Logistic Regression

Logistic Regression is used for classification. In linear regression the response varible (Y) is quantitative, but in many situations the response variable can be qualitative or categorical. Predicting a qualitative response for an observation can be referred to as classifying that observation, since it involves assigning the observation to a category, or class. On the other hand, often the methods used for classification first predict the probability of each of the categories of a qualitative variable, as the basis for making the classification. In this sense they also behave like regression methods.

There are many possible classification techniques, or classifiers, that one might use to predict a qualitative response. Three of the most qualitative classification classifiers widely-used are: logistic regression, linear discriminant analysis, and logistic K-nearest neighbors.

Example Classification problems:

1. A person arrives at the emergency room with a set of symptoms that could possibly be attributed to one of three medical conditions. Which of the three conditions does the individual have?
2. An online banking service must be able to determine whether or not a transaction being performed on the site is fraudulent, on the basis of the user's IP address, past transaction history, and so forth.
3. On the basis of DNA sequence data for a number of patients with and without a given disease, a biologist would like to figure out which DNA mutations are deleterious (disease-causing) and which are not.

We use the logistic function to output a value ranging from 0 to 1. Based off of the probability we assign a class.

Sigmoid function: always produces results in a probability from 0 to 1 of belonging in the 1 class. That means we can set a cutoff point at 0.5, anything below it results in class 0, anything above is class 1.

After you train a logistic regression model on some training data, you will evaluate your model's performance on some test data. You can use a confusion matrix to evaluate classification models.

Terminology

True Positive (TP) cases in which we perdicted positive and in reality the the test is positive

True Negative (TN) cases in which we perdicted negative and in reality the test is negative

False Positives (FP) cases in which we predicted positive and in reality the test is negative (type 1 error) (telling a man they are pregnant)

False Negatives (FN) cases in whch we predicted negative and in reality the test is positive (type 2 error) (telling a pregnant women that is visablly showing that she is not pregnant)

Misclassification rate (error rate) measures overall how often we are wrong.

Measuring Accuracy

To calculate the accuracy of the model you would take (TP + TN) / total

To calculate the misclassifiation rate you would take (FP + FN) / total

Getting started

You will need to import the following libraries:

1. tidyverse
2. ggthemes
3. Amelia

4. kableExtra

Explore the data

```
kable(head(df.train))
```

PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch
1	0	3	Braund, Mr. Owen Harris	male	22	1	0
2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Thayer)	female	38	1	0
3	1	3	Heikkinen, Miss. Laina	female	26	0	0
4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35	1	0
5	0	3	Allen, Mr. William Henry	male	35	0	0
6	0	3	Moran, Mr. James	male	NA	0	0

Source: Udemy - Data Science and Machine Learning Bootcamp with R (Jose Portilla)

An Intro to Statistical Learning with Applications in R (Gareth James, et al.) http://www-bcf.usc