

Week4 lecture - Classification

Dr Liwan Liyanage

School of Computing, Engineering and Mathematics

Classification: Definition

- Given a set of records (called the training set)
- Each record contains a set of attributes. One of the attributes is the class
- Find a model for the class attribute as a function of the values of other attributes
- Goal: Previously unseen records should be assigned to a class as accurately as possible
- Usually, the given data set is divided into training and test set, with training set used to build the model and test set used to validate it. The accuracy of the model is determined on the test set.

Classification Example: Direct Marketing

- Goal: Reduce cost of mailing by targeting a set of consumers likely to buy a new cell phone product
- Approach:
 - Use the data collected for a similar product introduced in the recent past.
 - Use the profiles of customers along with their {buy, didn't buy} decision. The latter becomes the class attribute.
 - The profile of the information may consist of demographic, lifestyle and company interaction.
 - Demographic - Age, Gender, Geography, Salary
 - Psychographic - Hobbies
 - Company Interaction - Recency, Frequency, Monetary
 - Use these information as input attributes to learn a classifier model

Classification Example: Fraud Detection

- Goal: Predict fraudulent cases in credit card transactions
- Approach:
 - Use credit card transactions and the information on its account holders as attributes (important information: when and where the card was used)
 - Label past transactions as {fraud, fair} transactions. This forms the class attribute
 - Learn a model for the class of transactions
 - Use this model to detect fraud by observing credit card transactions on an account

Classification Example: Customer Churn

- Goal: To predict whether a customer is likely to be lost to a competitor
- Approach:
 - Use detailed record of transaction with each of the past and current customers, to find attributes

How often does the customer call, Where does he call, What time of the day does he call most, His financial status, His marital status, etc. (Important Information: Expiration of the current contract).

- Label the customers as {churn, not churn}
- Find a model for Churn

Classification Example: Sky survey cataloging

- Goal: To predict class {star, galaxy} of sky objects, especially visually faint ones, based on the telescopic survey images (from Palomar Observatory)
 - 3000 images with 23,040 x 23,040 pixels per image
- Approach:
 - Segment the image
 - Measure image attributes (40 of them) per object
 - Model the class based on these features
- Success story: Could find 16 new high red-shift quasars (some of the farthest objects that are difficult to find) !!!

Classification problems

Classification problems occur often, perhaps even more so than regression problems. Some examples include

- A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of the three conditions does the individual have
- An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the users IP address, past transaction history, and so forth.
- On the basis of DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are disease causing and which are not.

Classification Continued

Consider a qualitative target or response variable Y and associated predictor variables X_1, X_2, \dots, X_p .

The Classification task is to build a function or a rule set in terms of X_1, X_2, \dots, X_p that takes as input and predicts its value (or category) for Y

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

- Eye color $\in \{\text{brown, blue, green}\}$
- Species $\in \{\text{versicolour, virginica, sethosa}\}$.
- Insurance claim $\in \{\text{fraudulent, legitimate}\}$.

Note that these Qualitative variables take values in an unordered set C :



Use Default Data set

```
library('ISLR')  
attach(Default)  
View(Default)  
dim(Default)
```

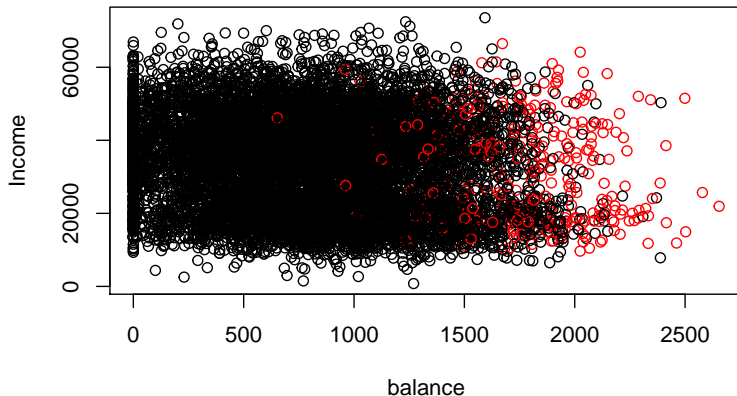
```
## [1] 10000      4
```

```
head(Default)
```

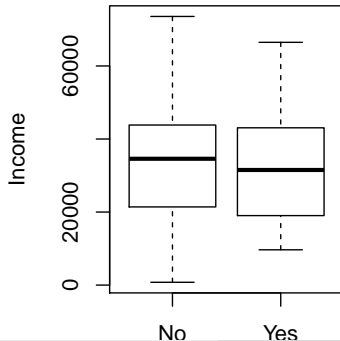
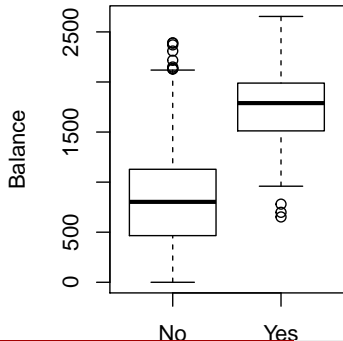
##	default	student	balance	income
## 1	No	No	729.5265	44361.625
## 2	No	Yes	817.1804	12106.135
## 3	No	No	1073.5492	31767.139
## 4	No	No	529.2506	35704.494
## 5	No	No	785.6559	38463.496
## 6	No	Yes	919.5885	7491.559

Explore Data

```
plot(income~balance,col=default, data=Default, ylab= "Income")
```



```
par(mfrow=c(1,2))
boxplot(balance~default, data=Default,
        ylab= "Balance", xlab= "Default")
boxplot(income~default, data=Default,
        ylab= "Income", xlab="Default")
```



Can we use Linear Regression when Y is qualitative?

If we are to classify customers according to credit card Default

Set Code

$Y = 0$, if default is No

$Y = 1$, if default is Yes

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

Limitations of using Linear Regression when Y is binary

Consider Y is default and X is balance,

Then simple linear regression model is

$$\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i \text{ for } i = 1, 2, \dots, n$$

In this case of a binary outcome, linear regression does a good job as a classifier Why????

$$E(Y_i) = P(Y = 1)1 + P(Y = 0)0 = P(Y = 1) = P(Defaul\textit{t})$$

$$\text{and } E(Y_i) = \alpha + \beta X_i$$

Therefore

$$P(Defaul\textit{t}) = \alpha + \beta X_i = E(Y_i)$$

NOTE: However, linear regression might produce probabilities less than zero or bigger than one. Next few slides we demonstrate this point.

Then we consider Logistic Regression

Simple Linear Regression Model requires a numeric X and Y variables

- How to change a factor variable to a numeric variable?
 - Add another variable named Defcode to table Default
 - check the levels of the new variable (It will be same class as the original variable)

```
Defcode = Default$default  
levels(Defcode)
```

```
## [1] "No" "Yes"
```

Change the levels as 1 for Yes and 0 for No

```
levels(Defcode) [levels(Defcode)=="No"]=0  
levels(Defcode) [levels(Defcode)=="Yes"]=1  
levels(Defcode)
```

```
## [1] "0" "1"
```

Still Defcode variable is a factor variable and cannot use as a numeric variable in regression setting.

To summarise a factor variable

```
Defcode = as.character(Default$default)
table(Defcode)
```

```
## Defcode
##      No   Yes
## 9667  333
```

```
Defcode[Defcode=="No"]=0
Defcode[Defcode=="Yes"]=1
table(Defcode)
```

```
## Defcode
##      0    1
## 9667  333
```


Change a factor variable to a numeric variable

```
Defcode = as.numeric(Defcode)  
class(Defcode)
```

```
## [1] "numeric"
```

Simple Linear Regression Model for Default data set with Coded Y

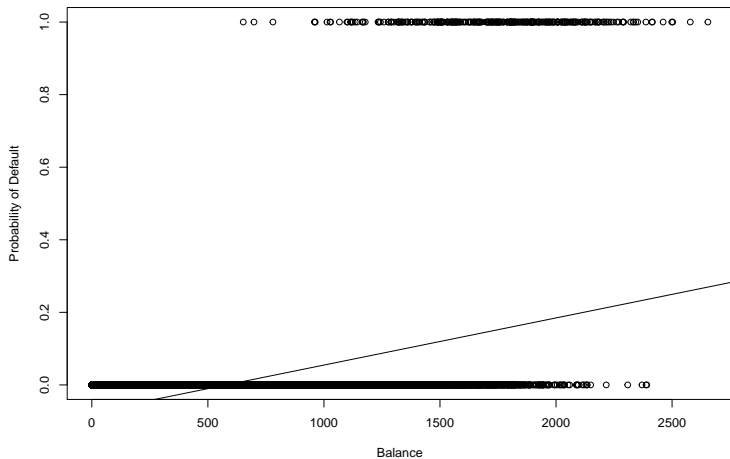
Set the codes for variable default

```
model1=lm(Defcode~balance)  
summary(model1)
```

```
##
## Call:
## lm(formula = Defcode ~ balance)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.23533 -0.06939 -0.02628  0.02004  0.99046
##
## Coefficients:
##              Estimate Std. Error t value Pr(>|t|)
## (Intercept) -7.519e-02  3.354e-03  -22.42  <2e-16 ***
## balance      1.299e-04  3.475e-06   37.37  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.1681 on 9998 degrees of freedom
## Multiple R-squared:  0.1226, Adjusted R-squared: 0.1225
## F-statistic: 1397 on 1 and 9998 DF,  p-value: < 2.2e-16
```

Plot to illustrate how Linear Regression Model produce probabilities less than zero or bigger than one.

```
plot(balance, Defcode,  
      ylab= "Probability of Default", xlab= "Balance")  
abline(a= -7.519e-02, b=1.299e-04)
```



Logistic Regression NOTE:

It's useful to treat simple logistic regression and Multiple Logistic Regression separately

- simple logistic regression, with only one independent variable
- multiple logistic regression, which has more than one independent variable

Classification Estimating the Probabilities

Often we are more interested in estimating the probabilities of Y assuming each of its category.

For example, it is more valuable to have:

- an estimate of the probability a customer with a given balance will default , than a classification default or not.
- an estimate of the probability an insurance claim is fraudulent, than a classification fraudulent or not.

Logistic Regression: Odds and LogOdds

$$Odds(Default) = \frac{P(Y=1/X)}{P(Y=0/X)} = \frac{P(Y=1/X)}{1-P(Y=1/X)} = \frac{P(X)}{1-P(X)}$$

$\log(Odds) = \log \frac{P(X)}{1-P(X)} = \alpha + \beta X$ which is called the logit transformation of $P(X)$

Therefore by rearranging we get the logistic regression form:

$$P(X) = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}$$

$e=2.71828$ is a mathematical constant Euler's number. It is easy to show $P(x)$ will always have values between 0 and 1 irrespective of the value of X .

Note that $P(x)$ is the probability that a person with balance X will default.

Contrast Between Logistic and Linear Regression

– In linear regression, the expected value of Y_1 given X_1 is

$$E(Y_i) = \alpha + \beta X_i \text{ for } i = 1, 2, \dots, n$$

Y_i has a normal distribution with standard deviation σ . It is the random component of the model, which has a normal distribution.

$\alpha + \beta X_i$ is the linear predictor.

- In logistic regression, the expected value of Y_1 given X_1 is

$$\text{logit}(E(Y_i)) = \alpha + \beta X_i \text{ for } i = 1, 2, \dots, n$$

$\text{logit}(E(Y_i))$ is the random component of the model

logit is the link function that relates the expected value of the random component to the linear predictor.

Note: $\text{logit}(\pi) = \log \frac{\pi}{(1-\pi)}$

Maximum Likelihood Estimation

- In linear regression we used the method of least squares to estimate regression coefficients.
- In generalized linear models we use another approach called maximum likelihood estimation.
- The maximum likelihood estimate of a parameter is that value that maximizes the probability of the observed data.
- We estimate $\hat{\alpha}$ and $\hat{\beta}$ by those values and that maximize the probability of the observed data under the logistic regression model.

Maximum Likelihood Estimates

- We use maximum likelihood to estimate the parameters

$$L(\alpha, \beta) = \prod_{i=0}^n [P(X_i)] \prod_{i=1}^n [(1 - P(X_{i'}))]$$

X_i when $Y_i = 1$ and $X_{i'}$ when $Y_i = 0$

- This likelihood gives the probability of the observed zeros and ones in the data.
- We pick $\hat{\alpha}$ and $\hat{\beta}$ to maximize the likelihood of the observed data.
- Most statistical packages can fit linear logistic regression models by maximum likelihood.
- In R we use the glm function.
- Maximum likelihood is a very general approach that is used to fit many of the non-linear models. In the linear regression setting, the least squares approach is in fact a special case of maximum likelihood

Logistic Regression model using glm function

```
model2= glm(Defcode~balance,data=Default,family=binomial)  
summary.glm(model2)
```

```
##
## Call:
## glm(formula = Defcode ~ balance, family = binomial, data = 
##
## Deviance Residuals:
##      Min        1Q      Median        3Q        Max
## -2.2697  -0.1465  -0.0589  -0.0221   3.7589
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.065e+01  3.612e-01  -29.49  <2e-16 ***
## balance      5.499e-03  2.204e-04   24.95  <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## (Dispersion parameter for binomial family taken to be 1)
##
##      Null deviance: 2920.6  on 9999  degrees of freedom
```

Making Predictions

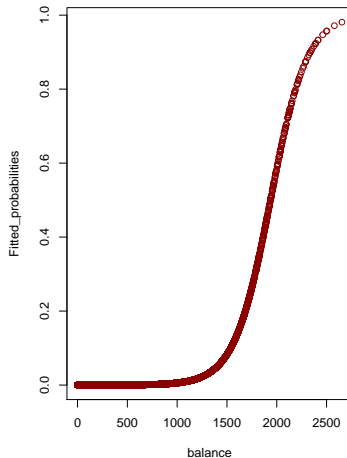
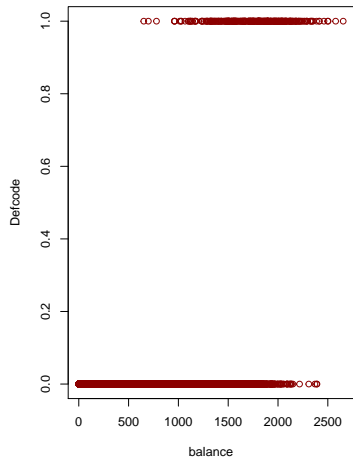
What is our estimated probability of default for someone with a balance of dollars 1000?

$$P(\hat{X}) = \frac{e^{\hat{\alpha} + \hat{\beta}X}}{1 + e^{\hat{\alpha} + \hat{\beta}X}} = ?$$

What is our estimated probability of default for someone with a balance of dollars 2000?

$$P(\hat{X}) = \frac{e^{\hat{\alpha} + \hat{\beta}X}}{1 + e^{\hat{\alpha} + \hat{\beta}X}} = ?$$

Plot of Fitted Probabilities of Default



Lets do it again, using student as the predictor

```
model3= glm(Defcode~student,data=Default,family=binomial)
summary.glm(model3)
```

```
##
## Call:
## glm(formula = Defcode ~ student, family = binomial, data =
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -0.2970  -0.2970  -0.2434  -0.2434   2.6585
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -3.50413    0.07071  -49.55  < 2e-16 ***
## studentYes   0.40489    0.11502   3.52 0.000431 ***
## ---
```


Predict Probability of Default using variable Student

$$\hat{Prob}(Default = Yes/Student = Yes) = \frac{e^{-3.50413+0.40489*1}}{1+e^{-3.50413+0.40489*1}} = ?$$

$$\hat{Prob}(Default = No/Student = No) = \frac{e^{-3.50413+0.40489*0}}{1+e^{-3.50413+0.40489*0}} = ?$$

Confounding

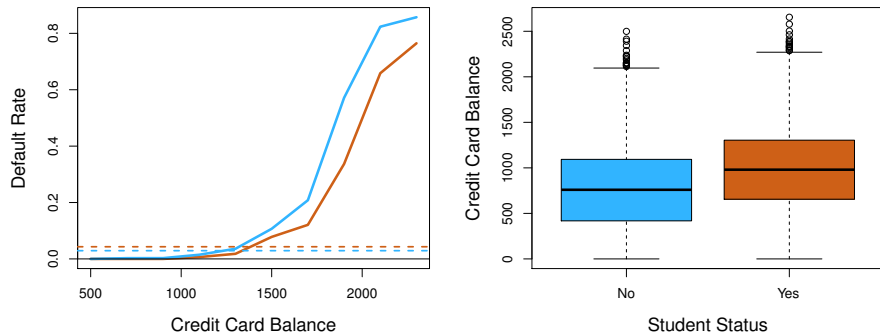


Figure 1:

Summary

- 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.

Logistic regression with several variables

Consider the Qualitative Binary Target Variable Y and several predictor X Variables $X_1, X_2, X_3, \dots, X_p$

Then the logistic regression model for the expected value of Y_i given $X_1, X_2, X_3, \dots, X_p$ is

$$\text{logit}(E(Y)) = \text{logit}(P(X)) = \log \frac{P(X_i)}{1-P(X_i)} = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi} \text{ for } i=1,2,\dots,n$$

$\text{logit}(E(Y))$ is the random component of the model

Then

$$P(X) = \frac{e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi}}}{1 + e^{\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \dots + \beta_p X_{pi}}}$$

Logistic regression with several variables

```
model4=glm(Defcode~student+balance+income, data=Default, family=
summary(model4)
```

```
##
## Call:
## glm(formula = Defcode ~ student + balance + income, family =
##      data = Default)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.4691  -0.1418  -0.0557  -0.0203   3.7383
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
## studentYes  -6.468e-01  2.363e-01  -2.738   0.00619 **
```

```
##
## Call:
## glm(formula = Defcode ~ student + balance + income, family =
##     data = Default)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.4691  -0.1418  -0.0557  -0.0203   3.7383
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
## studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
## balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
## income       3.033e-06  8.203e-06   0.370  0.71152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
##
## Call:
## glm(formula = Defcode ~ student + balance + income, family
##      data = Default)
##
## Deviance Residuals:
##      Min        1Q    Median        3Q        Max
## -2.4691  -0.1418  -0.0557  -0.0203   3.7383
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept) -1.087e+01  4.923e-01 -22.080  < 2e-16 ***
## studentYes  -6.468e-01  2.363e-01  -2.738  0.00619 **
## balance      5.737e-03  2.319e-04  24.738  < 2e-16 ***
## income       3.033e-06  8.203e-06   0.370  0.71152
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
```

Heart data

##	X	Age	Sex	ChestPain	RestBP	Chol	Fbs	RestECG	MaxHR	ExAr
## 1	1	63	1	typical	145	233	1	2	150	
## 2	2	67	1	asymptomatic	160	286	0	2	108	
## 3	3	67	1	asymptomatic	120	229	0	2	129	
## 4	4	37	1	nonanginal	130	250	0	0	187	
## 5	5	41	0	nontypical	130	204	0	2	172	
## 6	6	56	1	nontypical	120	236	0	0	178	
##	Ca	Thal	AHD							
## 1	0	fixed	0							
## 2	3	normal	1							
## 3	2	reversable	1							
## 4	0	normal	0							
## 5	0	normal	0							
## 6	0	normal	0							

Heart data

```
##
## Call:
## glm(formula = AHD ~ ., family = binomial, data = Heart)
##
## Deviance Residuals:
##      Min       1Q   Median       3Q      Max
## -2.7629  -0.5101  -0.1494   0.3460   2.7301
##
## Coefficients:
##              Estimate Std. Error z value Pr(>|z|)
## (Intercept)   -4.739690    2.931081  -1.617  0.10587
## X              0.002676    0.002221   1.205  0.22811
## Age          -0.013183    0.024785  -0.532  0.59479
## Sex            1.486941    0.519796   2.861  0.00423
## ChestPainnonanginal -1.755328    0.493018  -3.560  0.00037
## ChestPainnontypical -0.951481    0.560165  -1.699  0.08940
```

ANOVA

```
## Analysis of Deviance Table
##
## Model: binomial, link: logit
##
## Response: AHD
##
## Terms added sequentially (first to last)
##
##
```

	Df	Deviance	Resid. Df	Resid. Dev
## NULL			296	409.95
## X	1	0.800	295	409.15
## Age	1	15.661	294	393.49
## Sex	1	31.018	293	362.47
## ChestPain	3	72.702	290	289.77
## RestBP	1	5.391	289	284.38

Why Not Linear Regression for Y with more than two classes?

- Suppose that we are trying to predict the medical condition of a patient in the emergency room on the basis of her symptoms. In this simplified example, there are three possible diagnoses: stroke, drug overdose, and epileptic seizure. We could consider encoding these values as a quantitative response variable, Y , as follows:

$Y = 1$ if Stroke; $Y = 2$ if drug overdose; $Y = 3$ if epileptic seizure.

Using this coding, least squares could be used to fit a linear regression model to predict Y on the basis of a set of predictors X_1, \dots, X_p .

Unfortunately, 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.

Continued

If the response variable's values did take on a natural ordering, such as mild, moderate, and severe, and we felt the gap between mild and moderate was similar to the gap between moderate and severe, then a 1, 2, 3 coding would be reasonable. Unfortunately, in general there is no natural way to convert a qualitative response variable with more than two levels into a quantitative response that is ready for linear regression.

Logistic Regression for >2 Response Classes

The two-class logistic regression models discussed have multiple-class extensions, but in practice they tend not to be used all that often. One of the reasons is that the method discriminant analysis, is popular for multiple-class classification.

So we do not go into the details of multiple-class logistic regression here, but simply note that such an approach is possible, and that software for it is available in R

TEXTBOOK

Lecture notes are based on the textbook,
for further reference refer chapter 4;

Prescribed Textbook

– James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An Introduction to Statistical Learning: with Applications in R Springer.