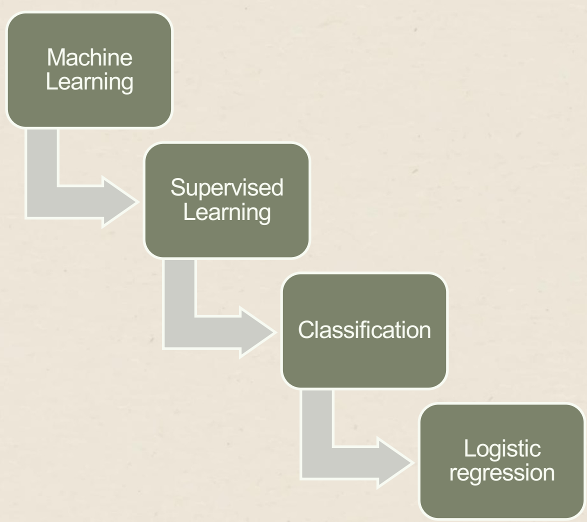
Data Mining and Machine Learning for Business

DSCI 5240

1

Logistic Regression



2

Classification

- Dependent variable (or response variable) takes value from a set, e.g.
  - result={win, lose}
  - Purchasedproduct={chocolate, ice cream, vegetable,…}
  - Fraud={Yes, No}
- Given a feature vector *X* and a response *Y* taking values in the set *C*, the classification task is to build a function *C(X)* that takes as input the feature vector *X* and predicts its value for *Y* ; i.e., $C(X) \in C$.
- Often, we estimate the probabilities that X belongs to each category in C

3

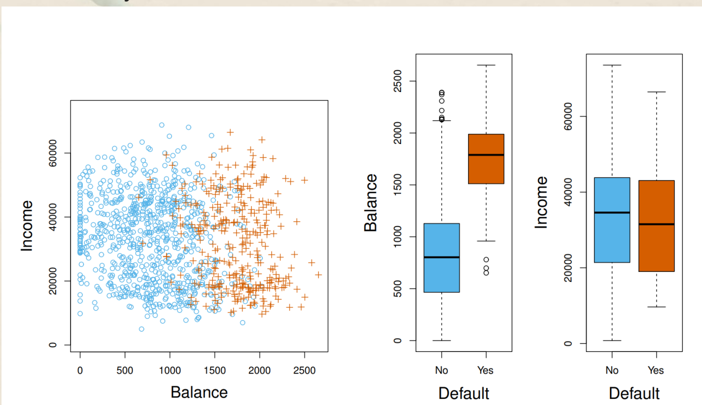
Popular classification techniques (classifiers)

- Logistic regression
- Tree-based Methods
  - Decision Tree
  - Random Forest
- Bayesian methods
  - Naïve Bayesian
  - Linear Discriminant Analysis
  - Quadratic Discriminant Analysis
- K-Nearest Neighbours
- Support Vector Machines
- Neural networks

4

An Example

- Y: whether an individual will default on his or her credit card payment
- Xs: monthly income and credit card balance



- Blue: Not default
- Orange: default

To explore the relationship between default and balance / Income, use boxplots across categories.

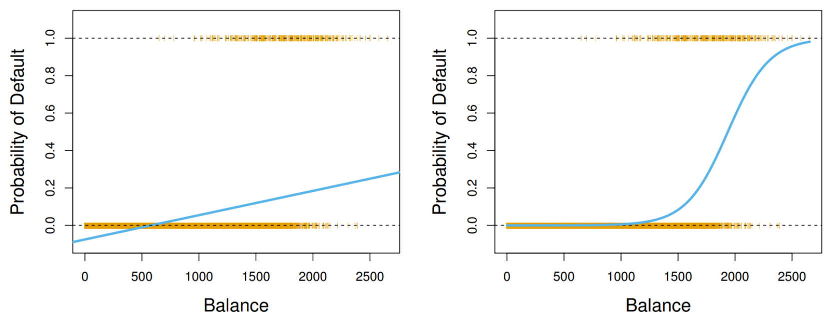
5

Can we use linear regression?

- Suppose for the Default classification task that we code
  - Y = 0 if No
  - Y = 1 if Yes
- Can we simply perform a linear regression of Y on X and classify as Yes if predY > 0.5?
- In this case of a binary outcome, linear regression does a good job as a classifier
- However, linear regression might produce probabilities less than zero or bigger than one.
- *Logistic regression* is more appropriate

6

Linear vs. Logistic Regression



The orange marks indicate the response Y, either 0 or 1. Linear regression does not estimate $\Pr(Y = 1|X)$ well. Logistic regression seems well suited to the task.

7

Logistic regression - Model

- Let's write *p(X) = Pr(Y = 1|X)* for short. Logistic regression uses the form

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

(e ≈ 2.71828 is a mathematical constant [Euler's number.])

- It is easy to see that no matter what values β0, β1 or X take, *p(X)* will have values between 0 and 1
- Rearrange the function, we have

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

- This monotone transformation is called the *log odds* or *logit* transformation of p(X)
- Logit has a linear relationship with X

8

How to find parameter values? – Maximum Likelihood

We use maximum likelihood to estimate the parameters.

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

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

Write the probability function of each Y and Xs using the logistic formula → Multiply all of the probability together to get the likelihood function → Maximize the likelihood by choosing the best parameter values

9

The output for a single variable

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

	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

The coefficient of balance: The odds of defaulting increases by $e^{0.55} = 1.73$ times when the balance increases by $1000.

The estimated standard deviation

The probability of observing such z-statistic given β_1 to be 0

10

Q1. Which factors are important in predicting Y?

H_0 : There is no relationship between X and Y

versus the *alternative hypothesis*

H_A : There is some relationship between X and Y.

- Mathematically, this corresponds to testing

$$H_0 : \beta_1 = 0$$

versus

$$H_A : \beta_1 \neq 0,$$

since if $\beta_1 = 0$ then the model reduces to $Y = \beta_0 + \epsilon$, and X is not associated with Y.

- The outcome of the test exhibits in p-values: if p-value is less than 5%, we are confident to reject the null hypothesis.
- That means, the alternative hypothesis is correct.

Given H_0 → Calculate z-statistic → P-value: whether to reject H_0 → There is / isn't a relationship between X and Y

11

Q2. How does each factor affect Y?

- If it is positive, it suggests that as X increases, the probability of Y = 1 also increases, and vice versa.
- There is no straightforward interpretation of the impact on the probability Y. Instead, we interpret β_1 as the average effect on the odds rather than the probability Y.
- The odds is defined as the ratio of the probability of Y = 1 to the probability of Y = 0

$the\ odds = \frac{p}{1-p} = e^{\beta_0+\beta_1 X+\varepsilon}$, where p is the probability of Y = 1

Examples: For probability of 0.5, the odds is 1.

- When X increases by 1, how will the odds change? We look at the change via *odds ratio*.

$$odds\ ratio = \frac{\overbrace{odds(X=x+1)}^{\text{After}}}{\underbrace{odds(X=x)}_{\text{Before}}} = \frac{\dfrac{p(X=x+1)}{1-p(X=x+1)}}{\dfrac{p(X=x)}{1-p(X=x)}} = \frac{e^{\alpha+\beta(X=x+1)+\varepsilon}}{e^{\alpha+\beta(X=x)+\varepsilon}} = e^{\beta}$$

12

Q2 Continued – For categorical variables

Lets do it again, using student as the predictor.

	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

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

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

- The odds of defaulting increases by $e^{0.4} = 1.51$ times when the borrower is a student, comparing to a non-student borrower.

13

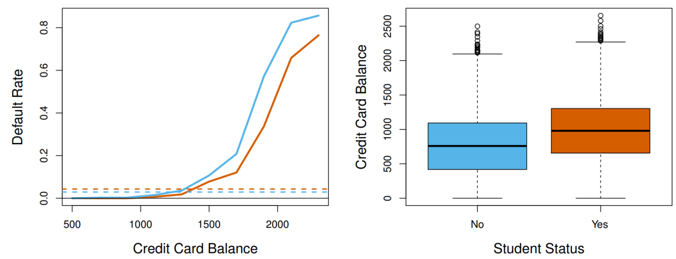
Multiple Logistic regression

	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, while it was positive in the one-variable model?
  - If we do not consider any other factors, students tend to default more.
  - If we take into consideration other factors, e.g., balance, further examination will find that these two are correlated: students have more balance than non-students
  - For the same level of balance, students are less likely to default
  - The positive effect in the single variable model captures the confounding effect of balance together with being a student
  - If we tease out the effect of balance by including it as a variable, we can find the effect of being a student by itself

14

The confounding effect of balance and student



- 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.

15

Q3. How to classify?

- It is straightforward to predict probability of Y = 1 using the estimated coefficients and the model
- The classification is based on probability values
  - Establishing cutoff level; If estimated prob. > cutoff, classify as “1”, e.g., if p>0.5, the prediction is classified as 1;
  - If the estimated probability is 0.67, the person is predicted to default the payment

16

Takeaways

- Classification and popular classifiers
- Logistic regression model specifications
- Key questions about the results
  - Which factors are more important predicting Y?
  - How does each factor affect Y?
  - How to classify?

17