

ICS 278 Homework 2

Data Mining, ICS 278, Spring 2006

Due Date: Tuesday May 9th, in class

Problem 1: Analyzing The Time and Space Complexity of Several Well-Known Classification Algorithms

Estimate the time and space complexity (“big 0” notation, worst-case analysis) for the models below, specifically:

- time and space complexity in training the model;
- space complexity in storing the model;
- time complexity in making 1 prediction.

In each case assume there are N data points, p predictor variables, and M class values. In some cases (e.g., for trees or k NN) you may need to make some additional assumptions (e.g., allowing for some additional “offline” time and space cost to pre-index the data)—if so, just be sure to clearly state what assumptions you are making and how such preprocessing is done.

1. Linear regression (linear in the p input variables) with p real-valued variables.
2. A naive Bayes classifier where each of the p variables are nominal, each taking K values.
3. A k -nearest neighbor classifier with $k = 1$, and all p variables are real-valued.
4. A decision-tree classifier that greedily builds trees of depth d using the Gini node-splitting function discussed in class, with all p variables again being real-valued (ignore the other aspects of typical tree learning algorithms such as pruning). Assume for simplicity that the tree is binary and approximately balanced (i.e., that approximately half of the examples are sent to each child node at each internal node in the tree).

Problem 2: Analysis of Data Mining Competition Results

There have been several large organized public data mining competitions in recent years, such as the annual KDD-Cup prediction competition that is held as part of the ACM SIGKDD conference. Below are 3 papers that analyze and discuss results from 3 such competitions (links are provided on the class Web site to these papers).

1. R. Kohavi et al, 2000, KDD-Cup 2000 organizers’ report: peeling the onion.
2. R. Caruana et al, 2004, KDD-Cup 2004: results and analysis.

3. Magical thinking in data mining: lessons from CoIL challenge, Charles Elkan, 2000.

Write a short discussion and critique of each paper. In your critique you should include discussion of each of the following points:

1. Did we learn anything useful about data mining methods and algorithms directly from the results of the competition itself?
2. Does this paper provide any useful additional insights into the competition results, beyond just the results reported by the competition participants?
3. If you were running the data mining prediction competition described in each paper, would you add or change anything about the way the competition is set up, the way results are reported, and so on? suggest at least one addition or change you would make.

Problem 3: Learning to Rank for Binary Classification Problems

In class we discussed how many problems in classification involving ranking the top-scoring examples with respect to a particular class of interest, e.g., in fraud detection, credit-scoring, document retrieval, and so on. Most classification algorithms are not designed to learn good rankings. Nonetheless there has been some research on “learning to rank” using classification and regression approaches. Your goal in this assignment is to find any two research papers of your choice on the topic of learning rankings for binary classification and to write a page or two that discusses and compares these two papers.

Topics that you could include in your discussion (these are suggestions, not required) might be: what are the strengths and weaknesses of the proposed approach? how much improvement do you think the proposed method provides over using a standard technique such as ranking using naive Bayes or logistic regression? is the method scalable to large data sets?

Some links to papers on this topic are provided on the class Web site.