

Lab 4

IMC 490: Machine Learning for IMC

4/19/2017

In this lab, we'll be going over:

- logistic regression
- making predictions
- logistic regression which requires data transformation (bottle return problem)

Logistic Regression

In this exercise, we will be using the Titanic data - a classic dataset with information on each passenger aboard the Titanic at the time of its sinking. We will build a model to predict the likelihood of survival for a passenger based on features like sex, age, class, and so on.

The dataset we are using is a cleaned version of the dataset used in the popular Kaggle competition, *Titanic: Machine Learning from Disaster*. <https://www.kaggle.com/c/titanic>

```
titanic = read.csv("titanic_cleaned.csv")
str(titanic)
```

```
## 'data.frame':    1045 obs. of  6 variables:
## $ pclass   : int  1 1 1 1 1 1 1 1 1 1 ...
## $ survived: int  1 1 0 0 0 1 1 0 1 0 ...
## $ name     : Factor w/ 1043 levels "Abbing, Mr. Anthony",...: 22 24 25 26 27 31 46 47 51 55 ...
## $ sex      : Factor w/ 2 levels "female","male": 1 2 1 2 1 2 1 2 1 2 ...
## $ age      : num  29 0.92 2 30 25 48 63 39 53 71 ...
## $ fare     : num  211 152 152 152 152 ...
```

In preparation for the Kaggle competition, we're going to get more hands on with handling the data in this lab. Download the dataset "titanic_cleaned.csv" from the Lab4 folder on Canvas, and do the following:

1. Read the data into R.
2. Inspect the structure of the data.
3. Get rid of the **name** column - there are too many levels to regress on, and including the names of the victims makes this exercise way too real.
4. The passenger class, **pclass**, should be a categorical feature, but since it is coded with the numbers 1-4, it is read as an integer field. Convert it to categorical.

```
# after cleaning
str(titanic)
```

```
## 'data.frame':    1045 obs. of  5 variables:
## $ pclass   : Factor w/ 3 levels "1","2","3": 1 1 1 1 1 1 1 1 1 1 ...
## $ survived: int  1 1 0 0 0 1 1 0 1 0 ...
## $ sex      : Factor w/ 2 levels "female","male": 1 2 1 2 1 2 1 2 1 2 ...
## $ age      : num  29 0.92 2 30 25 48 63 39 53 71 ...
## $ fare     : num  211 152 152 152 152 ...
```

Exercises:

0. Generate a contingency table to analyze the effect of a passenger's gender on probability of survival. Are males or females more likely to survive? hint: use `table()`
1. Fit a logistic regression model to predict the probability of survival using only `sex` as the predictor and print the summary.
2. Write the regression equation.
3. Obtain the odds ratio for β_{sex} . What is the interpretation of this value?
4. Fit a logistic regression model to predict the probability of survival using all of the available predictors.
5. Predict the probability of survival for Eric, a 21-year old male riding economy class (3rd) with a \$20 ticket.
6. Find the log odds value for the previous prediction. Use this to manually verify the predicted probability.

Logistic Regression (requiring data transformation)

A zoologist is researching squirrels. In a study on squirrels' favorite foods, the zoologist chooses three types of food - almonds, acorns, and cashews - and attempts to hand feed the squirrels around campus. The zoologist records data on the number of squirrels he approaches (`num_approached`), and the number of squirrels that take the food from him (`num_fed`). **Let's build a simple logistic regression to predict the probability of a squirrel taking food from the zoologist, depending on the type of the food.**

```
squirrels = data.frame(num_approached = c(12, 15, 13),  
                        food = as.factor(c("almonds", "acorns", "cashews")),  
                        num_fed = c(2, 13, 6))
```

Exercises:

1. Transform the data into a form suitable for logistic regression.
2. Fit the regression and print the summary.
3. Looking at the regression coefficients, which food appears to be the squirrels' favorite? Is this result consistent with your intuitive conclusion from the collected data?