CMSC 828G Homework 2

Data Mining,CMSC 828G, Spring 2002

Due Date: Thursday, March 7, at the start of class

Problem 1

An important aspect of data mining is being able to handle large data sets efficiently. It is not unusual to have applications where we have data sets D with very large numbers of observations n and/or variables p. For example, in Web data analysis we might easily have data on $n = 10$ million visitors in a week's worth of data at a large Web site. In credit-scoring applications there might be $p = 500$ different variables defined on each customer, derived from their credit history and their demographic information.

Imagine that you are a data mining consultant and you are often asked to analyze data sets D that are given to you as large "flat files" with n rows (observations) and p columns (variables). Assume that the following is always known for each data set D:

- n and p are both known
- Each of the p variables is either real-valued, categorical, or discrete (ordered integers), with r, c, d variables of each type respectively, where $r + c + d = p$. The values of the categorical variables are all represented as character strings of fixed width s.
- For the d discrete variables the maximum number of possible values is 100.
- For the c categorical variables the maximum number of possible values is 10.
- You are told which of the 3 types each variable is, and you are given a brief one-sentence description of each variable.

In an ideal world you would sit down with the "data-owner" and carefully discuss each variable, how it was measured, what its interpretation is, etc. Assume, however, that p is so large (e.g., $p = 1000$) that it is not practical to do this, and/or that perhaps you don't have direct access to the data-owner for some reason. In practice this is not uncommon. Again, in an ideal world, we would have much more information about the semantics of the data variables (i.e., what each variable means) in the form of "metadata" (or "data about data"): but more commonly we only have limited information available.

For each of the following please define (a) pseudocode to determine each of the quantities and (b) derive the total time and space complexity ("big O" notation) of calculating each set of quantities. In each case you can assume that these calculations are performed sequentially, so that if you need a quantity that has been calculated earlier in the sequence, you can just assume that it is known without being recalculated.

1. The maximum, minimum, and range, for each of the r real variables
2. The median for each of the d discrete variables
3. The number of unique values and a list of these values for each of the c categorical variables
4. The sample mean for each of the r real variables
5. The sample variance for each of the r real variables
6. A matrix for each of all pairwise correlation coefficients for the r real variables (see Equation 2.5 in the Text)
7. The estimated marginal probability tables for each of the c categorical variables
8. The estimated pairwise conditional probability tables for each of the possible pairs of the c categorical variables.

Problem 2

We will shortly be discussing in class the application of classification and clustering techniques for document classification. The most widely used approach for representing documents is to convert each document to a vector of p terms, where the p terms are often automatically selected from a collection of documents, e.g., all words that appear at least three times (to eliminate rare words) but ignoring the 30% most frequently occurring words such as "the" and "of", etc. (since they contain little information content). With a reasonably large document set, this can result in a selected set of words of size $p = 10,000$ or more (so our data will be very high-dimensional!).

For each document i, $1 \leq i \leq n$ we can then count whether one of the selected words k appears or not in that document, and if it does we set $x_k(i) = 1$ and otherwise we set $x_k(i) = 0$, where $1 \leq k \leq p$, and where $x_k(i)$ is the kth component of the vector $\underline{x}(i)$ that represents document i. This is sometimes referred to as the "bag of words" representation since it ignores all document structure (e.g., word order, headings, paragraphs, etc.).

The main advantage of this is that once we have created these "term vectors" we have standardized our documents and can now treat our set of n documents as a standard $n \times p$ data matrix D. For example, we can calculate various distances between any two documents in this representation, e.g.,

$$d(i,j) = \sum_{k=1}^{p} \Big(x_k(i) - x_k(j) \Big)^2.$$

Note that these data sets are often very sparse, i.e., the average number of ones in a typical row (or document) may be very much less than p. Also note that some of the original documents might be very short to begin with (e.g., 500 words, a single paragraph) while others might be very long (e.g., 100,000 words, a long article).

Let r be the density of 1's in D, i.e., the fraction of entries that are non-zero. For the questions below assume that $r < 0.01$, i.e., less than 1 percent of all entries are 1's (this is quite typical of real-world documents).

1. Describe briefly an efficient way to store the data D, that takes much less memory than np (define your method briefly and calculate how much memory it will take).

2. Describe briefly an efficient way to calculate the distance $d(i, j)$ above, using your memory structure from part 1: define the method (e.g., via pseudocode). How much more efficient is your method on average compared to a "brute-force" calculation? (e.g., for two documents i and j that each have n_1 and n_2 1's in their rows in D, respectively, where $n_1 < rp$ and $n_2 < rp$).

3. Since the $\underline{x}(i)$'s are vectors, we can define the sample mean of these vectors and the sample variance. What is your interpretation of the sample mean? of the sample variance? (i.e., define in words what you think these vectors represent, how they can be interpreted).

4. Let's say we now have 2 sets of documents, group 1 and group 2, each defined on the same set of words (columns) p. We can calculate the mean for each group, call them μ_1 and μ_2, respectively. What is your interpretation of $\mu_1 - \mu_2$?

5. Now back to the situation where we just have one group of documents again. Define a simple procedure that automatically finds an "interesting" document, using the vector representation above (and any functions of the vectors that you wish to calculate). Feel free to be creative in terms of how "interesting" is defined (there is no unique right answer to this question!). Clearly state one advantage and one potential disadvantage of your method (e.g., how your method might be "fooled" or sensitive to the vector representation).

6. As noted earlier some documents are much lengthier than others. Say we have 3 documents, i, j, m, where i and m were both very long in their original form (e.g., over 100,000 words each) and j was very short (e.g., only 200 words). Now assume that i and j are in fact similar (i.e., about the same topic), but that m is about a different topic. Comment on the limitations of the distance measure $d(i, j)$ defined above for automatically determining the similarity of i, j, and m. Can you think of an improved distance measure $d(i, j)$ that might be used?

Problem 3

In Example 4.10 in the text, the maximizing value (the "mode") of a Beta(α, β) distribution is said to occur at a value of

$$\theta = \frac{\alpha - 1}{\alpha + \beta - 2}.$$

1. Prove that this is in fact the value of θ that maximizes the density Beta(α, β) (as defined by Equation 4.11 in the text).

2. The posterior mean (not the mode) of the posterior density is defined in the text (with a binomial likelihood and Beta prior, again see Example 4.10 and class notes) as

$$E[\theta|D] = \frac{\alpha + r}{\alpha + \beta + n}.$$

Prove that $E[\theta|D]$ must lie between the mean of the prior and $\hat{\theta}_{ML}$ (the maximum likelihood estimate for θ).

MATLAB Exercises

If you are on one of the unix machines where you can access the junkfood cluster, add the pathname `/fs/www/class/spring2002/cmsc828g/hws/hw2` to your path in MATLAB. You can do this from the MATLAB command line, or by choosing File.Set Path. If you are using a PC or MAC, you will need to copy the files from the above directory to your machine. You can access them from the class web page, `www.cs.umd.edu/class/spring2002/cmsc828g/hws/hw2`

A. Demo Scripts

Type `demoparta` at the command line in MATLAB and a script will be executed to illustrate priors, likelihoods, and posterior densities for the binomial example we discussed in class. Please go through this carefully and make sure you understand what is being plotted at each step. Also please review your class notes and Example 4.10 in the text before you start this demo script.

Next work you way through the second demo, `demopartb`, that illustrates the calculation of marginal and joint probabilities on the census data. Please make sure that you carefully go through and understand the various probability tables that are printed out to the screen.

B. Data Exploration

You will need to look at the demo scripts and functions that are called to figure out how to do the following in MATLAB:

1. Generate plots of the prior, likelihood, and posterior, for each of the following cases: (1) $r = 0, n = 4, \alpha = 2, \beta = 2$, and (1) $r = 2, n = 4, \alpha = 2, \beta = 2$. Discuss briefly and compare these graphs.

2. Experiment with the conditionals.m function and find two variables with a conditional probability table you think is interesting. Print out the output of the function for these two variables, and write a brief commentary on your interpretation of the results. The two variables selected must be different from any pair used in the demo script.

Discussion

Available from the class web page are pointers to several recent articles which involve clustering in one way or other. Choose one of them and write a one page critique of the article. Briefly summarize the main results of the paper. Discuss the novelty of the proposed approach. Discuss interesting possible extensions.

If you like, you may choose an alternate paper (as long as it is about clustering!). If you choose to do this, please turn in a copy of the paper, along with your writeup.