Introduction to Data Science

DSA1101

Semester 1, 2018/2019
Week 11

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: example

- Suppose we toss the coin n times.
- For the i^{th} toss, a dollar coin ($x_i = 1$) or a non-dollar coin ($x_i = 0$) is used.
- For the i^{th} toss, let $y_i = 1$ if it comes up a head and $y_i = 0$ if it comes up a tail.
- So we observe data $x = c(x_1, x_2, ..., x_n)$ and $y = c(y_1, y_2, ..., y_n)$.
- How to incorporate the binary feature variable X into the MLE framework that we discussed?

Logistic Regression

- To incorporate the single binary feature variable X, let

$$\pi(X) = P(Y = 1|X) = \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)}$$

so that

$$P(Y = 0|X) = 1 - \frac{\exp(\beta_0 + \beta_1 X)}{1 + \exp(\beta_0 + \beta_1 X)} = \frac{1}{1 + \exp(\beta_0 + \beta_1 X)}$$

- In general, X can be binary, categorical or continuous.
- Before estimating β_0 and β_1 via MLE, we need to look at the form of the likelihood function first:

$$L(\beta_0, \beta_1) = \prod_{i=1}^{n} \left[\frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)}\right]^{y_i} \left[\frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)}\right]^{1-y_i}$$

Logistic Regression

$$
\begin{aligned}
L(\beta_0, \beta_1) &= \prod_{i=1}^{n} \left[\frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right]^{y_i} \left[\frac{1}{1 + \exp(\beta_0 + \beta_1 x_i)} \right]^{1-y_i} \\
&= \prod_{i=1}^{n} \frac{\exp(\beta_0 + \beta_1 x_i)^{y_i}}{1 + \exp(\beta_0 + \beta_1 x_i)} \\
&= \prod_{i=1}^{n} \frac{\exp[y_i(\beta_0 + \beta_1 x_i)]}{1 + \exp(\beta_0 + \beta_1 x_i)}
\end{aligned}
$$

Logistic Regression

- The log-likelihood function is:

$$\ln L(\beta_0, \beta_1) = \sum_{i=1}^{n} \{y_i(\beta_0 + \beta_1 x_i) - \ln(1 + \exp(\beta_0 + \beta_1 x_i))\}$$

Review: Derivatives of the Exponential Function

- Recall that

$$\frac{d}{dx}e^x = e^x$$

and

$$\frac{d}{dx}e^{f(x)} = e^{f(x)}\frac{d}{dx}f(x)$$

- For example,

$$\begin{aligned}\frac{d}{dx}e^{x^2} &= e^{x^2}\frac{d}{dx}x^2 \\ &= \left(e^{x^2}\right)2x\end{aligned}$$

Logistic Regression

- The log-likelihood function is:

$$\ln L(\beta_0, \beta_1) = \sum_{i=1}^{n} \{y_i(\beta_0 + \beta_1 x_i) - \ln(1 + \exp(\beta_0 + \beta_1 x_i))\}$$

- To obtain the *maximum likelihood* estimates, we take the derivatives $\frac{\partial L}{\partial \beta_0}$ and $\frac{\partial L}{\partial \beta_1}$.
- Note that

$$\frac{\partial}{\partial \beta_0} \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) = \sum_{i=1}^{n} y_i$$

and

$$\frac{\partial}{\partial \beta_1} \sum_{i=1}^{n} y_i(\beta_0 + \beta_1 x_i) = \sum_{i=1}^{n} x_i y_i$$

Logistic Regression

- In addition,

$$
\begin{aligned}
& \frac{\partial \log[1+\exp(\beta_0+\beta_1 X_i)]}{\partial \beta_0} \\
= & \frac{\partial[1+\exp(\beta_0+\beta_1 X_i)]/\partial\beta_0}{1+\exp(\beta_0+\beta_1 X_i)} = \frac{\partial \exp(\beta_0+\beta_1 X_i)/\partial\beta_0}{1+\exp(\beta_0+\beta_1 X_i)} \\
= & \frac{\exp(\beta_0+\beta_1 X_i)}{1+\exp(\beta_0+\beta_1 X_i)} = \pi[X_i] \\
& \frac{\partial \log[1+\exp(\beta_0+\beta_1 X_i)]}{\partial \beta_1} = \frac{\partial[1+\exp(\beta_0+\beta_1 X_i)]/\partial\beta_1}{1+\exp(\beta_0+\beta_1 X_i)} \\
= & \frac{\partial \exp(\beta_0+\beta_1 X_i)/\partial\beta_1}{1+\exp(\beta_0+\beta_1 X_i)} \\
= & X_i \frac{\exp(\beta_0+\beta_1 X_i)}{1+\exp(\beta_0+\beta_1 X_i)} = X_i \pi[X_i]
\end{aligned}
$$

x

Logistic Regression

- Therefore, the *maximum likelihood* estimates $\hat{\beta}_0$ and $\hat{\beta}_1$ are solutions to the equations

$$\sum_{i=1}^{n} \left\{ y_i - \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right\} = 0,$$

and

$$\sum_{i=1}^{n} \left\{ x_i y_i - x_i \frac{\exp(\beta_0 + \beta_1 x_i)}{1 + \exp(\beta_0 + \beta_1 x_i)} \right\} = 0.$$

- These equations are also known as score functions and are typically solved by numerical methods such as the *Newton-Raphson* algorithm.

Logistic Regression: interpretation

- Suppose $p = P(Y = 1)$ denotes the probability of an event happening (e.g. coin toss turns up as head).
- Then the odds of the event occurring is given by

P(happen)/ P(not happen)

fair coin -> odds = 1

$$\text{odds} = \frac{p}{1-p}.$$

- To incorporate feature variables, let $\pi(x) = P(Y = 1|x) = \frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}$
- Then the conditional odds of the event occuring is given by

$$\text{conditional odds} = \frac{\pi(x)}{1-\pi(x)}.$$

Logistic Regression: interpretation

- The conditional odds of the event occuring is given by

$$\text{conditional odds} = \frac{\pi(x)}{1-\pi(x)}.$$

$$\begin{aligned}\frac{\pi(x)}{1-\pi(x)} &= \left[\frac{\exp(\beta_0+\beta_1 x)}{1+\exp(\beta_0+\beta_1 x)}\right] \div \left[\frac{1}{1+\exp(\beta_0+\beta_1 x)}\right] \\ &= \exp(\beta_0+\beta_1 x) \\ \rightarrow \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] &= \beta_0+\beta_1 x.\end{aligned}$$

Logistic Regression: interpretation

- Therefore, we have that the log odds, $\ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1 x.$
- We want to compare the odds of the event occurring per unit change in the feature variable X.
- $\ln\left[\frac{\pi(x+1)}{1-\pi(x+1)}\right] - \ln\left[\frac{\pi(x)}{1-\pi(x)}\right] = \beta_0 + \beta_1(x+1) - (\beta_0 + \beta_1 x) = \beta_1$
- Therefore, β_1 equals to the log odds ratio for unit change in X:

$$\ln\left[\frac{\frac{\pi(x+1)}{1-\pi(x+1)}}{\frac{\pi(x)}{1-\pi(x)}}\right] = \beta_1$$

- Equivalently, the odds ratio equals $\exp(\beta_1)$:

$$\left[\frac{\frac{\pi(x+1)}{1-\pi(x+1)}}{\frac{\pi(x)}{1-\pi(x)}}\right] = \exp(\beta_1)$$

Calculating the odds ratio: example

- Suppose both the outcome Y and feature variable X are binary, then the odds ratio for $Y = 1$ occurring for $X = 1$ versus $X = 0$ is

$$\left[\frac{\frac{P(Y=1|X=1)}{1-P(Y=1|X=1)}}{\frac{P(Y=1|X=0)}{1-P(Y=1|X=0)}}\right]$$

Calculating the odds ratio: example

- Continuing with our example, suppose we tossed the coin 100 times, and the results are tabulated as follows

	$x = 1$	$x = 0$
$y = 1$	30	40
$y = 0$	20	10

- We can estimate the odds ratio using the tabular data above (also known as a contingency table)

Calculating the odds ratio: example

- Continuing with our example, suppose we tossed the coin 100 times, and the results are tabulated as follows

	$x = 1$	$x = 0$
$y = 1$	30	40
$y = 0$	20	10

$300/800 = 0.375$

- An estimate for $P(Y = 1|X = 1)$ is $\frac{30}{30+20} = 0.6$
- An estimate for $P(Y = 1|X = 0)$ is $\frac{40}{40+10} = 0.8$
- Therefore the estimated odds ratio is given by

$$\left[\frac{\frac{0.6}{1-0.6}}{\frac{0.8}{1-0.8}}\right] = 0.375.$$

Example in R: Customer Churn

- In our earlier 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
```

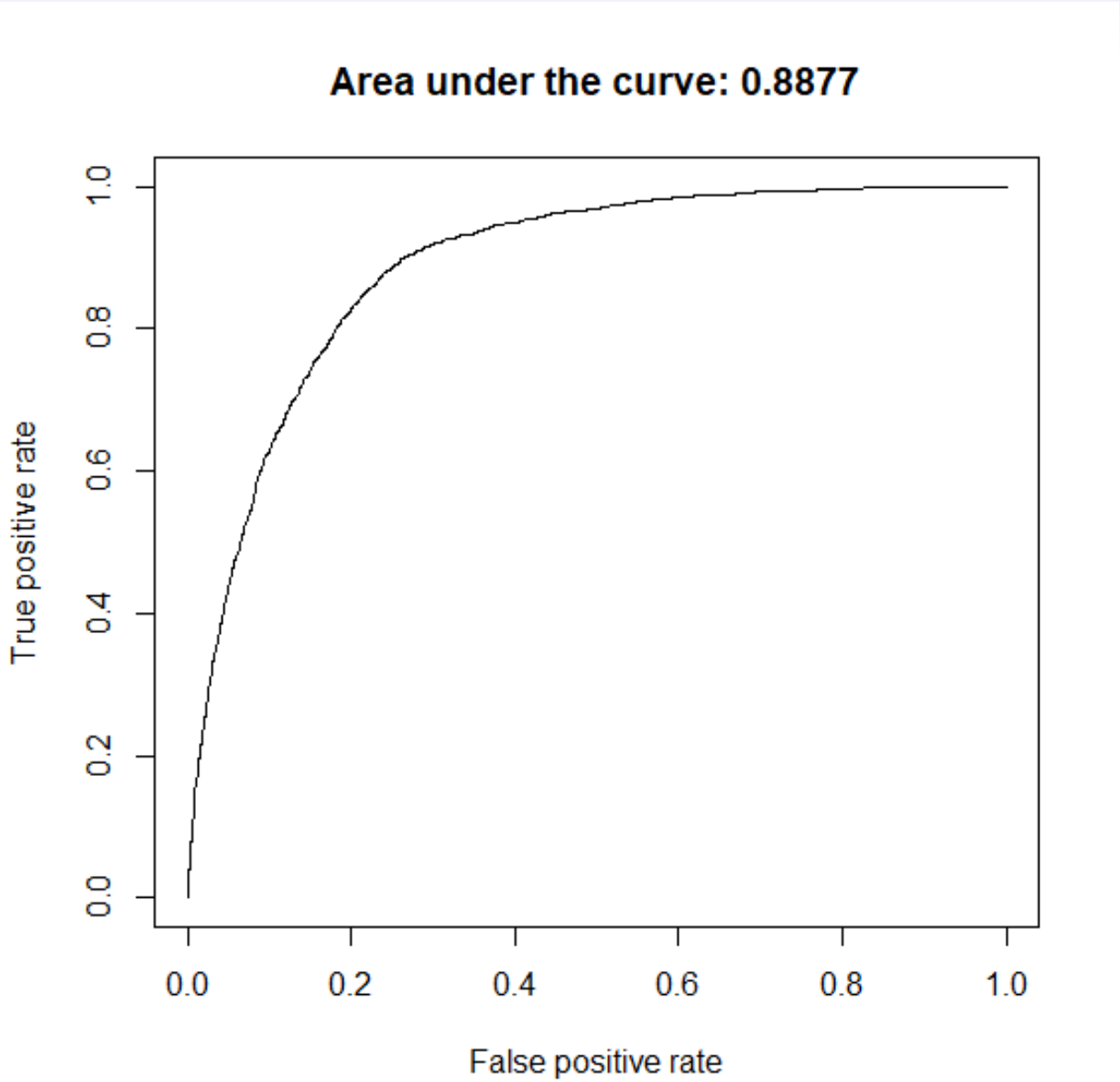
Example: Customer Churn

- We can also compute the AUC for classification based on logistic regression
-

```
library(ROCR)
pred = predict(Churn_logistic1, type="response")

predObj = prediction(pred, churn$Churned )
rocObj = performance(predObj, measure="tpr", x.measure="fpr")
aucObj = performance(predObj, measure="auc")
plot(rocObj, main = paste("Area under the curve:",
round(aucObj@y.values[[1]] ,4)))
```

Example: Customer Churn



Example: Customer Churn

- To illustrate how the FPR and TPR values are dependent on the threshold value used for the classifier, we can use the following R code to produce the plots:

```
# extract the alpha(threshold), FPR, and TPR values from rocObj
alpha <- round(as.numeric(unlist(rocObj@alpha.values)),4)
fpr <- round(as.numeric(unlist(rocObj@x.values)),4)
tpr <- round(as.numeric(unlist(rocObj@y.values)),4)
# adjust margins and plot TPR and FPR
par(mar = c(5,5,2,5))
plot(alpha,tpr, xlab="Threshold", xlim=c(0,1),
ylab="True positive rate", type="l")
par(new="True")
plot(alpha,fpr, xlab="", ylab="", axes=F, xlim=c(0,1), type="l" )
axis(side=4)
mtext(side=4, line=3, "False positive rate")
text(0.18,0.18,"FPR")
text(0.58,0.58,"TPR")
```

Example: Customer Churn