CSci 5523 Data Mining Midterm I

October 12, 2011 4 :00 PM – 5 :15 PM 100 points / 9 questions / 9 pages

Name: Robert Martin **ID Number:** 1505151

- Please do not begin until we have handed out all the exams.
- Check to make sure you have the correct number of questions and pages.
- Use your time wisely; don't spend too much time on any question.
- Keep your answers brief and to the point.
- You do not need a calculator. If a fraction is too hard to simplify, leave it as a fraction.
- If you feel something is ambiguous, please state your assumptions in your answer.
- All electronic devices must be turned off and stowed for the remainder of the test.

90

1) Attribute types (9 points) Classify the following attributes as binary, discrete, or continuous. Further classify the attributes as qualitative (nominal or ordinal) or quantitative (interval or ratio). Some of the cases may have more than one interpretation, so briefly indicate your reasoning if you think there may be some ambiguity.

Example: Age in years. **Answer: Discrete, quantitative, ratio**

a. Degree to which people like various hobbies as measured by the fraction of their income they spend on them.

continuous, ratio

b. State of a traffic light at an intersection where the light can be red, yellow, or green.

discrete, nominal (could be ordinal in certain cases like red light running + speed)

-3 c. Increase in profit of the current year over the profit obtained in year 2000.

discrete (if measured in currency), ratio
continuous (if measured as ratio)

2) Similarity (8 Points)

a. (4 points) Which one of the following similarity measures will be least appropriate for document data? Briefly explain.

i. Correlation
ii. Cosine
iii. Jaccard

i. correlation - correlation looks at linear relationships between vectors whereas cosine + Jaccard compare the distance between the vectors themselves.

-4

b. Multiple Choice. (4 Points) If the correlation of two attributes is -0.95, which of the following statements is true.

i. They provide almost the same information.
ii. They are completely unrelated.
iii. [✓ None of these.]
iv. provide nearly the opposite

3) Data (15 points) A group of biologists conducts a field study to evaluate the relative occurrence of different types of birds in a number of states. There are many types of birds, but only a relatively small number of species in any state. The data is collected in a table where
1) the rows correspond to locations,
2) the columns correspond to different species of birds,
3) the ijth entry is the number of birds of the jth species in the ith state.

The following table indicates the representation, but is intended only for illustration.

Location/Species	Blue Jay	Crow	...	Robin	Sparrow
Alabama	0	30		50	0
Arkansas	0	15		0	50
...					
Wisconsin	20	10		0 the state bird?!	0

a. **(5 points)** To which type of data set described in Chapter 2 (Section 2.1.2) is this data set most similar?

document-term matrix

b. **(5 points)** Which proximity measure would you use if you wanted to find states that were similar in terms of the number of birds of each species? Briefly explain.

Cosine or extended Jaccard. We do not care about zero counts but do care about the magnitude of non-zero counts so regular Jaccard will not do.

c. **(5 points)** Suppose that you only care about the presence or absence of a bird species in a state. How would you change the data representation and to which type of data set in Chapter 2 would this correspond?

the columns would become binary and the document term matrix would reduce to a transaction data matrix

4) Decision Trees (**9 points**) Figure 1 shows three different decision tree classifiers built on a training sample of the same data set. The dataset consists of instances of class 'X' and instances of class 'O'. The decision boundary of the classifiers is indicated by the boundary line inside the rectangle and the classification decision is made as follows: everything above the boundary is classified as 'O' and everything below the boundary is classified as 'X'. Assume the future instances of the data are similar in terms of distribution and class composition.

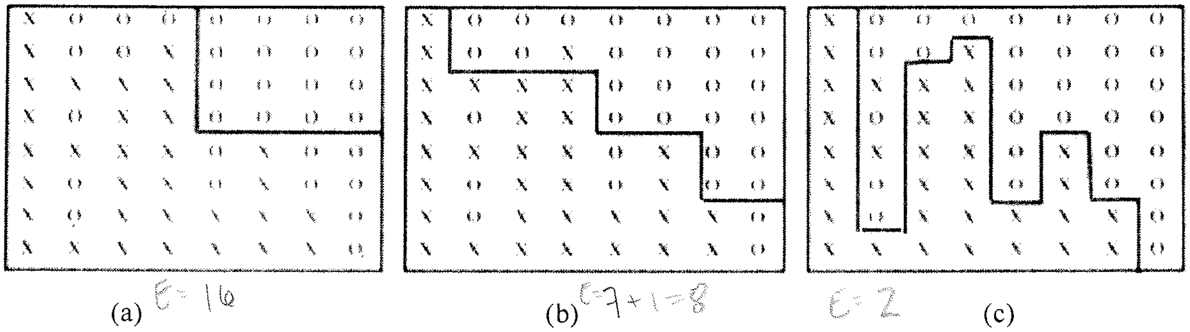

Figure 1. Three decision tree classifiers, a, b, and c. Border of the classifier is indicated by the boundary line inside the rectangle.

a) Which decision tree is the best fit to the training sample and why?

c. It has the smallest training error

b) Which model will best predict future data? Explain.

b) it makes both F1 and F2 type errors so doesn't over-generalize (1a) or over-specify (1c)

c) What are the phenomena in Figure 1a) and 1c) called?

over/under fitting

5) True / False. (8 points) Answer True or False, and provide a brief explanation.

a. **(4 points)** It is a good idea to standardize an attribute (subtract the mean and divide by the standard deviation) when the attribute is constant except for small variations due to noise.

F, standardize when you have two attributes with different means & ranges.

b. **(4 points)** Equal frequency discretization is always better than equal width.

F; if we have unbalanced classes then we may miss the correct decision boundary

7) Classifier selection (short answer) (24 points)

a. **(6 points)** If you had to choose between the naïve Bayes and k-nearest neighbor classifiers, which would you prefer for a classification problem where there are numerous missing values in the training and test data sets? Indicate your choice of classifier and briefly explain why the other one may not work so well?

NB; NB is elegant in the way it can still make decisions in the absence of data. KNN has problems with comparing two vectors w/ different missing attributes which is compounded w/ large data sets and/or large K.

b. **(6 points)** A realtor is studying housing values in the suburbs of Boston, and has given you a dataset with the following attributes: crime rate, proximity to Charles River, number of rooms per dwelling, age of unit, distance to five Boston employment centers, pupil-teacher ratio by town, house value (the target variable with values *high* and *low*). The realtor would like you to build a classification model that not only performs well, but is also easy to interpret. Between K-NN and C4.5, which one would you pick? Indicate your choice of classifier and briefly explain why the other is not a good choice. Please give at least two reasons.

C.4.5, a type of decision tree is better. First, the resulting tree is intuitive and easy to interpret. Second, there is a mix of variable types & ranges that make it difficult to measure distances between data points

-6

c. **(6 points)** Consider the problem of predicting whether a person is a good credit risk given the following attributes: hair color, income, weight, time in current job, marital status, height, age, and birth month. If you had to choose between a Bayesian classifier and decision tree, which would you prefer? Indicate your choice of classifier and briefly explain why the other one may fail?

The Bayesian would be the better choice due to irrelevant attributes like hair color & birth Month.

d. **(6 points)** Consider the problem of predicting whether a movie is popular or not, given the following attributes: Format (DVD/Online), Movie Category (Comedy/Documentaries), Release Year, Number of world-class stars, Director, Language, Expense of Production and Length. If you had to choose between RIPPER and a k-nearest neighbor classifier, which would you prefer and why? Briefly explain why the other one may not work so well?

RIPPER handles mixed data types more easily than KNN and also handles possible irrelevant data like Format & Release Year

8) Classifier selection (short answer) (12 points)

Consider the following classification methods: Decision Trees, RIPPER, Naïve Bayes, k-Nearest Neighbor (kNN).
Which of the above methods is most appropriate in the following situations? If there is more than one possible answer, choose the one that you consider most appropriate and give a brief explanation.

a. The class distribution is very skewed.

RIPPER, designed to handle unbalanced classes

b. Many attributes are correlated

decision trees, will select one of correlated attributes & ignore others resulting in less complex tree

c. Computation time for model-building is to be minimized.

KNN, does not build model

d. The data contains missing values

NB, provided there aren't too many missing values, the result of NB is still an easy & straightforward computation

9) Classification evaluation (15 points)

You are working with a doctor to evaluate how well a new, inexpensive blood test can detect a particular type of cancer. 1000 subjects are recruited from a population at high risk for the cancer and evaluated for cancer using a very expensive, but 100% accurate medical procedure. 100 subjects are found to have cancer. The 100 subjects with cancer and another 100 subjects without cancer are given the inexpensive blood test. Results are shown in the following confusion matrix.

Predicted by Blood Test

Actual	*Cancer*	*No Cancer*	
Cancer	90	10	100
No Cancer	10	90	100
	100	100	200

Cancer: precision = 90/(90 + 10) = 0.90, recall = TPR = 90/(10 + 90) = 0.90
No Cancer: precision = 90/(10 + 90) = 0.90, recall = 90/(10 + 90) = 0.9, FPR = 10/(10 + 90) = 0.10

The doctor is very excited about these results, but wants to see what the results will be after all the blood tests are evaluated. The remaining 800 subjects (none of which have cancer) are given the blood test. The confusion matrix for all 1000 subjects is given below.

Predicted by Blood Test

Actual	*Cancer*	*No Cancer*	
Cancer	90	10	100
No Cancer	90	810	900
	180	820	1000

Cancer: precision = 90/(90 + 90) = 0.50, recall = TPR = 90/(10 + 90) = 0.90
No cancer: precision = 810/(10 + 810) = 0.988, recall = 810/(90 + 810) = 0.9, FPR = 90/(90 + 810) = 0.1

a. Which of the measures (precision, recall, TPR, FPR) have changed and which have stayed the same? Comment on why some measures were affected and others were not.

Same: recall, FPR
different: precision

recall + FPR stayed the same since the same test was used adding the right proportion to the final results. Precision, or ability to to make correct predictions, changed since we only added patients w/o cancer. Or to say, recall stayed the same since the prediction stayed the same but precision changed because the test population distribution changed

recall 90%
specific 90%

b. You were disappointed by the change you observed in one of the measures from the first confusion matrix to the second, but the doctor was not. The doctor tells you to compute sensitivity (recall of the positive class) and specificity (1-FPR) for both confusion matrices. Note that cancer is the positive class. Comment on why these measures are preferred in the medical profession.

They are more concerned about missing cancer than being right. High precision means that of the cancer predicted, there was actually cancer found*. High recall means that of the patients w/ cancer the test found many of them.

— 5

c. When might precision be the preferred measure of the classification performance?

Document retrieval. If I search for documents about Naive Bayes, I don't need an exhaustive list, I just need at least one that is correct.

*a precision of 100% could mean they found one true cancer patient but this says nothing about how many they missed, possibly many.

+2

Bonus question (Please do only if you have time.)

(2 points) Suppose you measure the similarity between you and your friends based on the common friends between you. Thus, if Anna and Bob have 5 common friends, then the similarity between them is 5. Is this measure a metric? Why or why not?

Not a metric. Let's introduce Conrad he has 10 friends in common with Anna but Ø in common with Bob

$S(\text{Anna}, \text{Bob}) = 5$

$S(\text{Anna}, \text{Conrad}) = 10$

$S(\text{Bob}, \text{Conrad}) = 0$

Triangle In.

$S(\text{Anna}, \text{Conrad}) \leq S(\text{Anna}, \text{Bob}) + S(\text{Bob}, \text{Conrad})$

$10 \not\leq 5 + 0$

Does not hold triangle inequality