Week 5

Agenda

1. Async Review
2. Logistic Regression discussion
3. Evaluation discussion
4. Breiman Paper discussion
5. Time permitting: notebook or case study

Quizzes as Warmup

Logistic regression is an appropriate method to use when trying to predict a continuous dependent variable given one or more binary independent variables

True

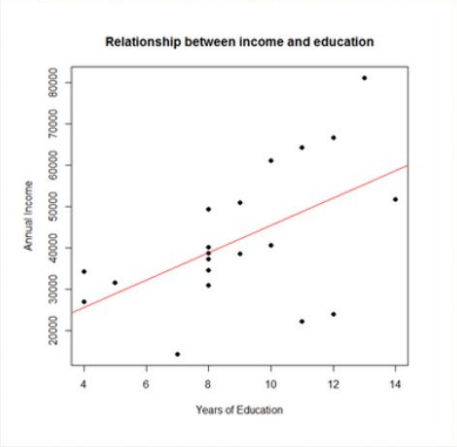
False

The regression line can be completely summarized by two coefficients: the _____ and the _____.

The first blank is

The second blank is

Based on the data and regression line in the figure below, what is the actual income for the individual with 7 years of education? What is the predicted income for an individual with 7 years of education?



Ordinary Least Squares regression minimizes

The total absolute error

The sum of the squared error

The slope and intercept of the regression line

The computational complexity of the algorithm

Linear & Logistic Regression

Types of Regression

- **Linear regression**: assumes a linear relation between independent and dependent variables
- **Bivariate**: exactly one independent and dependent variable

$$y_i = a + bx_i + \varepsilon_i$$

- **Multiple**: linear regression with multiple independent variables

$$y_i = \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i = \mathrm{x}_i^{\mathrm{T}}\beta + \varepsilon_i,\ i = 1, \ldots, n$$

- **Logistic regression**: binary dependent variable

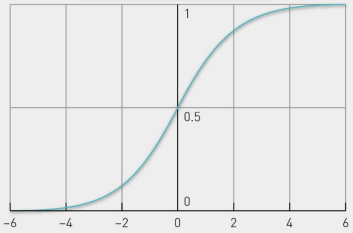
$$\operatorname{logit}\left(p_i\right) = \ln\left(\frac{p_i}{1-p_i}\right) = \beta_0 + \beta_1 x_{1,i} + \cdots + \beta_M x_{m,i}$$

Interpreting α and β

- In linear regression, β is the causal effect of a one-unit increase in x on y.
- In logistic regression, β is the change in the odds ratio.
  - For a one-unit increase in x, we expect to see a $1 - e^{\beta}\%$ change in y.

1. Have you trained linear regression models in the past? Tell us about your experience.
2. What are the parameters? Hyperparameters?
3. How do you interpret the parameters?
4. What are some of the characteristics of the logistic function?

How Does Logistic Regression Work?



$$P = \frac{e^{\alpha+\beta X}}{1+e^{\alpha+\beta X}}$$

- Transforms the continuous infinite scale into a scale between 0 and 1.

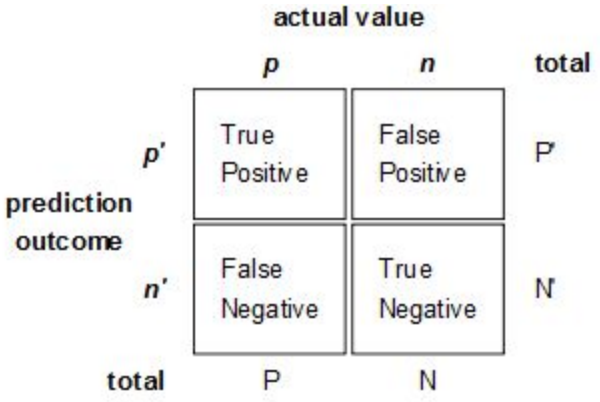
Logistic Regression

Maximum Likelihood Estimation

1. Computer picks initial parameters, α and β.
2. Determines likelihood of data, given chosen parameters.
3. Improves parameter estimates incrementally (e.g., Newton's method or gradient descent).
4. Recomputes likelihood of data, given these new parameters.
5. When parameters cease to change significantly, we tell the computer to stop presuming we have reached a minimum or maximum.

1. What makes one set of parameters better than another?
2. Describe the MLE approach to finding a good set of parameters.
3. What are alternatives to MLE?

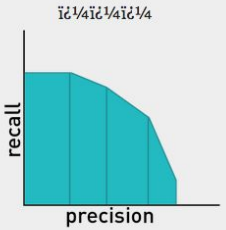
Evaluation



Video Slide Presentation

in Supervised Learning

- If you have supervised data, you will want to maximize an objective function.
  - **Precision**: $TP \div (TP + FP)$ % positives correctly identifed
  - **Recall**: $TP \div (TP + FN)$ % existing positives identified
  - **Optimal point** on ROC (precision/recall) curve
  - **Accuracy**: $(TP + TN) \div (TP + TN + FP + FN)$
  - **F-test**: $2 \cdot (P \cdot R) \div (P + R)$



- Training data allows you to maximize your objective.

Using and Evaluating Logistic Regression Models

Confusion Matrix

- Breaking down accuracy: TP, TN, FP, FN
- Many lenses built on top of confusion matrix to provide different views of goodness of classifier

Precision/Recall (P/R)

- What is recall?
- What is precision (accuracy @ threshold)?
- Most important for spam detection?
- Most important for credit worthiness prediction?
- Most important for Google search results?

Thresholds

- Threshold setting reflects concern over precision vs. recall
- Calibration of probabilities and retraining a model
- Let's talk about ROC...

The AUC

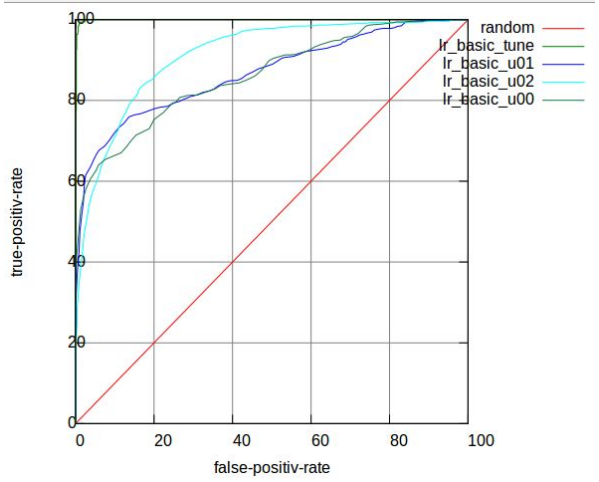
sensitivity or true positive rate (TPR)

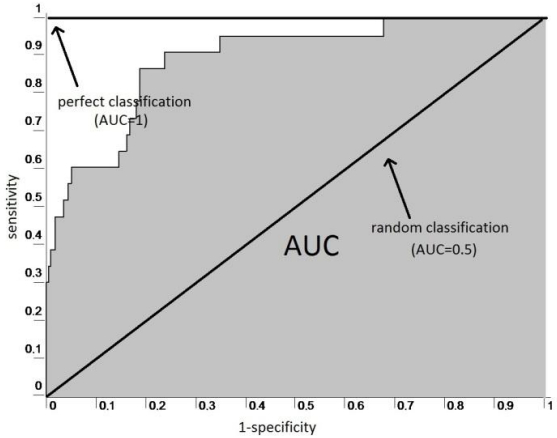
eqv. with hit rate, recall

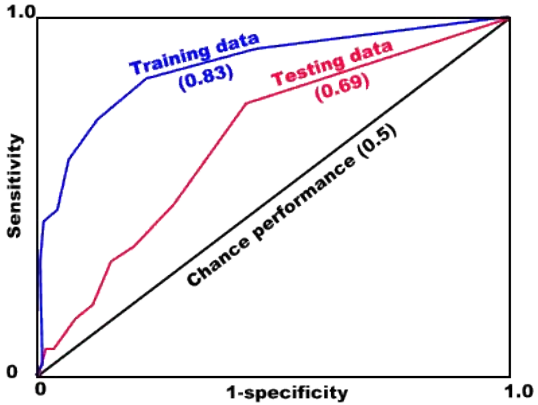
$$TPR = TP/P = TP/(TP + FN)$$

specificity (SPC) or true negative rate

$$SPC = TN/N = TN/(TN + FP)$$





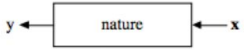


1. Describe the role of the threshold when comparing classifiers.
2. What is the TPR? FPR?
3. Why is the curve of TPR vs FPR generated by changing the threshold monotonically increase?
4. When is the AUC useful and when might it not be?

‘Two Cultures’ Paper

1. INTRODUCTION

Statistics starts with data. Think of the data as being generated by a black box in which a vector of input variables **x** (independent variables) go in one side, and on the other side the response variables **y** come out. Inside the black box, nature functions to associate the predictor variables with the response variables, so the picture is like this:



There are two goals in analyzing the data:

Prediction. To be able to predict what the responses are going to be to future input variables;
Information. To extract some information about how nature is associating the response variables to the input variables.

There are two different approaches toward these goals:

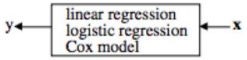
The Data Modeling Culture

The analysis in this culture starts with assuming a stochastic data model for the inside of the black box. For example, a common data model is that data are generated by independent draws from

response variables = f(predictor variables, random noise, parameters)

Leo Breiman is Professor, Department of Statistics, University of California, Berkeley, California 94720-4735 (e-mail: leo@stat.berkeley.edu).

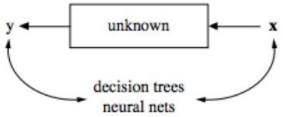
The values of the parameters are estimated from the data and the model then used for information and/or prediction. Thus the black box is filled in like this:



Model validation. Yes–no using goodness-of-fit tests and residual examination.
Estimated culture population. 98% of all statisticians.

The Algorithmic Modeling Culture

The analysis in this culture considers the inside of the box complex and unknown. Their approach is to find a function $f(\mathbf{x})$—an algorithm that operates on **x** to predict the responses **y**. Their black box looks like this:



Model validation. Measured by predictive accuracy.
Estimated culture population. 2% of statisticians, many in other fields.

In this paper I will argue that the focus in the statistical community on data models has:

- Led to irrelevant theory and questionable scientific conclusions;

- Kept statisticians from using more suitable algorithmic models;
- Prevented statisticians from working on exciting new problems;

I will also review some of the interesting new developments in algorithmic modeling in machine learning and look at applications to three data sets.

between inputs and outputs than data models. This is illustrated using two medical data sets and a genetic data set. A glossary at the end of the paper explains terms that not all statisticians may be familiar with.

3. PROJECTS IN CONSULTING

1. What are the ‘Two Cultures’?
2. Give examples of when one approach might be more reasonable than the other?

8. RASHOMON AND THE MULTIPLICITY OF GOOD MODELS

Rashomon is a wonderful Japanese movie in which four people, from different vantage points, witness an incident in which one person dies and another is supposedly raped. When they come to testify in court, they all report the same facts, but their stories of what happened are very different

9. OCCAM AND SIMPLICITY VS. ACCURACY

Occam's Razor, long admired, is usually interpreted to mean that simpler is better. Unfortunately, in prediction, accuracy and simplicity (interpretability) are in conflict. For instance, linear regression gives a fairly interpretable picture of the $\mathbf{y}, \mathbf{x}$ relation. But its accuracy is usually less than that of the less interpretable neural nets. An example closer to my work involves trees.

10. BELLMAN AND THE CURSE OF DIMENSIONALITY

The title of this section refers to Richard Bellman's famous phrase, "the curse of dimensionality." For decades, the first step in prediction methodology was to avoid the curse. If there were too many prediction variables, the recipe was to find a few features (functions of the predictor variables) that "contain most of the information" and then use these features to replace the original variables. In procedures common in statistics such as regression, logistic regression and survival models the advised practice is to use variable deletion to reduce

Final Thoughts?