Logistic Regression

Akshata Mohan, Valerie Roth

WiDS Oct 20th 2017

References

http://www.stat.cmu.edu/~cshalizi/uADA/12/lectures/ch12.pdf

Previously what we did in Linear Regression

- Predicting continuous outcomes - conditioning Y on some values of X - conditional expectations
- How do we get Y to be discrete taking on only certain values?

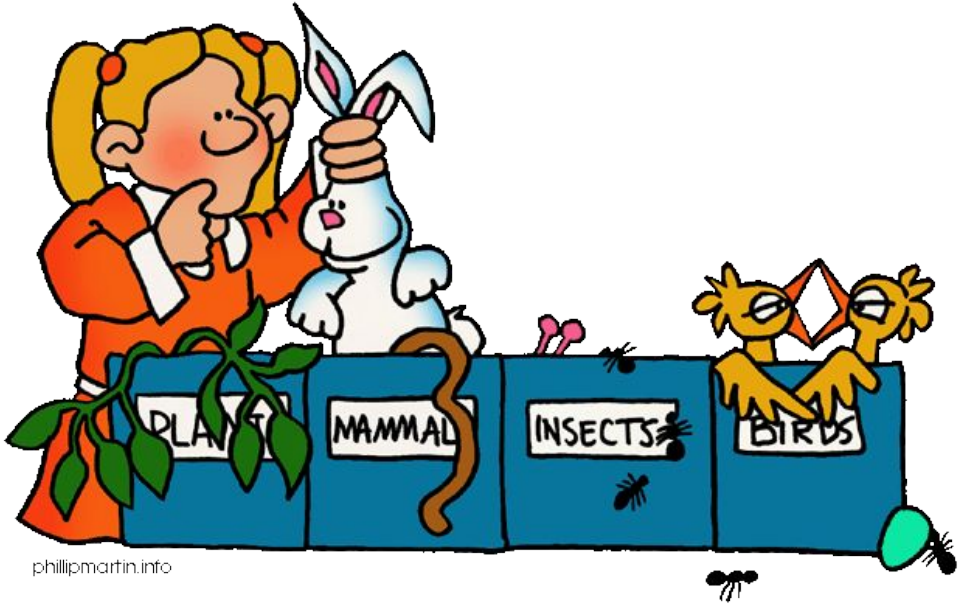
Why wasn’t it enough?

- Linear regression cannot give you discrete outcomes.
- Assigning class labels to data - classification
- Basic model - “yes” or “no”
- Logistic regression outputs probabilities - quantifies uncertainty
- Conditional distribution of Y given input X - Y is response class, X are predictors. Pr(Y/X) this is what logistic regression gives us.
- P(Y|X) - measure “precision” of prediction

Bottom line -

Image credits - clipartpanda.com

CLASSIFICATION!

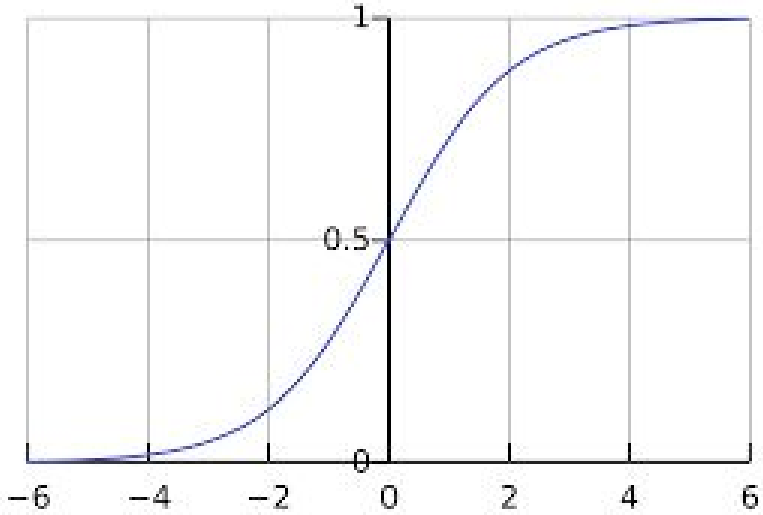


Secret sauce

$$\log \frac{p(x)}{1-p(x)} = \beta_0 + x \cdot \beta$$

Sigmoid function

Image credits - Wikipedia



How does this make sense?

- Linear dependence but 0,1 as o/p
- Sigmoid function [0,1]
- t is the linear regression equation

$$S(t) = \frac{1}{1 + e^{-t}}.$$

Fitting the parameters and model assumptions

- We can fit it using gradient descent or Newton’s method or MLE
- Y’s are independent and are either a 0 or a 1 => Bernoulli distribution
- A linear relationship between logit of Y and X’s. And NOT between Y and X’s
- Cannot have homoscedasity of errors. But errors must be independent

Decision boundary

- Find B’s
- We plug coefficient estimates and find the equation representing decision boundary.
- $S(t) > 0.5$ for $Y = 1$

Coefficient interpretation

- The effect of a unit change in xi is to increase the odds of a response multiplicatively by the factor exp(Bi)

Logistic regression using sklearn

See ipython notebook

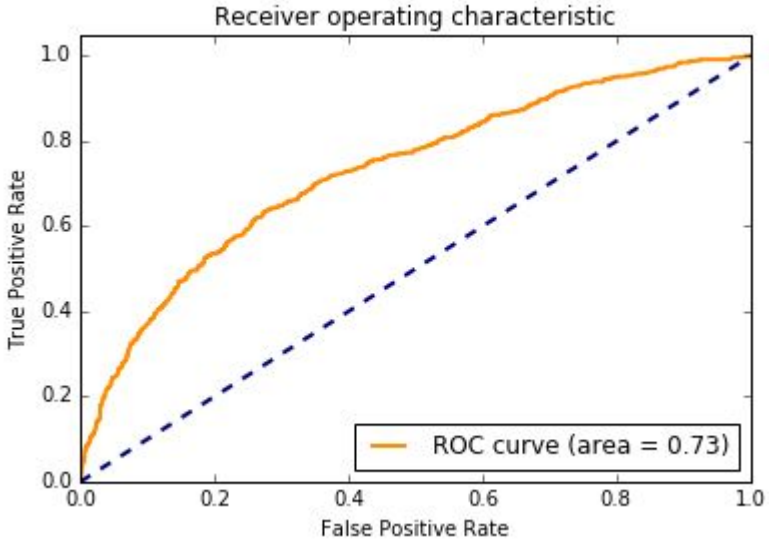
Measuring performance

- Accuracy [tp + tn /(tp+tn+fp+fn)]
- ROC - receiver operating characteristic curves - tells us about the performance of a classifier as the threshold is changed

- TPR or sensitivity (y axis) vs FPR or specificity (x axis). TPR - recall is the fraction of relevant instances that are retrieved. FPR is probability of false alarm

ROC curves

Image from ipython notebook



Multi class logistic regression

- K classes, K-1 different logistic regression equations are estimated.
- The test record is assigned to the class with the highest probability from one of the K-1 different equations