Data Mining

Homework 1

Due: 3/11/2024, 23:59

Instructions

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

The first homework has to be done by each **person individually**.

Handing in: You must hand in the homeworks by the due date and time by an email to Gianluca (`decarlo@diag.uniroma1.it`) that will contain as attachment **(not links to some file-uploading server!)** a .zip file with your answers. The filename of the attachment should be `DM_Homework_1__StudentID_StudentName_StudentLastname.zip`;
for example:
`DM_Homework_1__1235711_Robert_Anthony_De_Niro.zip`.
The email subject should be
`[Data Mining] Homework_1 StudentID StudentName StudentLastname`;
For example:
`[Data Mining] Homework_1 1235711 Robert Anthony De Niro`.
After you submit, you will receive an acknowledgement email that your project has been received and at what date and time. If you have not received an acknowledgement email within 2 days after the deadline then contact Gianluca.

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

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

Problem 1. We shuffle a standard deck of cards, obtaining a permutation that is uniform over all 52! possible permutations.

1. Define a proper probability space Ω for the above random process. What is the probability of each element in Ω?
2. Find the probability of the following events:
   (a) The first four cards include at least one club.
   (b) The first seven cards include exactly one club.
   (c) The first three cards are all of the same suit (i.e., they are clubs, or all spades, or all hearts, or all diamonds).
   (d) The first three cards are all sevens.
   (e) The first five cards form a *straight*, (e.g., 8, 9, 10, J, Q independently of the suit—but not all suits can be the same, as this would be a *straight flush*).
3. (Optional) Develop some small programs in Python to perform simulations to check your answers.

Problem 2. 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 and define the probabilities to the points in the sample space.

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 3. The following "paradox" is an example of how some concepts can be counter intuitive when dealing with probabilities, even in our everyday life.

Data-miningitis is a very rare disease that each person has probability 1 in a 100,000 to have it. Luckily, there exists a very good test for the disease, which has accuracy 99.9%.

Your instructor, who is afraid of having it, went to take the test and he was found positive. Given the result of the test, what is the probability that he indeed suffers from data-miningitis? First define an appropriate sample space, then the events required so that you can compute the probability.

Problem 4. Aris and Gianluca each pick a different page from the textbook IIR. Your goal is to find who chose the page that appears earlier in the book. Note that we do not choose the pages randomly; in fact, we may choose the pages adversarially so that we make your task really hard (for example, by choosing pages 1 and 2). However, to help you we give you the chance to ask one of us randomly (each with probability 1/2) what is the page number he has chosen. Note that it is easy to find a strategy with probability of winning exactly 1/2; for example, pick Aris or Andrea with probability 1/2. Surprisingly, there are ways to achieve a probability of winning strictly greater then 1/2. Devise a strategy that does that.

Hint: Assume that you know that a given number x is between the two page numbers. Then you know that if the random person you picked reveals a number $y > x$ then you need to respond that the other person has picked the smallest page number. Of course you do not really know x but this hint should help you.

Problem 5. The Erdős-Rényi $G_{n,p}$ random-graph model, is a mathematical model for creating random graphs. Fix a positive integer n and a value $p \in [0,1]$. Then a graph created according to the $G_{n,p}$ model, has n nodes, and each pair of distinct nodes is connected with an edge with probability p; all pairs being independent from each other.

1. Define an appropriate probability space Ω to describe the $G_{n,p}$ model. What is its size?

2. What is the probability of each element of Ω?

3. What is the probability that a graph created according to $G_{n,p}$ contains exactly two node-disjoint cycles of length $n/2$ and no other edges (assume that n is even for this)? A cycle is a sequence of vertices $v_1, v_2, \ldots, v_\ell, v_1$ ($\ell \geq 3$), with $v_i \neq v_j$ for $i,j \in \{1, \ldots, \ell\}$, such that each edge (v_i, v_{i+1}) exists in the graph (as well as (v_ℓ, v_1)).

4. What is the probability that a graph created according to $G_{n,p}$ contains exactly two node-disjoint cycles of any length (the two cycles can have different length and of course they have to have length at least 3) with no nodes in common, no isolated nodes, and no other edges?

5. In a graph created by the $G_{n,p}$ model, what is the expected degree of a node?

6. In a graph created by the $G_{n,p}$ model, what is the expected number of edges?

7. We define a *papillon subgraph* to be a subgraph of 5 nodes, $v_1, \ldots, v_5$ in which the edges $\{v_1,v_2\}, \{v_1,v_3\}, \{v_2,v_3\}, \{v_1,v_4\}, \{v_1,v_5\}, \{v_4,v_5\}$ exist, and no other edge between them exists (see Figure 1). What is the expected number of house subgraphs in a random graph according to $G_{n,p}$?

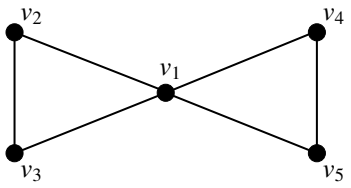

Figure 1: The *papillon subgraph*.

Problem 6. 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-2024/protected/beers.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 7. 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.)