Lecture 7 Logistic Regression I

STAT 441/505: Applied Statistical Methods in Data Mining

Linglong Kong

Department of Mathematical and Statistical Sciences
University of Alberta

Winter, 2016

Outline

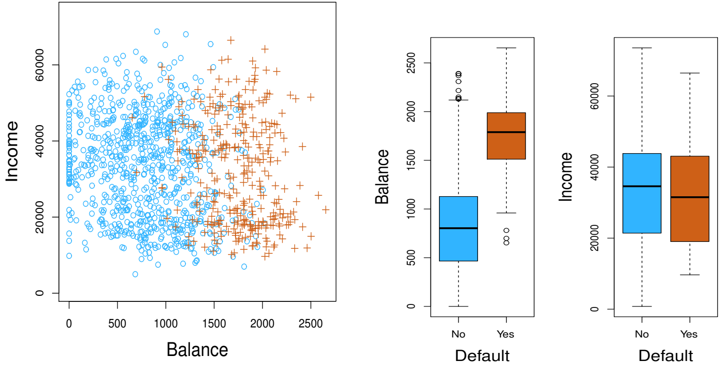
UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Qualitative Response

- There are many qualitative response taking values in an unordered set $\mathcal{C}$ such as
  `eye color ∈ {brown; blue; green}`.
- Given a feature vector X and a qualitative response Y taking values in the set $\mathcal{C}$, the classification task is to build a function $C(X)$ (learn a rule) that takes as input the feature vector X and predicts its value for Y; i.e. $C(X) \in \mathcal{C}$.
- Often we are more interested in estimating the probabilities that X belongs to each category in $\mathcal{C}$.
- 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.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Cedit Card Defualt



Individuals who defaulted in a given month in orange, and did not in blue.

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

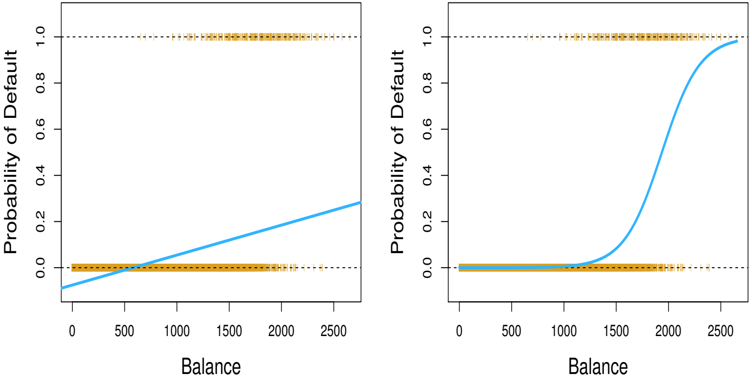
Linear Regression Model

- Suppose for the Default classification task that we code

$$Y = \begin{cases} 1 & \text{if Yes} \\ 0 & \text{if No} \end{cases}.$$

  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 discuss later.
- Since in the population $\mathrm{E}(Y_j|X = x) = \Pr(Y_j = 1|X = x)$, 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.

Credit data example



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

UNIVERSITY OF ALBERTA
EDMONTON·ALBERTA·CANADA

Logistic Regression

- Denote $p(X) = \Pr(Y_j = 1|X)$ consider using balance to predict default. Logistic regression uses the form

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

- 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.
- A bit of rearrangement gives

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

- This monotone transformation is called the log odds or logit transformation of $p(X)$.

Estimation

- We use maximum likelihood to estimate the parameters.

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

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

```
> glm.fit=glm(default~balance,data=defaultData,family=binomial)
> summary(glm.fit)
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
```

Interpretation

- Interpreting what β_1 means is not very easy with logistic regression, simply because we are predicting $\Pr(Y_j = 1|X)$ and not Y.
- If $\beta_1 = 0$, this means that there is no relationship between Y and X.
- If $\beta_1 > 0$, this means that when X gets larger so does the probability that $Y = 1$.
- If $\beta_1 < 0$, this means that when X gets larger, the probability that $Y = 1$ gets smaller.
- But how much bigger or smaller depends on where we are on the slope.

Hypothesis Testing

- We still want to perform a hypothesis test to see whether we can be sure that are β_0 and β_1 significantly different from zero.
- We use a z test instead of a t test, but of course that doesn't change the way we interpret the p-value
- Here the p-value for balance is very small, and β_1 is positive, so we are sure that if the balance increase, then the probability of default will increase as well.

```
                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
```

Prediction

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

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.5613+0.0055\times 1000}}{1+e^{-10.5613+0.0055\times 1000}} = 0.006.$$

- The predicted probability of default for an individual with a balance of \$1000 is less than 1%.
- For a balance of \$2000, the probability is much higher, and equals to $0.586(58.6\%)$.

$$\hat{p}(X) = \frac{e^{\hat{\beta}_0+\hat{\beta}_1 X}}{1+e^{\hat{\beta}_0+\hat{\beta}_1 X}} = \frac{e^{-10.5613+0.0055\times 2000}}{1+e^{-10.5613+0.0055\times 2000}} = 0.586.$$

```
> predict(glm.fit, list(balance = c(1000,2000)),type="response")
          1           2
0.005752145 0.585769370
```

Summary and Remark

- Introduction
- Logistic regression and estimation
- Hypothesis testing and prediction
- Read textbook Chapter 4 and `R` code
- Do `R` lab