

470 Data Mining. Final exam. 2010.

Total XP: 250

HOW TO SUBMIT: Because I am old and have a hard time reading the script of the young, please submit your answers typed (PDF preferred) to submit.o.bot@gmail.com. (subject line: 470 exam submission)

1. naive Bayes (30 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 (30 XP)

Using the data in 1, Draw the decision tree using either the basic algorithm or C4.5 (in your answer specify which you used). (Note: you can do this by hand or use weka).

3. 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.

3. zWeb (30 XP)

I have a news aggregation app for the iPad (and soon for the Android). I just started using Facebook's useful (but creepy) feature that gives me the public information of Facebook users, if they use my app while still being logged into Facebook on their browser. Minimally, this lets me uniquely identify users. With this information I maintain a server log, containing page view information (what pages/articles they looked at and for how long). Articles are also classified by a tag (for example, a page might be classified with the two tags *education* and *philanthropy*).

- a) How can I use this information to improve my app?
- b) What specific algorithm should I use
- c) Do I need to clean or normalize the data in any way?

4. 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.

5. Visualization (30XP)

- Identify at least 2 advantages and 2 disadvantages of using color to visually represent information.
- What are the arrangement issues that arise with respect to 3-dimensional plots?
- Describe how you would create a visualization to display information that describes the change in occupation of workers in countries around the world over the last thirty years. Assume you have a variety of yearly information about each person including gender and level of education.

6. Lightning Round (50XP)

Give short 1-2 sentence answers for the following:

- What are the disadvantages of hierarchical clustering?

- b) Give one example of when we shouldn't normalize data.
- c) You've developed a nearest-neighbor recommendation system using the basic algorithm. Unfortunately, now that you have lots of users the algorithm is very slow. What might be a way to speed this up?
- d) What is the difference between k-means clustering and hierarchical clustering?
- e) What are the attributes (and values) when data mining unstructured text?