CS490D Spring 2004 Final Solutions, May 3, 2004
Prof. Chris Clifton

Time will be tight. If you spend more than the recommended time on any question, **go on to the next one**. If you can't answer it in the recommended time, you are either going in to too much detail or the question is material you don't know well. You can skip one or two parts and still demonstrate what I believe to be an A-level understanding of the material.

The answers I've given are one possible answer - not necessarily the only one, or even the best. *The exam is out of 35 points. My rough feeling (subject to change) is that I'd expect an A student to get at least 31, a B student to get at least 25. I'm not concerned about anyone not demonstrating at least a C level knowledge of the material, so I won't even think about the C/D cutoff.*

Note: It is okay to use abbreviations in your answers, as long as the abbreviations are unambiguous and reasonably obvious.

During the exam, you will make use of the following table, a hypothetical customer relationship management (CRM) database for a company that makes computer parts.

Name	*Publicly Held?*	*Market Cap. ($Mil)*	*Employees*	*Sales Channel*	*Customer since*	. . .	*Year*	*Units Sold*	*Profit*	. . .
Dell	Y	75000	37200	Direct	1997		2002	120000	$1.2M	
Dell	Y	88000	39100	Direct	1997		2003	109000	1.1M	
Gateway	Y	2300	11000	Direct	1995		2001	60000	.9M	
Gateway	Y	1400	11500	Direct	1995		2002	70000	1.0M	
Gateway	Y	1700	9600	Retailers	1995		2003	65000	1.9M	
Compaq	Y	10000	50000	Retailers	1993		2002	30000	1.4M	
Hewlett-Packard	Y	35000	95000	Retailers	1994		2002	80000	2.2M	
Hewlett-Packard	Y	60000	141000	Retailers	1994		2003	100000	2.5M	
MA Micro	N	?	80	Direct	1995		2003	400	3500	

. . .

This contains information on the companies we supply our products to, both general information about them (such as their market capitalization - the total value of their outstanding stock, the way they sell their products); and information about our sales to them and the profit we get from those sales. Our company's marketing department wants to use this information to better target their campaigns, thus increasing the sales of our company's products and the profit earned. The table you are shown is not complete - you can assume there are a lot more attributes both describing the company, and describing our sales to the company (represented by . . . in the table.) The information shown above will be sufficient for you to answer the questions on the exam.

1 Choosing the right data mining techniques for the job (8 minutes, 4 points)

The people in marketing would like a better understanding of their different customers. They want to know what distinguishes customers – what are the key attributes that make a customer unique? The idea isn't to group similar customers, but to identify the attributes that set customers apart.

What data mining technique you would use to answer this question? Include a sentence or two justifying your answer. Also give a couple of sentences describing how you would relate the raw data mining results back to the question asked by marketing.

Principle Component Analysis would be an appropriate technique, after dropping such "uninteresting" attributes as company name. The primary components used to generate the most important vectors would be the information to provide back to marketing.

An alternative would be to use a decision tree, regression tree, or other such "transparent" classifier to predict the company names. I'd first bin continuous values into a small number of bins, build the decision tree, then see what the top few nodes are. Also of interest would be nodes that seem to have a relatively even split between the number of entities at that point in the tree. Note that there may be multiple entries with a single company name, which helps to make this approach interesting.

Scoring: one point for a method, 1-2 points for a reasonable description of why it is appropriate, 1-2 points for how you would interpret the results.

2 Types of clustering (9 minutes, 6 points)

Give an advantage / strong point of each of the following types of clustering. (Your answers can be general - they don't need to be specific to the CRM database above.)

- K-means clustering

  Gives a "prototype" description of the cluster, the entity that would be constructed by taking the cluster mean.

- Hierarchical clustering

  Number of clusters can be based on various parameters, including intra-cluster distance, number of clusters. Gives a measure of which clusters are close to each other, which are far apart.

- Density-based clustering

  Handles odd-shaped clusters: Items can be distant from each other and still be placed in the same cluster, if appropriate.

Scoring: 1 point each for showing evidence that you know what the method is, 1 for a good advantage it has over other methods.

3 Clustering for CRM

Marketing is going to run three independent advertising campaigns, each addressing a different segment of current or potential customers. They want you to cluster the companies to help to build these advertising campaigns.

3.1 Choice of Technique (8 minutes, 3 points)

What clustering technique/algorithm would you use if your goal was to describe key characteristics of each cluster? Give a brief reason why.

I'd use k-medoid clustering, with $k = 3$. This approach handles continuous and discrete attributes (provided I define a distance function), and it would be easy for marketing to understand when I said "here is a typical company for this cluster".

Scoring: One for a valid clustering technique, 1-2 for good reasons why.

3.2 K-Means (15 minutes, 3 points)

Assume you were to cluster the companies using k-means, with $k = 3$. For just the data you are provided, describe ONE cluster, i.e., list the companies that are in the cluster and calculate the means for that cluster.

One cluster consists of MA Micro. Since it is a cluster of a single company (quite distant from any others), the mean would be that entity itself. It isn't meaningful to talk about "k-means" of categorical attributes, but for the numeric attributes the mean would be 80 employees, since 1995, year 2003, 400 items, $3500 profit.

Scoring: One for giving a reasonable set of companies for a cluster, one for showing evidence you know how to calculate the mean, one for getting a correct mean.

3.3 Is this the right question? (15 minutes, 2 points)

The goal is to maximize revenue. By using clustering of existing customers to develop the advertising campaign, the marketing department may miss something important. Describe something they might miss given the data you have and the data mining technique you suggested, and what else you would need so they wouldn't miss it.

This approach will only help us define campaigns for customers similar to those we already have. It wouldn't help to identify new market segments. I'd want to include information on potential customers that we don't currently sell to (e.g., our competitors' customers.)

Scoring: One for something reasonable that it wouldn't provide, one for what you'd do about it.

4 Data Preparation

In 2002, Compaq and Hewlett-Packard merged (i.e., Hewlett-Packard bought Compaq.) This distorts the data – you'll notice that sales to Hewlett-Packard jumped significantly in 2003, not surprisingly to roughly the combined level of the two companies in 2001. You could handle this in various ways: Combine the historical records of the two companies, try to split them out in more recent data, or just ignore the change.

4.1 Ignoring the change (8 minutes, 2 points)

Give a data mining problem/scenario/technique where it would be appropriate to just take the data as is – where the merger wouldn't make a difference.

Building a model to predict profit on a new customer from public information. Since this is independent of history, treating Compaq as an independent company would be fine.

Scoring: One for a reasonable example, one for explanation.

4.2 Merge the history (8 minutes, 2 points)

Give a data mining problem/scenario/technique where it would be appropriate to combine the companys' data prior to the merge, e.g., add the sales of Compaq to Hewlett-Packard and eliminate the old records for Compaq.

Predicting next year's profit for existing customers. Without combining past HP/Compaq data, there would be no background on the combined company to use for effective prediction.

Scoring: One for a reasonable example, one for explanation.

5 Data Mining Process / Cost

The company has said "we have the data in a data warehouse, all you need to do is install a data mining tool and run it. Shouldn't take you more than a couple of days."

5.1 True or False? (0.1 minute, 1 point)

False

5.2 Reasoning (8 minutes, 4 points)

How would you justify your answer to 5.1? Remember, you are out to convince the company - not me. Give a description of the resources you would need and the reasoning you would use estimate the time/cost required. Also suggest what resources or arguments you might use to convince the company you are correct. You don't need to do an estimate, just say briefly how you would go about it.

There are many tasks that need to be done beyond gathering the data in a data warehouse. These are a significant part of the effort, generally much greater than installing and running a tool. For documented evidence, see the CRoss-Industry Standard Process for Data Mining (CRISP-DM). Some of the tasks that need to be done are:

Business Understanding: What are the questions to be answered? Access to domain experts will be needed.

Data Analysis / preparation: The data may be in the warehouse, but is it clean? What about missing values? How about discretizing / smoothing the data so the results are meaningful? At the very least, I'd want statistics on the correctness and completeness of the data so I could estimate the time required to address these issues.

Result analysis: Relating data mining results back to the original business questions requires considerable effort, and collaboration between data mining and business domain experts.

Scoring: one point for each task, one for describing what you would need to either estimate or accomplish it. One point for giving some pointer to "expert sources" to back you up.

6 Time Series / Sequential Associates (8 minute, 4 points)

Demonstrate your knowledge of time series mining and sequential association mining by giving an example, based on the CRM database, of something that would qualify as:

- Time series mining

  Repeat buying patterns. For example, which companies can we rely on to keep us going in an economic downturn?

- Sequential associations

  Prediction of company-specific changes or trends, e.g., indicators that we may lose a customer.

Scoring: One each for demonstrating a knowledge of the difference between the two, one for a reasonable example.

7 Text Mining

Text mining commonly uses the *vector space*, or "bag of words" model. To represent a set of documents in a traditional "flat" format, each document is treated as a row. The words are the columns (attributes.) For a given document, the value of an attribute is a weight, such as the number of times that word occurs in the document.

Naïve Bayes has proven effective for text classification. However, Naïve Bayes has limitations.

7.1 Limitation Example (8 minutes, 2 points)

Give an example where Naïve Bayes would *not* be effective, but some other method would. Hint: Naïve Bayes has a general limitation, or assumption about characteristics of the data, that applies – you can describe this limitation as opposed to a specific example.

Naïve Bayes doesn't capture correlation between items. For example, the words "Naïve" and "Bayes" appearing together in a document are strongly indicative that the document is about data mining. However, either one alone is likely to be about a lot of things other than data mining.

Scoring: One for demonstrating some understanding of Naïve Bayes, one for a clear discussion of what it fails to capture.

7.2 Alternatives (8 minutes, 2 points)

Describe briefly how another classification method would overcome the limitation you described in Question 7.1.

Decision trees capture such correlation. For example, we could have “Naï ve” as one node of a decision tree. The “yes” branch could go on to “Bayes”. The yes from that would have “data mining” as the class. A no branch from either node would not lead to data mining.

Scoring: One point for naming a method that handles your objection to Naïve Bayes, one for a solid discussion of how it handles the problem.