

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?