University of Mary Washington

CPSC419 Data Mining Spring 2014

FINAL EXAM

- This exam is due 2:30pm on Monday 28 April. You are to submit the exam to the gmail address submit.o.bot. If your exam is delivered after this deadline, it will not be accepted.
- You can submit your exam in any of the following formats:
  - plain text file
  - pdf
  - postscript

  No other formats will be accepted.
- This test is to be completed **individually** without outside help. This includes no help from peers or the Internet. You are free to use any resources used for the class including the textbook, handouts, and any class notes you made. If there is evidence that any part of the exam was completed with outside help you will receive a 0 for the entire exam.
- To discourage guessing and brain-dump style answers you will receive 20% of the XP for problems you leave completely blank. If you attempt a problem you start at zero XP. By *problem* I mean any numbered problem--subproblems do not count. This means that if you attempt one part of a problem it is best to answer the remaining parts. If you are less than 50XP away from the grade you desire for the class you can submit a blank exam and you will receive 50XP. (**Keep in mind that the minimum score you can receive for the exam is 0**)
- UMW Policy requires that you turn in a final exam, even if it is blank.

1. naive Bayes   (60 XP)

Consider the following data set of new graduates Google hired for programming jobs:

major	main language	experience w/ versioning	capstone project?	Hired
CS	Python	no	no	no
CS	Python	no	yes	no
CIS	Python	no	no	yes
Computer Engineering (CE)	Java	no	no	yes
CE	C++	yes	no	yes
CE	C++	yes	yes	no
CIS	C++	yes	yes	yes
CS	Java	no	no	no
CS	C++	yes	no	yes
CE	Java	yes	no	yes
CS	Java	yes	yes	yes
CIS	Java	no	yes	yes
CIS	Python	yes	no	yes
CE	Java	no	yes	no

We are trying to predict who is hired.

a) Construct the table of probabilities for Naive Bayes.
b) Using this table, give the equations to classify the following instances (and perform the classification):

i) CE, Python, yes, yes

ii) CS C++, no, yes

2. Decision Tree (60 XP)

Using the data in 1, Draw the decision tree.

3. kNN (50 XP)

Part 1. Describe how you could use kNN for the problem in 1 (15xp).

Part 2. Perform the classification using kNN (35xp).

4. Army uniforms (30 XP)

Recently, I read on HuffingtonPost that men's pants vary in their sizing. They compared the size of a men's 34 inch waist pair of pants. The size 34 pants from Old Navy were 39 inches, size 34 Dockers were 36 inches, and Levi's I think actually were 34. A few weeks back I was with my wife when she was shopping for clothes at Boot Barn. Women's clothing has all sorts of classifications. even sizes, odd sizes, junior, petite, and so on. It's a mess.

A few years back the U.S. Army decided to redesign women's uniforms. The Army's goal was to have better fitting uniforms and also to reduce the number of different sizes they needed in their uniform.

Researchers collected 100 different measurements on 3,000 women.

Describe how you might use data mining techniques to help the Army in this task. Be as specific as possible.

5. Cars (50XP)

I have the following data

car	MPG	HP
Nissan Altima Hybrid	35	198
Honda Civic	40	110
Lexus GS 450	22	132
Mazda MX-5 Miata	28	167
Nissan 370G	25	332
Hyundai Genesis Coupe	30	210
Ford Fiesta	37	120
Ford Fusion	36	156

Please perform a hierarchical clustering of this data. Normalize using standard scores with absolute standard deviation.

PART A:

Fill in the standard scores:

car	standardized MPG	standardized HP
Nissan Altima Hybrid		
Honda Civic		
Lexus GS 450		
Mazda MX-5 Miata		
Nissan 370G		
Hyundai Genesis Coupe		
Ford Fiesta		
Ford Fusion		

PART B:

Draw the dendrogram.