Data Mining

Homework 1

Due: 22/3/2015, 23:59.

Instructions

You must hand in the homeworks electronically and before the due date and time.

Handing in: You must hand in the homeworks by the due date and time by an email to the instructor that will contain as attachment (not links!) a .zip or .tar.gz file with all your answers and subject

`[Data Mining class] Homework #`

where `#` is the homework number. After you submit, you will receive an acknowledgement email that your homework has been received and at what date and time. If you have not received an acknowledgement email within 1 day after you submit then contact the instructor.

The solutions for the theoretical exercises must contain your answers either typed up or hand written clearly and scanned.

The solutions for the programming assignments must contain the source code, instructions to run it, and the output generated (to the screen or to files).

For information about collaboration, and about being late check the web page.

Most of the questions are not very hard but require time and thought. **You are advised to start as early as possible, to work in groups, and to ask the instructor in case of questions.**

Problem 1. A family has two kids, each being a boy or a girl with probability 1/2 and born in a random day of the week.

1. Define a sample space sufficient to answer the third question.
2. If we know that one kid is a girl, what is the probablity that the other kid is a girl?
3. If we know that one kid is a girl born on Sunday, what is the probablity that the other kid is a girl?

Problem 2. You are in an airplane that falls in the jungle and you manage to survive. In the jungle there are are two tribes, the *Randomukee* and the *Bugiardukee*. The Randomukee are twice as many as the Bugiardukee. Each time you ask a question to a Randomukee he will say the truth with probability 3/4, whereas, each time you as a question to a Bugiardukee he will lie. As you try to find your way out of the jungle, you find a random person from the two tribes. You ask him the question "To get out of the jungle, I have to go left or right?"

1. Define an appropriate probability space that can be used to answer the questions that follow.
2. Assume that the person gave you the answer "right." What is the probability that the answer is correct?
3. You ask the same person again, and he gives you the same answer. Show that the probability that the answer is correct is 1/2.

4. You ask a third time and you get again the same answer. What is now the probability that the answer is correct?

5. Finally, you ask a fourth time and you get again the answer "right." Show that the probability that the answer is correct is 27/70.

6. Assume that the first three times the answer was "right" but that the fourth one it was "left." Show that the probability that the correct answer is "right" is 9/10.

Hint: You need to define several events to answer the questions.

Problem 3. Assume that a monkey sits in front of a keyboard and hits randomly the 26 letters, each with the same probability. Assume that it types 100,000,000,000 letters. Let X be the number of times that the word "mining" appears? What is the expectation of X?

Problem 4. Quite often you can analyze your data just by using simple unix tools. Some usefull commands are the `grep`, `sort`, `uniq`, `cut`, `sed`, `awk`, `join`, `head`, `tail`, `wget`, `curl`. You can find more information using the `man` command or by checking the web. Shell scripting can help you even more.

As a simple exercise do a simple analysis of the reviews in
`http://aris.me/contents/teaching/data-mining-2015/protected/ratebeerProcessed.txt`.
After you download and unzip the file, use some of the commands above to find the 10 beers with the highest number of reviews. (**Hint:** You can do it with a single command line, by chaining commands through pipes!)

Problem 5. We will now go one step further and start practicing with Python. Write a Python program to find the top-10 beers with the highest average overall score among the beers that have had at least 100 reviews. (You may need to preprocess the file first.)

Problem 6. We will run locality-sensitive hashing (LSH) on a set of announcements that we will obtain from the kijiji web site. In this exercise we do the first step, which is downloading and parsing the web pages with the announcements.

For downloading the web pages you may use the package `Requests` or the package `urllib2`. To parse the page you can either use regular expressions through the package `re` (it is anyway a good idea become familiar with regular expressions), or, probably better, use an HTML/XML parser. The `Beautiful Soup` package is a good one but it loads the whole file in memory. This is fine for this problem, since the pages to parse are small, but be careful if you want to use it on large XML files; for those ones check the `lxml` library and the tutorial at

`http://www.ibm.com/developerworks/xml/library/x-hiperfparse/`

Write a program that will download from `http://www.kijiji.it` and parse all the job positions in Lazio about *Informatica/Grafica/Web*. Download regular and top announcements, but not sponsored ads. Save in a tab-separated value (TSV) file, for every job (one line per job), the *title*, *short description* (from the job summary page), *the location*, the *publication date* of the job announcement, and the *URL link* to its web page. **Because you will make a lot of calls to the kijiji site, make sure that you have a delay (use: `sys.sleep()`) between different downloads of kijiji pages, to avoid being blocked.**

Optional, extra credit: Download and store also the *full text* of the ads. (You will need to visit the ad pages for that.)