we can use MLE to estimate.

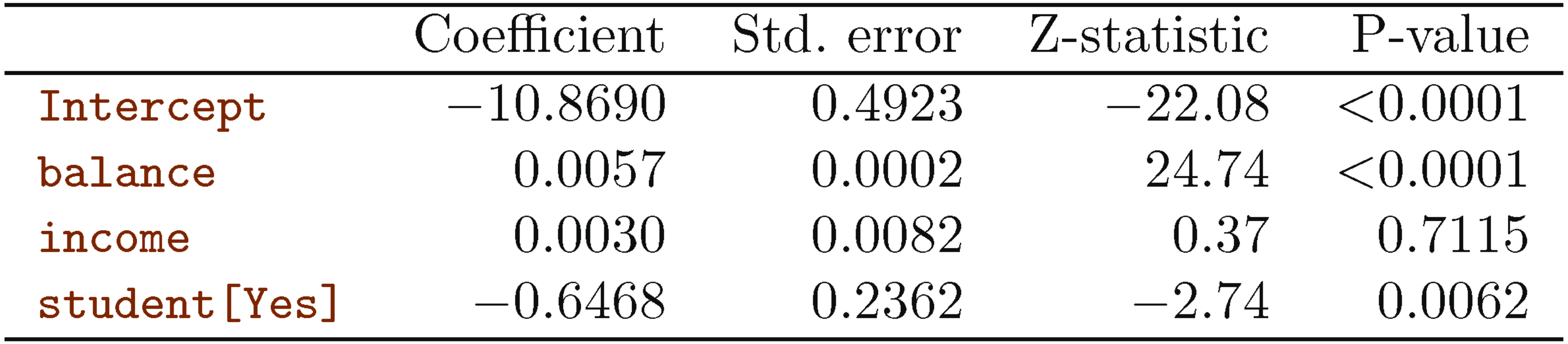
Logistic Regression

- Multiple logistic regression

Predicting a binary response using multiple predictors.

Example: the default data set

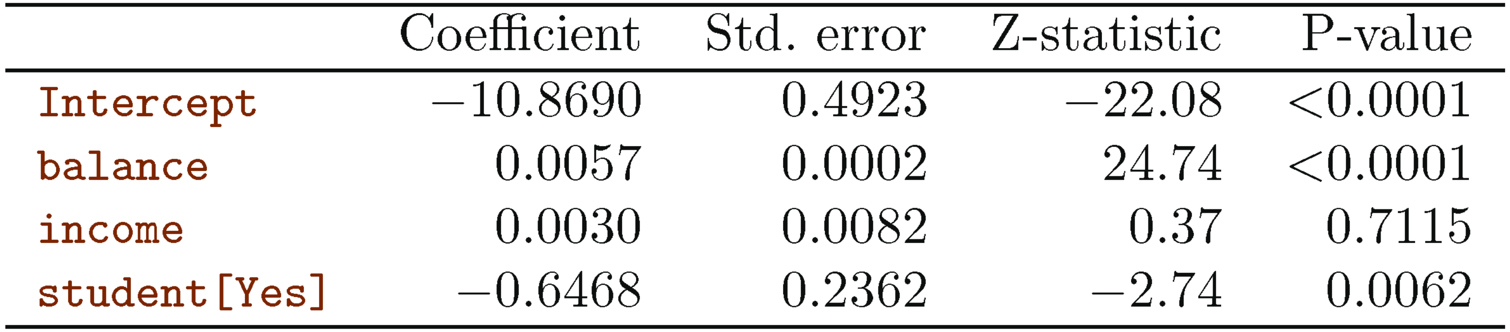
Goal: we are interested in predicting whether an individual will default on his or her credit card payment, on the basis of balance, income (in thousands of dollars), and student status.



	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

Logistic Regression

- Multiple logistic regression

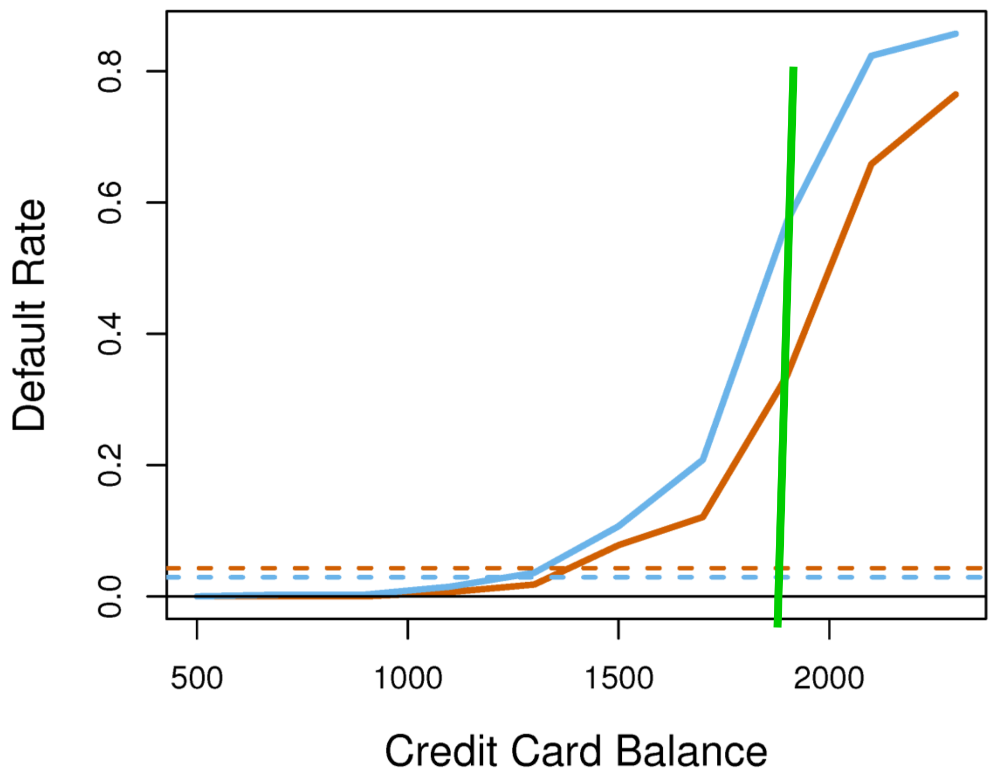


	Coefficient	Std. error	Z-statistic	P-value
Intercept	−10.8690	0.4923	−22.08	<0.0001
balance	0.0057	0.0002	24.74	<0.0001
income	0.0030	0.0082	0.37	0.7115
student[Yes]	−0.6468	0.2362	−2.74	0.0062

Why is coefficient for student negative (students are less likely to default than non-students) while it was positive (students are more likely to default than non-students) before?

Logistic Regression

- Multiple logistic regression



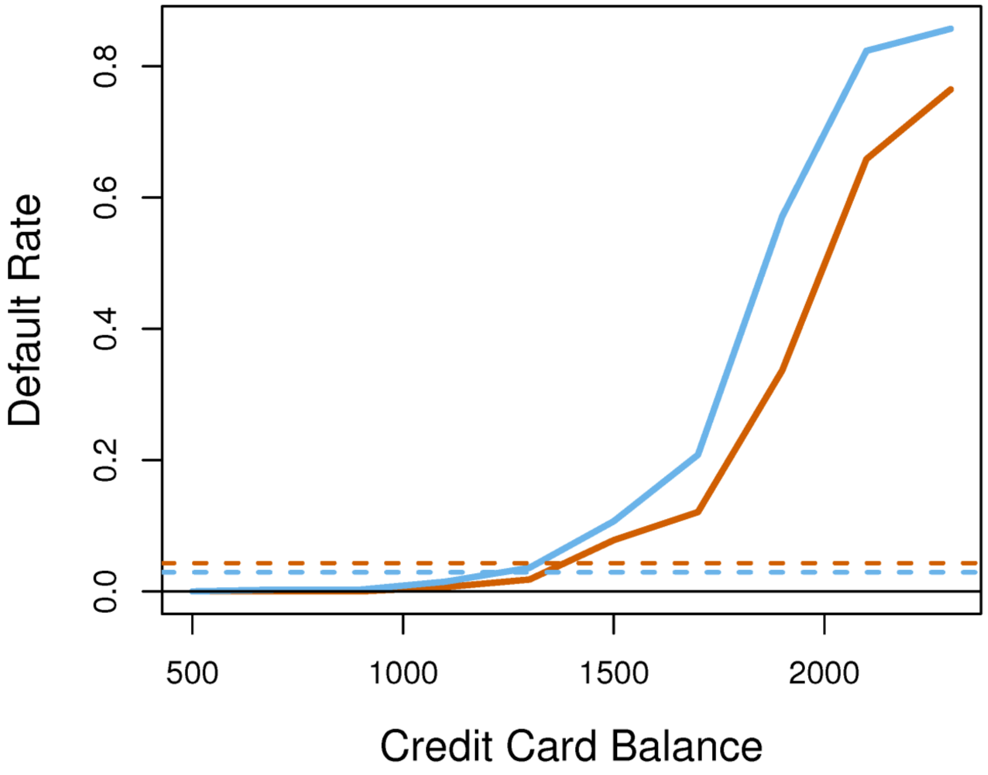
Orange solid line – the average default rates for students as a function of credit card balance

Blue solid line – the average default rates for non-students as a function of credit card balance

Conclusion: For a fixed value of balance and income, a student is less likely to default than a non-student

Logistic Regression

- Multiple logistic regression



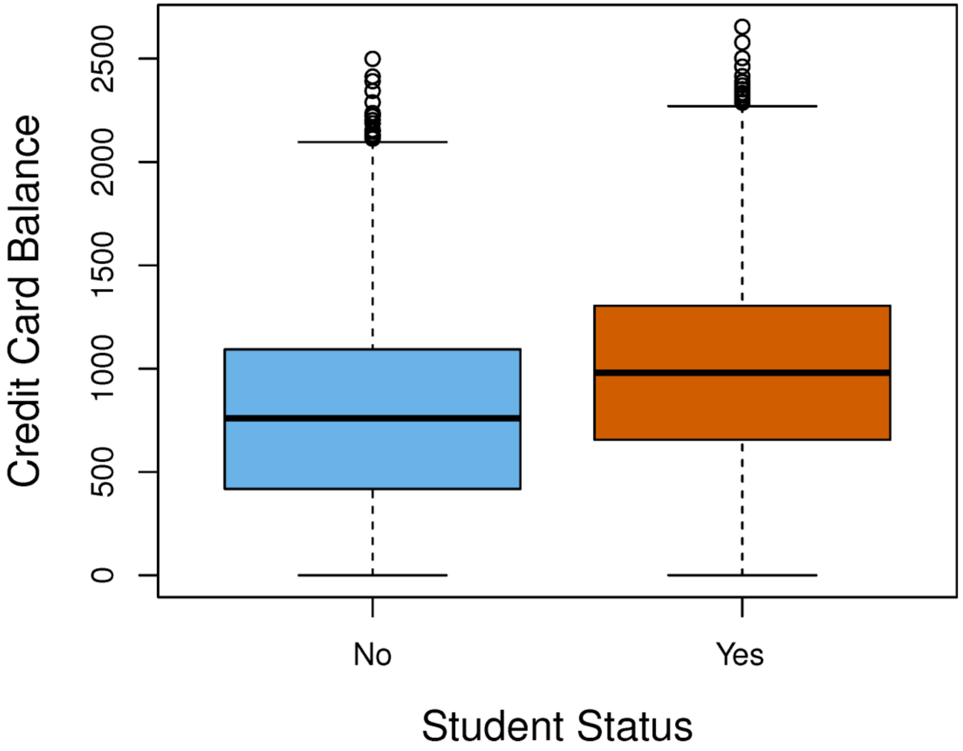
Orange dash line – the default rates for students averaged over all values of balance and income

Blue dash line – the default rates for non-students averaged over all values of balance and income

Conclusion: the overall student default rate is higher than the non-student default rate

Logistic Regression

- Multiple logistic regression



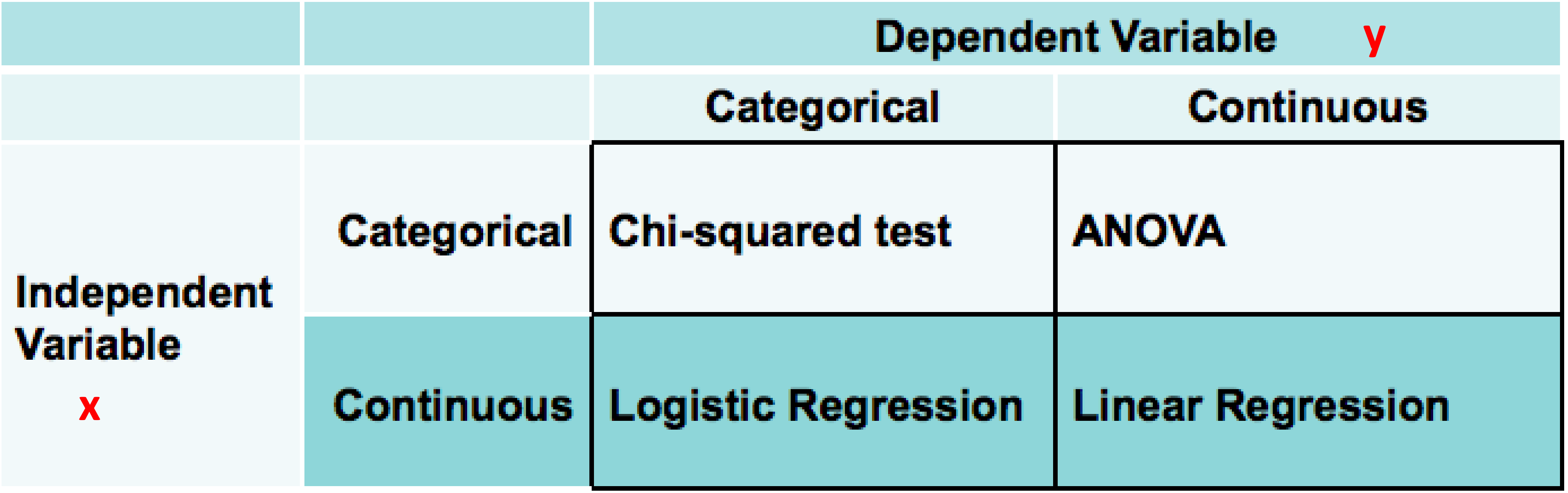
The variables student and balance are correlated, or more correctly, associated.

Students are more likely to have large credit card balances, which, tend to associated with high default rates.

Confounding!

Logistic Regression

- Multiple logistic regression

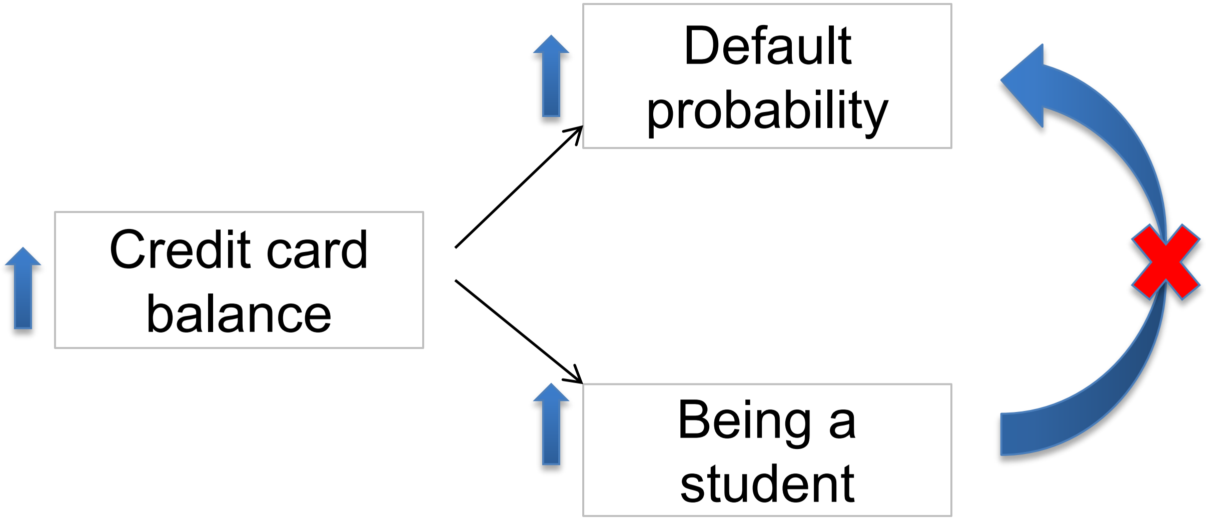


		Dependent Variable y	
		Categorical	Continuous
Independent Variable x	Categorical	Chi-squared test	ANOVA
	Continuous	Logistic Regression	Linear Regression

https://www.quora.com/How-can-I-measure-the-correlation-between-continuous-and-categorical-variables

Logistic Regression

- Multiple logistic regression



Logistic Regression

- Multiple logistic regression

Conclusions:

- A student is riskier (more likely to default) than a non-student if no information about the student's credit card balance is available.
- However, that student is less risky (less likely to default) than a non-student with the same credit card balance.

Logistic Regression

- Multiple logistic regression

The default probability of a student with a credit card balance of $1,500 and an income of $40,000 is:

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1500+0.003\times 40-0.6468\times 1}}{1+e^{-10.869+0.00574\times 1500+0.003\times 40-0.6468\times 1}} = 0.058$$

The default probability of a non-student with the same balance and income is:

$$\hat{p}(X) = \frac{e^{-10.869+0.00574\times 1500+0.003\times 40-0.6468\times 0}}{1+e^{-10.869+0.00574\times 1500+0.003\times 40-0.6468\times 0}} = 0.105$$

Logistic Regression

- Logistic regression for >2 response classes

Classify a response variable that has more than two classes.

$$Y = \begin{cases} 1 & \text{if stroke} \\ 2 & \text{if drug overdose} \\ 3 & \text{if epileptic seizure} \end{cases}$$

Logistic Regression

- Logistic regression for >2 response classes

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

Classify a response variable that has more than two classes (K classes).

$$\log\left(\frac{\Pr(Y=1\,|\,X)}{\Pr(Y=K\,|\,X)}\right) = \beta_{01} + \beta_{11}X_1 + \ldots + \beta_{p1}X_p$$

$$\log\left(\frac{\Pr(Y=2\,|\,X)}{\Pr(Y=K\,|\,X)}\right) = \beta_{02} + \beta_{12}X_1 + \ldots + \beta_{p2}X_p$$

.

.

$$\log\left(\frac{\Pr(Y=K-1\,|\,X)}{\Pr(Y=K\,|\,X)}\right) = \beta_{0K-1} + \beta_{1K-1}X_1 + \ldots + \beta_{pK-1}X_p$$

The model is specified in terms of K-1 log-odds or logits.

Logistic Regression

- Logistic regression for >2 response classes

Classify a response variable that has more than two classes (K classes).

$$\Pr(Y=k\mid X)=\frac{e^{\beta_{0k}+\beta_{1k}X_1+\ldots+\beta_{pk}X_p}}{1+\sum_{l=1}^{K-1}e^{\beta_{0l}+\beta_{1l}X_1+\ldots+\beta_{pl}X_p}},\ k=1,\ldots,K-1$$

$$\Pr(Y=K\mid X)=\frac{1}{1+\sum_{l=1}^{K-1}e^{\beta_{0l}+\beta_{1l}X_1+\ldots+\beta_{pl}X_p}}$$

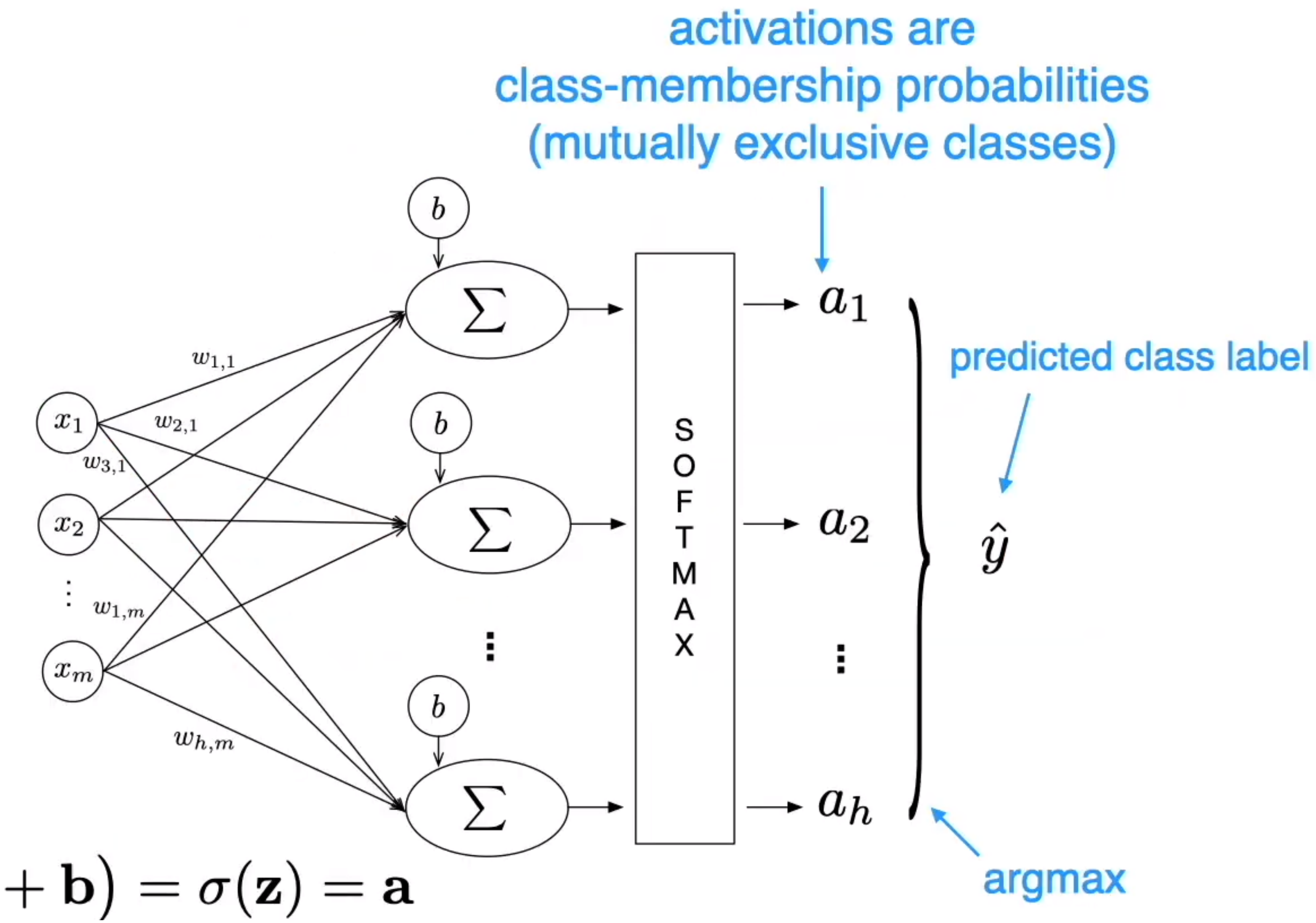
⬇

$$\sum_{l=1}^{K}\Pr(Y=l\mid X)=1$$

$$\hat{p}(X)=\frac{e^{\hat\beta_0+\hat\beta_1X}}{1+e^{\hat\beta_0+\hat\beta_1X}}=\frac{e^{-10.6513+0.0055\times1000}}{1+e^{-10.6513+0.0055\times1000}}=0.00576$$

Logistic Regression

- Multinomial Logistic Regression/ SoftMax Regression



activations are class-membership probabilities (mutually exclusive classes)

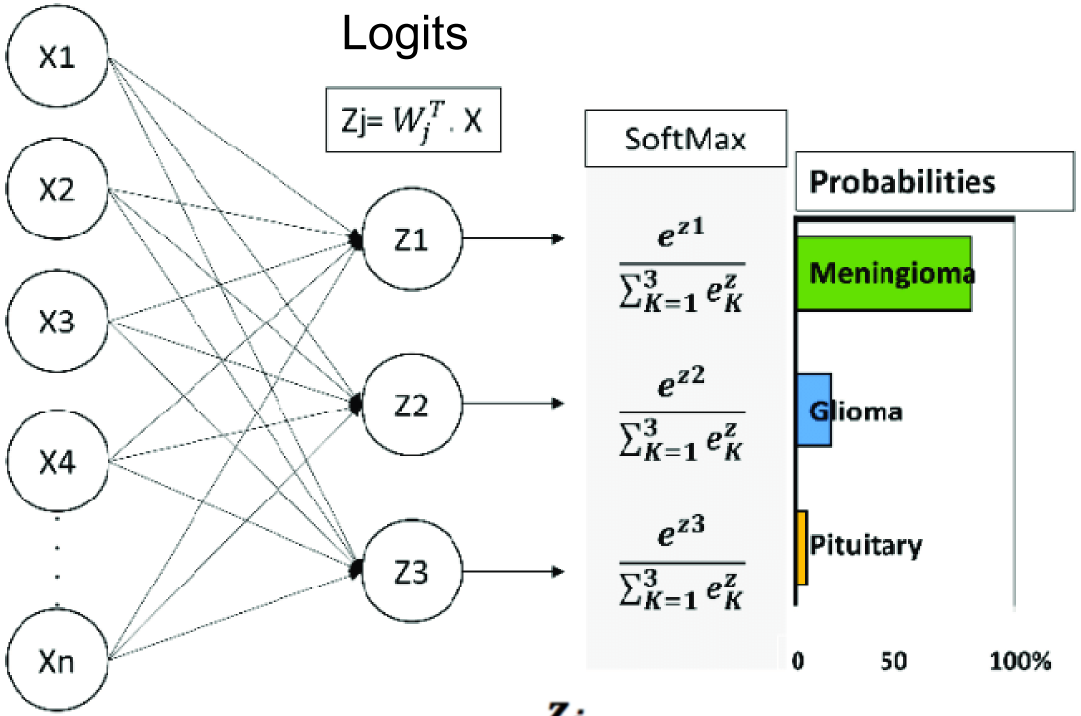
predicted class label

argmax

$$\sigma(\mathbf{Wx}+\mathbf{b}) = \sigma(\mathbf{z}) = \mathbf{a}$$

Logistic Regression

- Relationship between Logistic Regression and SoftMax.



Logits
Zj= $W_j^T . X$
SoftMax
Probabilities
X1
X2
X3
X4
Xn
Z1
Z2
Z3
$\frac{e^{z1}}{\sum_{K=1}^{3} e_K^z}$
$\frac{e^{z2}}{\sum_{K=1}^{3} e_K^z}$
$\frac{e^{z3}}{\sum_{K=1}^{3} e_K^z}$
Meningioma
Glioma
Pituitary
0
50
100%

SoftMax function: $$\sigma(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{k} e^{z_k}} \ for\ j = 1, \dots, k$$

The above is the softmax formula, which takes each Logits and find the probability. The numerator is the e-power values of the Logit and the denominator calculates the sum of the e-power values of all the Logits.

Linear Discriminant Analysis (LDA)

In logistic regression, we model $\Pr(Y = k \mid X = x)$ using the logistic function (direct approach).

In discriminant analysis, we model the distribution of X in each of the classes separately $\Pr(X = x \mid Y = k)$, and then use *Bayes' theorem* to flip things around and obtain $\Pr(Y = k \mid X = x)$ (indirect approach)

Linear Discriminant Analysis (LDA)

- Using Bayes' theorem for classification

$$\Pr(X = x \mid Y = k) \xleftrightarrow{\textit{Bayes' theorem}} \Pr(Y = k \mid X = x)$$

HOW?

$$\Pr(Y = k \mid X = x) = \frac{\Pr(X = x \mid Y = k) \cdot \Pr(Y = k)}{\Pr(X = x)}$$

Linear Discriminant Analysis (LDA)

$$\Pr(Y=k \mid X=x) = \frac{\Pr(X=x \mid Y=k)\cdot \Pr(Y=k)}{\Pr(X=x)}$$

- Using Bayes' theorem for classification

In discriminant analysis, we write

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

--- the *posterior* probability that an observation $X=x$ belongs to the *k*th class.

$f_k(x) = \Pr(X=x \mid Y=k)$ --- the *density* for X in class k. In LDA, we will use normal densities, for these, separately in each class.

$\pi_k = \Pr(Y=k)$ --- the marginal or *prior* probability for k class.

- A simple estimate of the **prior**: the fraction of the training observations that belong to the *k*-th class.
- Estimating the **density** is more challenging, unless we assume some simple forms for these densities.

贝叶斯方法的核心在于：　**后验概率 = 先验概率 * 来自数据的信息**。

例子：假设A和B和C是3个人，3个人都不认识；让A和B打牌，C来猜谁赢的可能性大；没打牌之前C会猜A和B赢的几率各为50%；刚打完一次 A赢了，这时候C就会认为A的技术可能会更好赢的几率会大于50%，A和B继续打牌 一会A连续几次赢一会B连续赢几次，C在这个过程中有时候认为A技术好 点有时候会认为B技术好点；C的判断随着打牌次数不断变化，就是贝叶斯概率原理。

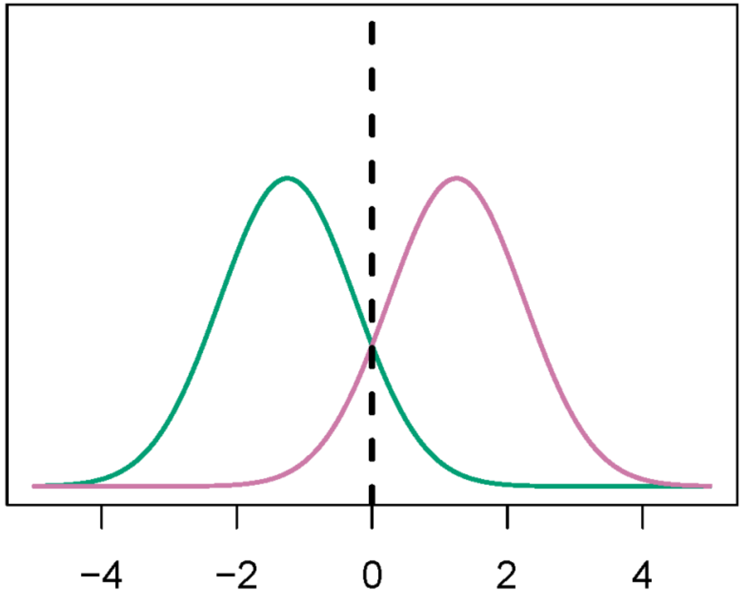
贝叶斯方法最好的地方就在于：你不用知道全局，甚至你一开始可以错的离谱，但你可以越来越接近真理。

Linear Discriminant Analysis (LDA)

- Using Bayes' theorem for classification

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

π_1=.5, π_2=.5



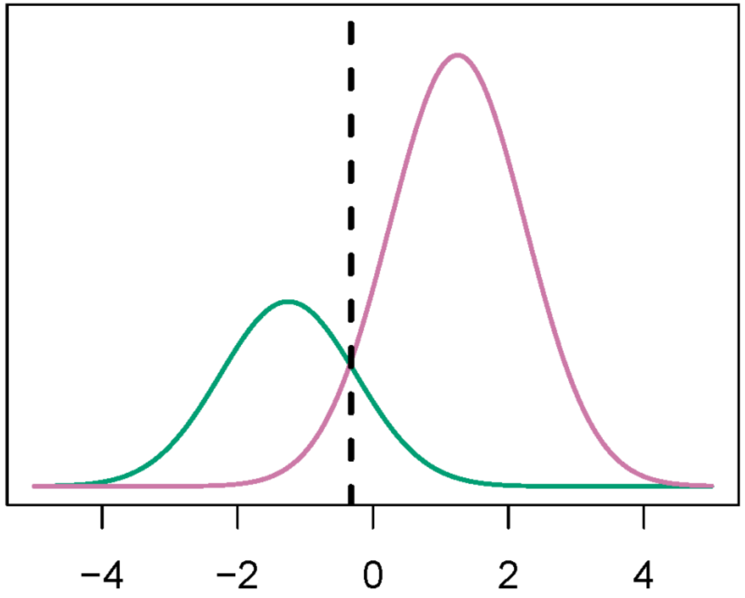
When the priors are equal, we classify a new point according to which density is highest.

Linear Discriminant Analysis (LDA)

- Using Bayes' theorem for classification

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

π_1=.3, π_2=.7



When the priors are different, we take them into account as well, and compare $\pi_k f_k(x)$.

Linear Discriminant Analysis (LDA)

- Using Bayes' theorem for classification

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

Comment: The Bayes classifier classifies an observation to the class for which $p_k(x)$ is largest. It has the lowest possible error rate out of all classifiers. Therefore, if we can find a way to estimate $f_k(X)$, then we can develop a classifier that approximates the Bayes classifier.

Linear Discriminant Analysis (LDA)

- Why discriminant analysis, given that we already have logistic regression?

1. When the classes are well-separated, the parameter estimates for the logistic regression model are surprising unstable. LDA does not suffer from this problem.

2. If n is small and the distribution of X is approximately normal in each of the classes, LDA is again more stable than logistic regression.

3. LDA is popular when we have more than two response classes.

Linear Discriminant Analysis (LDA)

- The main idea of LDA:

1. In LDA, we assume that the data is normally distributed and that each class has its own mean and covariance matrix.

2. The goal of LDA is to find a projection of the data that maximizes the separation between the classes while minimizing the variance within each class.

3. LDA can also be used for dimensionality reduction such as reducing the number of features in a dataset.

Linear Discriminant Analysis (LDA)

- LDA for p=1 这里的P是predictor的意思，也就是自变量x只有一个

When p=1, the Gaussian (normal) density has the form

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

μ_k and ${\sigma_k}^2$ are the mean and variance parameters for the *k*th class. We will assume that all the ${\sigma_k}^2 = \sigma^2$ are the same.

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

Linear Discriminant Analysis (LDA)

- LDA for p=1

To classify at the value of $X = x$, 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*:

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

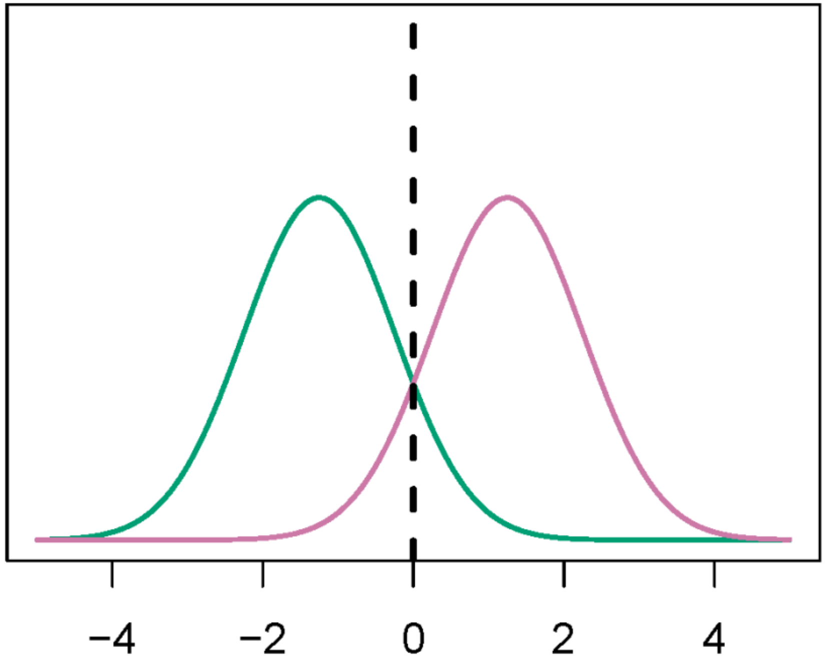
$\delta_k(x)$ is a *linear* function of x --- that's why it is called **linear** DA!

Question: If there are two classes with equal prior probabilities, then what is the decision boundary?

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

Linear Discriminant Analysis (LDA)

- LDA for p=1



$$\mu_1 = -1.25, \mu_2 = 1.25$$
$$\sigma_1{}^2 = \sigma_2{}^2 = 1$$
$$\pi_1 = \pi_2 = 0.5$$

- The Bayes classifier assigns the observation to class 1 if x<0 and class 2 otherwise.
- In this case, we can compute the Bayes classifier because we know that X is drawn from a Gaussian distribution within each class, and we know all the parameters involved. In real-life situation, we are not able to calculate the Bayes classifier.
- In practice, under the assumption of normal distributions, we still need to estimate the parameters $\mu_1, \mu_2, ..., \mu_K$, $\pi_1, \pi_2, ..., \pi_K$ and σ^2.

Linear Discriminant Analysis (LDA)

- LDA for p=1

Parameter estimations in LDA

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

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

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

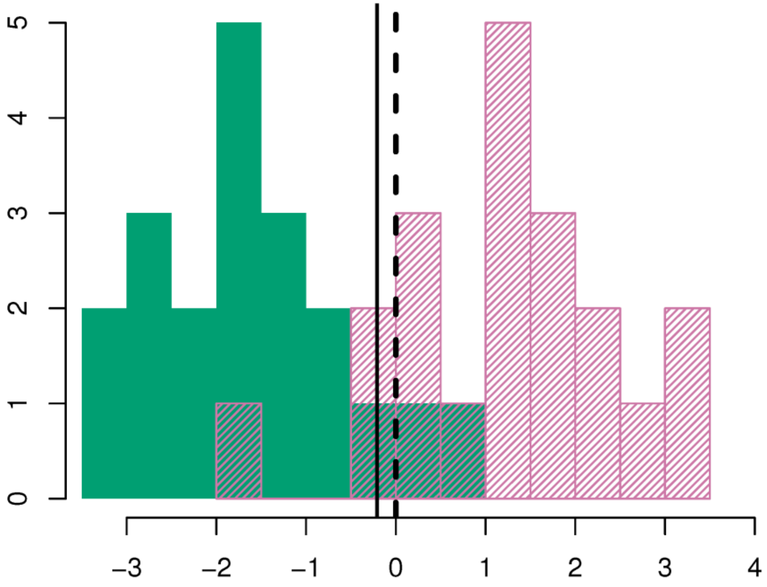
$$= \sum_{k=1}^{K} \frac{n_k - 1}{n-K} \cdot \hat{\sigma}_k{}^2$$

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

- n -- the total number of training observations.
- n_k -- the number of training observations in the k-th class.
- $\hat{\sigma}_k{}^2$ -- the estimated variance in the k-th class.
- $\hat{\sigma}^2$ -- a weighted average of the sample variance for each of the K classes.

Linear Discriminant Analysis (LDA)

- LDA for p=1



A sample of 20 observations for each of the two classes (green versus pink)

Dash – Bayes decision boundary;
Solid – LDA decision boundary;

- Since $n_1 = n_2 = 20$, we have $\hat{\pi}_1 = \hat{\pi}_2 = 0.5$
- The decision boundary corresponds to the midpoint between the sample means for the two classes, $(\hat{\mu}_1 + \hat{\mu}_2)/2$
- The LDA decision boundary is very close to the Bayes decision boundary, indicating the LDA is performing pretty well on this dataset.

Linear Discriminant Analysis (LDA)

- LDA for p>1

We assume that $X = (X_1, X_2, \ldots, X_p)$ is drawn from a *multivariate Gaussian* (multivariate normal) distribution, with a class-specific mean vector and a common covariance matrix.

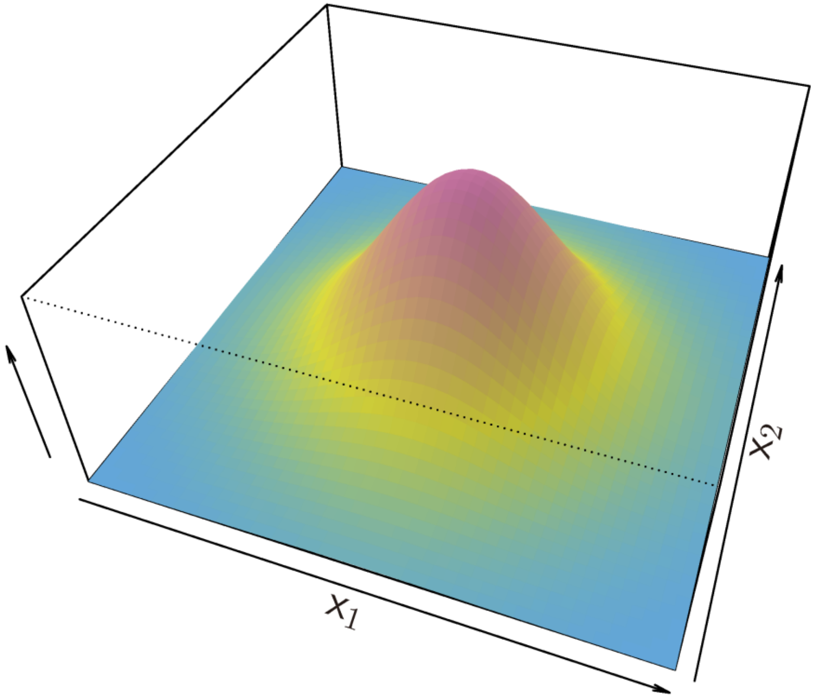
The multivariate Gaussian assumes that each individual predictor follows a ***one-dimension normal distribution****, with* ***some correlation*** *between each pair of predictors.*

Linear Discriminant Analysis (LDA)

- LDA for p>1

multivariate Gaussian

Two predictors are uncorrelated



The surface has a characteristic *bell shape*

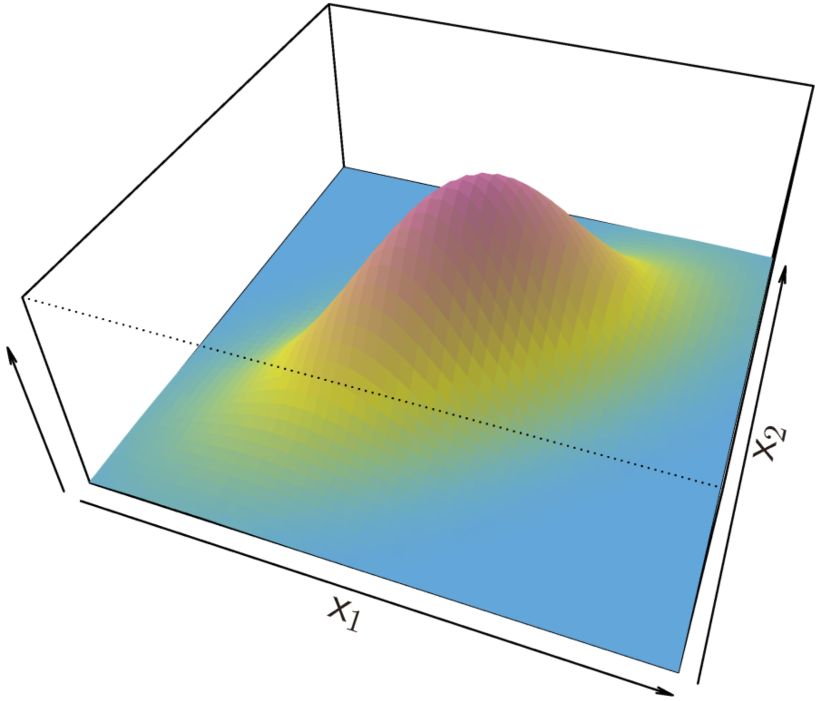
$Var(X_1) = Var(X_2),\ Cor(X_1, X_2) = 0$

Linear Discriminant Analysis (LDA)

- LDA for p>1

multivariate Gaussian

Two predictors have a correlation of 0.7

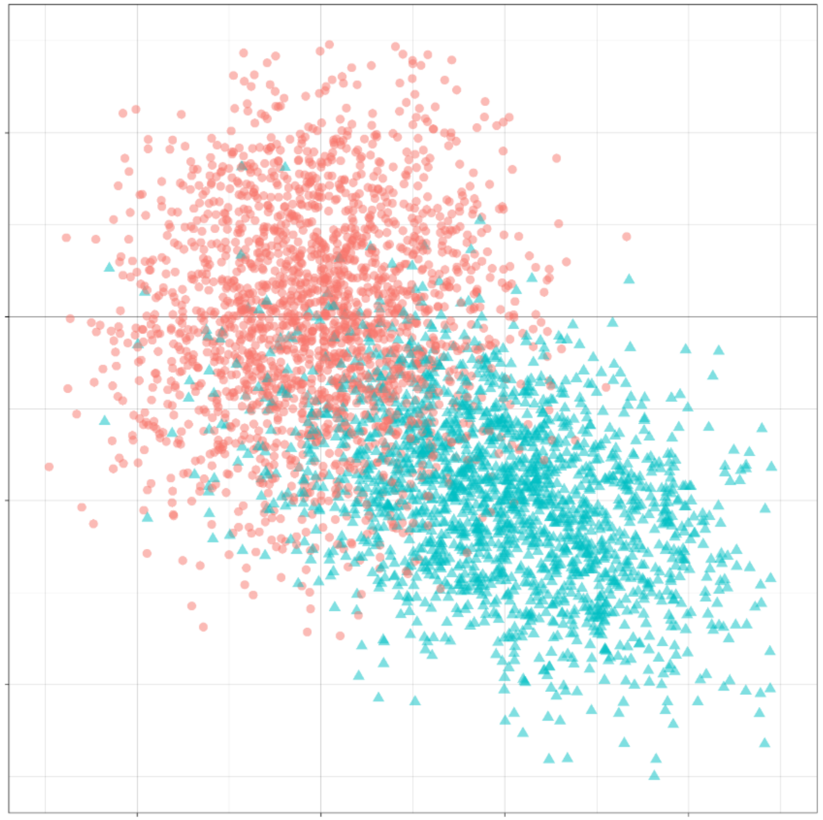


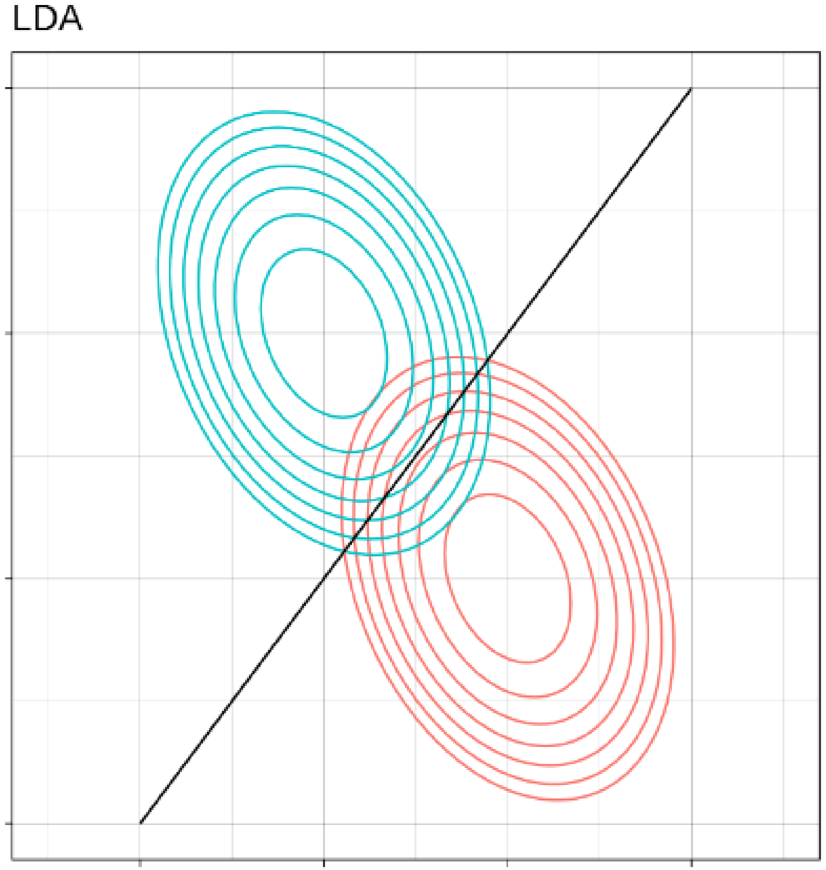
The *bell shape will be distorted if the predictors are correlated or have unequal variances*

Linear Discriminant Analysis (LDA)

- LDA for p>1

multivariate Gaussian





Linear Discriminant Analysis (LDA)

- LDA for p>1

multivariate Gaussian

$X \sim N\left(\mu, \Sigma\right)$

$E(X) = \mu$ -- the mean of X (a vector with p components)

$Cov(X) = \Sigma$ -- the covariance of X (a matrix of size $p \times p$)

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

Linear Discriminant Analysis (LDA)

- LDA for p>1

multivariate Gaussian

$$X \sim N\left(\mu, \Sigma\right)$$

In the case of p>1 predictors, the LDA classifier assumes that the observations in the *k*-th class are drawn from a multi-variate Gaussian distribution $N\left(\mu_k, \Sigma\right)$, where μ_k is a class-specific mean vector, and $\sum$ is a covariance matrix that is common to all K classes.

Linear Discriminant Analysis (LDA)

- LDA for p>1

multivariate Gaussian

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

$$X \sim N(\mu, \Sigma)$$

Plugging into $p_k(x) = \Pr(Y = k \mid X = x) = \dfrac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)}$, we have that the

Bayes classifier assigns an observation $X = x$ to the class for which

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

$$\delta_k(x) = c_{k0} + c_{k1} x_1 + \ldots + c_{kp} x_p$$ -- a linear function (this is the reason for the word *linear* in LDA)