Department of Statistics

Stats 784 Data Mining

Mid-term test 2016: Model answers

Answer all 10 questions in the boxes provided. Keep your answers short and succinct. Each question is worth 5 marks.

1. In the first lecture, several definitions of data mining were given. What were the common themes in these definitions?

Discovery of unknown patterns/relationships
Discovery of useful information
Big data
Results should be actionable
Computer-driven

2. We discussed the '' data mining process". What were the steps in this process?

Question posed
Data requirements identified
Data acquired, cleaned, formatted
Model building/prediction
Communication
Implementation

3. Define what is mean by **apparent error** and **test error**. How do they differ? Which one would you expect to be bigger than the other? Why?

Apparent error: fit model on training set, use training set to compute PE

Test error: fit model on training set, use test set to compute PE

Apparent error usually less than test error (as the noise in the training set is modelled to some extent)

4. What is the difference between 5-fold and 10-fold cross-validation? Which has the smaller bias? The smaller variance?

In 5-fold CV, the data is divided into 5 parts, each part used in turn as a test set, the rest as training. In 10-fold CV, the data are divided into 10 parts.

5-fold CV has more bias, less variance than 10-fold CV

5. Describe how the err+opt bootstrap estimate of prediction error is calculated.

Let PE(data1, data2) be the PR calculated using data1 to train the predictor, and data2 to calculate the PE. Then the err+opt estimate is

PE(training, training) + Average{ PE(bs, training) - PE(bs, bs) }

where we average over several bootstrap samples (denoted by bs)

6. What changes have to be made to the rpart algorithm when changing from the prediction of a continuous target to classification?

First, we change the splitting criterion: for classification, we choose the split that leads to the biggest decrease in node impurity (as measured by the Gini index).

Second, the value of the function on a region is the majority class of all points in the region.

7. What are basis functions? What basis functions are used in MARS?

Basis functions are a set of functions used to model more general functions as linear combinations of the basis functions.

For MARS, the basis functions are the "broken stick" functions of the form

$$\beta_1(x-t)_+ + \beta_2(t-x)_+$$

and products of these.

8. Describe the backfitting algorithm. You may assume there are only two features.

Step 1. Set $\phi_1(x_1) = x_1$, $\phi_2(x_2) = x_2$.

Step 2. Calculate $r = y - \phi_2(x_2)$ and smooth the plot of r versus x_1 to get ϕ_1.

Step 3. Calculate $r = y - \phi_1(x_1)$ and smooth the plot of r versus x_2 to get ϕ_2.

Repeat 2-3 until no change.

9. Describe the basic idea behind bagging. When applying bagging to trees, what extra feature is used (over basic bagging) in the random forest implementation? What is its purpose?

Bootstrap Aggregation: We calculate predictors based on different bootstrap samples and average (or "majority vote") the results.

For random forests, we select the splits using a random sample of predictors at each stage.

This means the predictions corresponding to different bootstrap samples are less correlated and we get more benefit from the averaging.

10. When boosting trees we use small trees. In random forests, do we use large or small trees? Why?

We use large trees as they are less biased. They are more variable, but we reduce the variance by the averaging process.