Introduction to Data Science

DSA1101

Semester 1, 2018/2019
Week 10

Logistic regression

Logistic regression

- In linear regression modeling, the outcome variable is a continuous variable.
- When the outcome variable is categorical in nature, logistic regression can be used to predict the likelihood of an outcome based on the input variables.
- Although logistic regression can be applied to an outcome variable that represents multiple values, in this course we will focus on the case in which the outcome variable is binary (e.g. true/false, pass/fail, or yes/no).

Logistic regression

- For example, a logistic regression model can be built to determine if a person will or will not purchase a new automobile in the next 12 months.
- The training set could include input variables for a person's age, income, and gender as well as the age of an existing automobile.
- The training set would also include the outcome variable on whether the person purchased a new automobile over a 12-month period.
- The logistic regression model provides the likelihood or probability of a person making a purchase in the next 12 months.

Example: Medical

Source: The Straits Times

- Develop a model to determine the likelihood of a patient's successful response to a specific medical treatment or procedure. Input variables could include age, weight, blood pressure, and cholesterol levels.

Example: Finance

- Using a loan applicant's credit history and the details on the loan, determine the probability that an applicant will default on the loan. Based on the prediction, the loan can be approved or denied, or the terms can be modified.

Example: Marketing

Source: The Straits Times

- Determine a wireless customer's probability of switching carriers (known as churning) based on age, number of family members on the plan, months remaining on the existing contract, and social network contacts.

Example: Engineering

- Based on operating conditions and various diagnostic measurements, determine the probability of a mechanical part experiencing a malfunction or failure. With this probability estimate, schedule the appropriate preventive maintenance activity.

Logistic function

- Logistic regression is based on the logistic function

$$f(z) = \frac{\exp(z)}{1 + \exp(z)}, \text{for } -\infty < z < \infty.$$

- Note that as $z \to \infty$, $f(z) \to 1$.
- Also, as $z \to -\infty$, $f(z) \to 0$.

e^(-infinity) → 0

0/(1+0) = 0

f: R → (0, 1)

z consist of feature variables, do not want to restrict range → R.
(0, 1) so that we can easily compute the conditional probability for binary outcome.

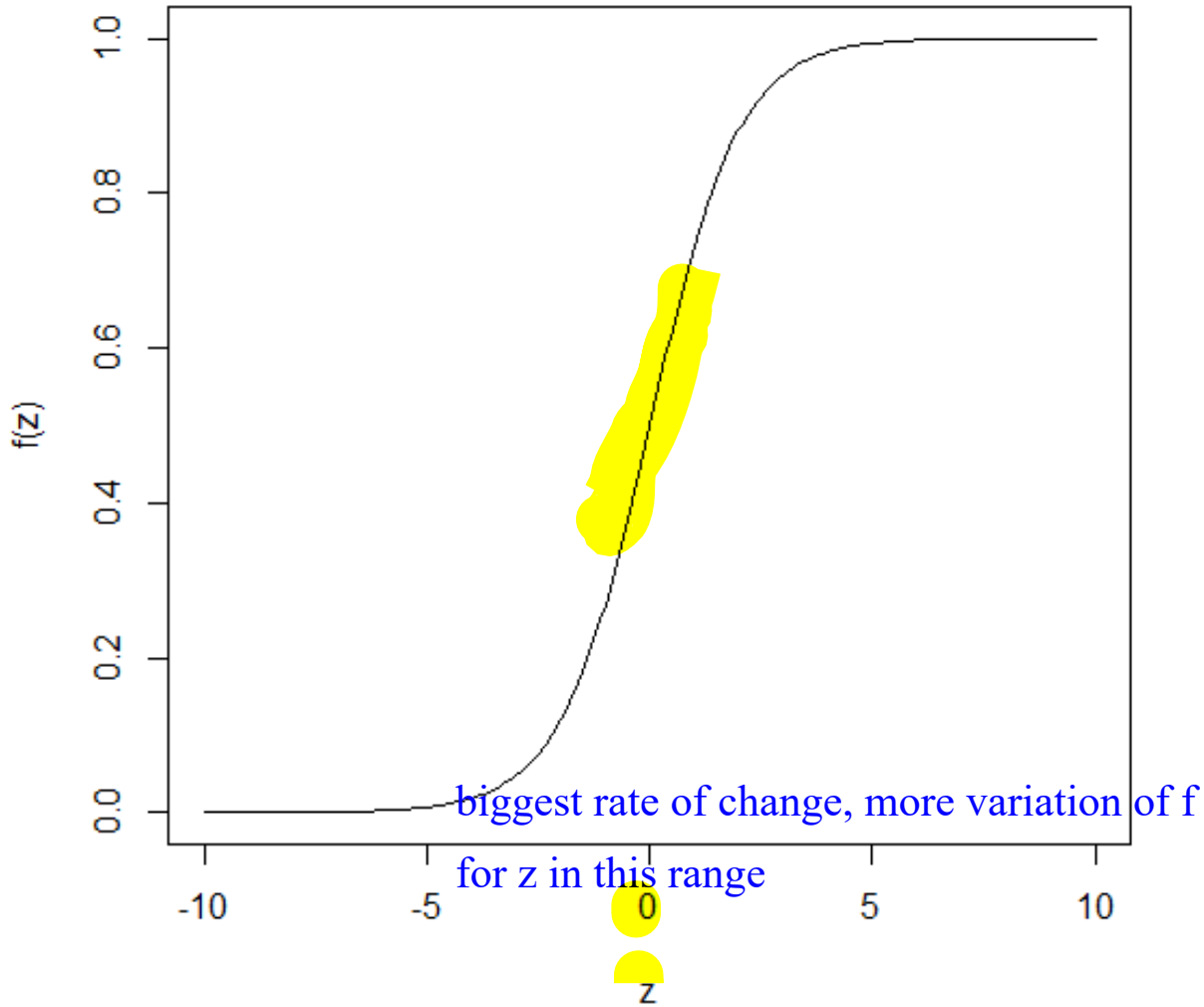
Logistic function

- We can plot the logistic function in R to visualize these properties.
-

```
logistic = function(z) {
  exp(z)/(1+exp(z))
}

z = seq(-10,10,0.1);
plot(z, logistic(z), xlab="z", ylab="f(z)", lty=1,
     type='l')
```

Logistic function



Logistic regression

- Recall that in (binary) naïve Bayes Classification, we compute probability score which is proportional to the conditional probability $P(Y = 1|X)$
- The conditional probability is generally bounded, $0 < P(Y = 1|X) < 1$.
- Because the range of $f(z)$ is $(0, 1)$, the logistic function appears to be an appropriate function to model $P(Y = 1|X)$ directly.
- As the value of z increases, the probability of the outcome occurring increases.

f(z)

Logistic function

- In any proposed model, to predict the likelihood of an outcome, z needs to be a function of the input or feature variables X.
- In logistic regression, z is expressed as a linear function of the input variables:

$$z = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p$$

- Then in logisitc regression, $P(Y = 1|X_1, X_2, ..., X_p)$ can be expressed as plug into the input, not output like linear regression

$$P(Y = 1|X_1, X_2, ..., X_p) = \frac{\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}{1 + \exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + ... + \beta_p X_p)}$$

Logistic regression: bounded function (0,1)

Linear regression: unbounded, may give very big output value.

Logistic function

- Just like linear regression, in logistic regression the parameters $\beta_0, \beta_1, \ldots, \beta_p$ need to be estimated based on training data.
- Instead of the method of *least squares*, parameter estimation in logistic regression is based on the method called *maximum likelihood estimation* (MLE).
- We will focus more on MLE in next week's lecture; for now let us look at a few examples of logistic regression in R.

Example: Customer Churn

- In our first example, a wireless telecommunications company wants to predict whether a customer will churn (switch to a different company) in the next six months.
- With a reasonably accurate prediction of a person‘s churning, the sales and marketing groups can attempt to retain the customer by offering various incentives.

Example: Customer Churn

- Data on 8,000 current and prior customers was obtained. The variables collected for each customer follow:

(i) Age (years)

(ii) Married (true/false)

(iii) Duration as a customer (years)

(iv) Churned_contacts (count)-Number of the customer's contacts that have churned (count)

(v) Churned (true/false)-Whether the customer churned

Example: Customer Churn

- The customer churn dataset is available as the CSV file 'churn.CSV' on the course website
-

```
> churn = read.csv("churn.CSV")
> head(churn)
  ID Churned Age Married Cust_years Churned_contacts
1  1       0  61       1          3                 1
2  2       0  50       1          3                 2
3  3       0  47       1          2                 0
4  4       0  50       1          3                 3
5  5       0  29       1          1                 3
6  6       0  43       1          4                 3
> summary(as.factor(churn$Churned))
   0    1
6257 1743
```

- About 21.8% of the customers churned.

Example: Customer Churn

- Logistic regression can be performed using the Generalized Linear Model function, `glm()`, in R
- Specify the family to be binomial, with the logit link.
-

```
1 Churn_logistic1 <- glm (Churned~Age + Married +
      Cust_years +
2 Churned_contacts, data=churn,
3 family=binomial(link="logit"))
```

Example: Customer Churn

- As in the linear regression case, there are tests to determine if the coefficients are significantly different from zero.
- Such significant coefficients correspond to small values of $\Pr(> |z - value|)$, which denote the p-value for the hypothesis test to determine if the estimated model parameter is significantly different from zero.
-

```
> summary(Churn_logistic1)

Call:
glm(formula = Churned ~ Age + Married + Cust_years + Churned_contacts,
    family = binomial(link = "logit"), data = churn)

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       3.415201   0.163734  20.858   <2e-16 ***
Age              -0.156643   0.004088 -38.320   <2e-16 ***
Married           0.066432   0.068302   0.973    0.331
Cust_years        0.017857   0.030497   0.586    0.558
Churned_contacts  0.382324   0.027313  13.998   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Customer Churn

- Rerunning this analysis without the 'Cust_years' variable, which had the largest corresponding p-value, yields the following:

```
> Churn_logistic2 <- glm (Churned~Age + Married + Churned_contacts,
+ data=churn, family=binomial(link="logit"))
> summary(Churn_logistic2)

Call:
glm(formula = Churned ~ Age + Married + Churned_contacts, family =
      binomial(link = "logit"),
    data = churn)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4476  -0.5178  -0.1972  -0.0723   3.3776

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       3.472062   0.132107  26.282   <2e-16 ***
Age              -0.156635   0.004088 -38.318   <2e-16 ***
Married           0.066430   0.068299   0.973    0.331
Churned_contacts  0.381909   0.027302  13.988   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Customer Churn

- Because the *p*-value for the 'Married' coefficient remains quite large, the 'Married' variable is dropped from the model.
- The following R code provides the third and final model, which includes only the 'Age' and 'Churned_contacts' variables:
-

```
> Churn_logistic3 <- glm (Churned~Age + Churned_contacts,
+ data=churn, family=binomial(link="logit"))
> summary(Churn_logistic3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.4599  -0.5214  -0.1960  -0.0736   3.3671

Coefficients:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)       3.502716   0.128430   27.27   <2e-16 ***
Age              -0.156551   0.004085  -38.32   <2e-16 ***
Churned_contacts  0.381857   0.027297   13.99   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

Example: Customer Churn

- We can use the estimated parameters from logistic regression to predict the probability of a customer churning based on the customer's age and the number of churned contacts.
- For example, for a customer who is 50 years old with 5 churned contacts, an estimate for the probablity of churning, $P(Y = 1|X)$, is

$$\approx \frac{\exp(\overbrace{3.50}^{\beta 0} - \overbrace{0.157}^{\beta 1} \times 50 + \overbrace{0.382}^{\beta 2} \times 5)}{1 + \exp(3.50 - 0.157 \times 50 + 0.382 \times 5)} = 0.08$$

Example: Customer Churn

- We can similarly predict churning for the customer with a specified threshold.
- For example, predict the customer to churn if the estimate for $P(Y = 1|X)$ exceeds 0.5.
- ROC curve can be similarly constructed by varying the threshold.