Data Mining

Homework 1

Due: 15/10/2023, 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 you submit 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 two cards include at least one ace.
   (b) The first five cards include at least one ace.
   (c) The first two cards are a pair of the same rank (they are the same number or both are J, or both are Q, etc.)
   (d) The first five cards are all diamonds.
   (e) The first five cards form a full house (three of one rank and two of another rank).
3. (Optional) Develop some small programs in Python to perform simulations to check your answers.

Problem 2. In a hospital, at 11pm, there are 4 newborn boys and some number of girls. At midnight one new baby was born. Next morning a data miner, picked a random child from the hospital (including the newborn) and he found it to be a boy.

1. Design an appropriate probability space for the above process.
2. What is the probability that the baby born at midnight was a boy?

Problem 3. You throw a set of 3 regular dice again and again, until for the first time you see a sum of 11 or a sum of 16.

1. Design an appropriate probability space for the above process.
2. What is the probability that you stop because you see a sum of 16?

Problem 4. 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 5. If the probability to see a bicycle passing by a given spot of Via Nazionale in 45min is 97%, what is the probability to see it in 15min? Assume a constant probability over time that a bicycle passes.

Problem 6. 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 only one triangle (three nodes connected with each other and no other edges exist)?
4. What is the probability that all the n nodes are connected in a line (with no other edges present)?
5. Assume that we create a graph according to $G_{n,p}$, what is its expected number of edges?
6. Consider a random graph $G = (V,E)$ with $|V| = n$ nodes created according to the $G_{n,p}$ model. Let us define a *3-star* to be a subgraph $V' = \{v_0,v_1,v_2,v_3\} \subseteq V$ of V such that all the edges (v_0,v_i) for $i = 1,2,3$ exist in G, and none other of the edges (v_1,v_2), (v_1,v_3), and (v_2,v_3) exist in the graph (but other edges may exist, including edges between nodes in V' and other nodes in V). What is the expected number of 3-stars in G?
7. Define similarly a k-star. What is the expected number of k-stars?

Problem 7. The goal of this problem is to dreate a Python program to download weather forecasts for a specific city from a website and extract relevant data for further analysis.
Description:

- Go to the *Meteomatics* website and get a free API account (link: `https://www.meteomatics.com/en/sign-up-weather-api-free-basic-account/`).

- Use the username and password provided by them to access their API service.
- Look at their documentation (link: `https://www.meteomatics.com/en/api/getting-started/`) and use *requests* Python module to download the weather forecast page for a location of your choice.
- Choose your favourite return format (CSV, JSON, HTML, ...) and parse accordingly the data returned by the API.
- Gather information about at least three parameters (here the list of the available parameters: `https://www.meteomatics.com/en/api/available-parameters/`) and plot their trend over 24 hours. To plot your data try avoiding using the standard *Matplotlib.pyplot* library but instead find and learn how to use a fancier one, such as *Plotly* (`https://plotly.com/`), or some other that you prefer.

Requirements: The program should be able to handle any errors during web page downloads or data analysis.