Homework 1 Data Mining

Davide Fortunato 1936575

November 2024

1 Problem 1

1.1 Answer to 1.1

Since the experiment consist in shuffling a standard deck of 52 cards, a proper sample space is represented by the set containing all possible permutations of the deck, meaning that it will contain all the 52! possible orderings of the cards. To formalize: $\Omega = \{\omega_1, \omega_2, ..., \omega_{52!}\}$ where each ω_i is a permutation of the deck. Each element ω_i in the sample space has the same probability to occur, which is $P(\omega_i) = \frac{1}{52!}$.

1.2 Answer to 1.2

1.2.1 a)

The probability that the first four cards include at least one club can be expressed using the complement rule. Specifically, we have:

$P(\text{at least one club in the first four cards}) = 1 - P(\text{no clubs in the first four cards}),$

where the complementary event of "at least one club" is "no clubs at all". So, the probability that there are no clubs in the first four cards is

$$P(\text{no clubs in the first four cards}) = \frac{39}{52} \times \frac{38}{51} \times \frac{37}{50} \times \frac{36}{49} \approx 0.31$$

so, finally, we can claim that

$$P(\text{at least one club in the first four cards}) = 1 - 0.31 = 0.69$$

1.2.2 b)

The probability that the first seven cards include exactly one club is:

$$P(\text{exactly one club}) = 7 \times \frac{13}{52} \times \frac{39}{51} \times \frac{38}{50} \times \frac{37}{49} \times \frac{36}{48} \times \frac{35}{47} \times \frac{34}{46} \approx 0.317$$

I adopted the following methodology to compute the result: at first, you can choose any card among the 13 clubs (13/52). Then each time you need to pick

non-club cards and you have 52-13 = 39 options for the first non-club choice, then 38, 37 till 34. Everything is multiplied by 7 because the club card can be at every position of the 7 available.

1.2.3 c)

$$P(\text{first three cards are all of the same suit}) = 4 \times \frac{13}{52} \times \frac{12}{51} \times \frac{11}{50} = 0.0517$$

Basically, since each suit has 13 cards, we can choose 13 out of 52 cards for the first choice, then 12 out of 51 and 11 out of 50. Everything must be multiplied by 4 since cards can be of any of the 4 suits.

1.2.4 d)

$$P(\text{the first three cards are all seven}) = \frac{4}{52} \times \frac{3}{51} \times \frac{2}{50} = 0.00018$$

Given that there are four sevens in a standard deck of 52 cards, the probability of drawing a seven on the first card is $\frac{4}{52}$. If a seven is drawn, the probability of drawing another seven on the second card becomes $\frac{3}{51}$, and if a second seven is drawn, the probability of drawing a third seven on the third card is $\frac{2}{50}$.

1.2.5 e)

I analyzed both the case where the order in which we pick cards matter and the case where it doesn't (for instance, if our first five cards are {5,3,1,2,4} in the case where order of pick matters this isn't a straight, while in the second it is since after ordering we have {1,2,3,4,5}). For the first case, I computed the probability in this way:

$$P(\text{straight with order}) = \frac{4}{51} \times \frac{4}{50} \times \frac{4}{49} \times \frac{3}{48} = 3.2 \cdot 10^{-5}$$

The first pick can be any card so I didn't included the first pick in the calculation. Instead the second pick must be the consecutive of the first and there are 4 of them out of the 51 cards (52 - 1 the one already picked). The same applies for the third pick and then for the last card there are again 4 options but we must exclude the one with the same suit of all the other cards (in the worst case, all first 4 cards will have same suit, so the last one must have a different suit). For the case where the order of pick doesn't matter, we have

$$P(\text{straight without order}) = \frac{10 \times (4^5 - 4)}{\binom{52}{5}} = 0.0039$$

where the numerator is 10 (we have 10 possible straights) multiplied by all ways the five cards can be arranged in terms of suits minus the only 4 cases where they have the same suit.

1.3 Answer to 1.3

See the code provided.

2 Problem 2

2.1 Answer to 2.1

In order to answer the question 2.3, a proper sample space is defined as follows:

$$\Omega = \{G, B\} \times \{Mon, Tue, Wed, Thu, Fri, Sat, Sun\} \times \{G, B\} \times \{Mon, Tue, Wed, Thu, Fri, Sat, Sun\}$$

So, each point of the sample space will encode the sex of the first child together with her/his day of birth, and the sex of second child together her/his day of birth. For instance, a sample point can be (G, Mon, B, Wed) encoding that the first child is a girl born on a Monday and the second child is a boy born on a Wednesday. Notice that $|\Omega| = 196$ since, given that for each child we have 14 possible outcomes, we get 14×14 possible combinations for both the children. Each sample space point will have the same probability that is $P(\omega_i) = \frac{1}{196}$.

2.2 Answer to 2.2

P(both girls|one kid is a girl) $= \frac{1}{3}$ since, given that one kid is a girl, we restrict our possible outcomes to be {GG, BG, GB} removing the possibility of having BB, encoding the fcat that both children are boys (notice that here the sample space is different from the one defined in the previous point). Among these possible outcomes, the only favorable case is GG, by which I derived $\frac{1}{3}$.

2.3 Answer to 2.3

To determine the possible outcomes given that at least one child is a girl born on a Sunday, we start by identifying the constraints imposed by this precondition. Specifically, we can only consider tuples of the form (G, Sun, _, _) or (_, _, G, Sun), where _, _ represent the other child's gender and day of birth. For these empty spots, there are 14 possible outcomes: {(G, Mon), (G, Tue), ..., (G, Sun), (B, Mon), (B, Tue), ..., (B, Sun)}. Since we have two couples of empty slots to fill, we'd initially consider 14 + 14 = 28 outcomes. However, we must subtract 1 to account for the overlap in the tuple (G, Sun, G, Sun), which was counted twice. This gives us a total of 27 unique possible outcomes. At this point, we must count all the favorable cases for the probability we are interested in, that is P(both kids are girls|one kid is a girl born on a Sunday). Since for each empty spot we have 7 possible choices to fulfill the fact that the other kid is a girl ({(G, Mon), (G, Tue), ..., (G, Sun)} we have 7 + 7 -1 = 13 favorable outcomes where also here the -1 refers to the overlap of (G, Sun, G, Sun) tuple. So we can conclude that P(both kids are girls|one kid is a girl born on a Sunday) $= \frac{13}{27}$

3 Problem 3

I defined the following sample space $\Omega = \{TP, TN, DM, NDM\}$ where TP and TN encode that the fact that the test was positive and negative respectively, while DM and NDM mean that the patient has the Data-Miningitis and the patient doesn't have the disease respectively. At this point, we are interested in computing $P(DM|TP) = \frac{P(DM \cap TP)}{P(TP)} = \frac{P(TP|DM)P(DM)}{P(TP)}$. We know that $P(DM) = \frac{1}{100000}$ and that the test accuracy is 99.9%, which means that $P(TP|DM) = 0.999$ and $P(TP|\neg DM) = 0.001$. We can compute $P(TP) = P(TP|DM)P(DM)+P(TP|\neg DM)P(\neg DM) = 0.999 \times 0.00001 + 0.001 \times 0.99999 = 0.00101$. Finally,

$$P(DM|TP) = \frac{P(DM \cap TP)}{P(TP)} =$$

$$\frac{P(TP|DM)P(DM)}{P(TP)} = \frac{0.999 \times 0.00001}{0.00101} = 0.00989 \approx 1\%$$

4 Problem 4

Given a probability distribution D that is positive everywhere, the strategy for having probability of winning strictly greater that $\frac{1}{2}$ begins by choosing a random number n according to the distribution D and adding it a decimal component. At this point, we have three possible cases: supposing that x and y are the number chosen by Aris and Gianluca, n can be less than both x and y, or n can be greater than both x and y, or n can be in the middle of the two numbers. The latter case is the most interesting, because we should notice that in this case if we follow the strategy suggested in the question we are gonna guess correctly who has the greater number whatever is the person to which we ask to reveal his number: in fact, if n is such that $x \leq n \leq y$ and we are revealed the number x that is less than n, we can conclude that y is the greater number, or if we are revealed the number y that is greater than n, we can conclude that y is the greater so we are gonna guess correctly in all the cases, no matter who we choose to reveal a page number. In the other two cases, where n is less or greater than both numbers, our chance to guess correctly drops down to 50%. For example, if numbers of Aris and Gianluca are 4 and 10 and our random number is 15.3 and we choose Gianluca to reveal his number, 10, according to the strategy, since 10 is less than 15.3 we would conclude that the number of Aris is the biggest among the two, making a mistake. The same example can be made with both numbers of Aris and Gianluca are bigger than the random number n. Finally, I claim that if S is the event success and refers to the fact of guessing correctly the winner, LT, M and GT are the events that the random number is generated such that is less than, in the middle of and greater than both numbers of Aris and Gianluca

$$P(S) = P(S|LT)P(LT) + P(S|M)P(M) + P(S|GT)P(GT)$$

$$= \frac{1}{2}P(LT) + 1P(M) + \frac{1}{2}P(GT)$$

$$= \frac{1}{2}P(LT) + \frac{1}{2}P(M) + \frac{1}{2}P(M) + \frac{1}{2}P(GT)$$

$$= \frac{1}{2}(P(LT) + P(M) + P(GT)) + \frac{1}{2}P(M) = \frac{1}{2} \cdot 1 + \frac{1}{2} \cdot P(M)$$

But since the random number is generated according to a probability distribution that is positive everywhere, $P(M) > 0$ so we conclude that by following this strategy we have a probability that is strictly greater than 0.5, although being aware that if Aris and Gianluca choose two numbers that are close each other the probability that the random number falls between those two numbers is less than if the two numbers where far each other and so in this case is not a big advantage.

5 Problem 5

5.1 Answer to 5.1

To describe the $G_{n,p}$ model an appropriate probability space Ω can be defined as the set of containing all possible graphs with n nodes. $|\Omega| = 2^{\binom{n}{2}}$

5.2 Answer to 5.2

Each element of Ω is a specific graph with a certain number of edges, suppose this number is m. The probability of each graph G having m edges will be

$$P(G) = p^m(1-p)^{\binom{n}{2}-m}$$

where p is the probability that an edge is generated and it is a parameter of the $G_{n,p}$ model.

5.3 Answer to 5.3

Supposing that G is a graph containing two node-disjoint cycles each of length $\frac{n}{2}$,

$$P(G) = \frac{1}{2}\binom{n}{\frac{n}{2}} \frac{(\frac{n}{2}-1)!}{2} \frac{(\frac{n}{2}-1)!}{2} p^n(1-p)^{\binom{n}{2}-n}$$

where $\frac{1}{2}\binom{n}{\frac{n}{2}}$ consists of all the possible ways to partition the nodes in two groups of size $\frac{n}{2}$ and $\frac{(\frac{n}{2}-1)!}{2}$ is the number of all unique cycles (where with unique I mean all cycles without counting symmetric ones and rotated - the cycle ABC will not appear in the form of CBA or BCA) and this factor is repeated twice since we have two groups of size $\frac{n}{2}$. So here I computed the probability of having a graph with exactly n edges and multiplying it by the number of possible ways we can achieve the specific configuration required.

5.4 Answer to 5.4

$$\sum_{k=3}^{\frac{n}{2}-1} \binom{n}{k} \frac{(k-1)!}{2} \frac{(n-k-1)!}{2} p^n (1-p)^{\binom{n}{2}-n} + \frac{1}{2} \binom{n}{\frac{n}{2}} \frac{(\frac{n}{2}-1)!}{2} \frac{(\frac{n}{2}-1)!}{2} p^n (1-p)^{\binom{n}{2}-n}$$

My idea here was that we need to sum all possible ways to partition the n nodes into two groups, this time without the constraint that the two groups should be of length $\frac{n}{2}$. At first I thought that the sum should arrive till n-3 but I prefer to impose the limit of $\frac{n}{2} - 1$ because otherwise we consider partitions already counted: for instance, if we have 8 nodes, and k= 3, I think it is a mistake to reconsider the ways of partitioning nodes in groups with k=5 since the groups of 5 were already counted in the part $\frac{(n-k-1)!}{2}$ when k was 3. In the end, I kept the case where $k = \frac{n}{2}$ separated because when groups are indistinguishable the first factor must be divided by 2.

5.5 Answer to 5.5

I imagined the random variable D, degree of a node, to follow a Binomial distribution with parameters n-1 (a node can be linked with at most all the others n-1 nodes) and probability p, the same of the $G_{n,p}$ model

$$P(D = i) = \binom{n-1}{i} p^i (1-p)^{n-1-i}$$

$$E(D) = \sum_{i}^{n-1} iP(D = i)$$

It is well known that the expected value of a binomial distributed variable is the product of its two parameters, so

$$E(D) = (n-1)p$$

6 Answer to 5.6

I applied same reasoning as before, which is trying to imagine the random variable as a binomial distributed one and in this case the parameters will be $\binom{n}{2}$, since each possible edge can be seen as a Bernoulli trial, and p is the same of the $G_{n,p}$ model. So in this case, if N is the random variable encoding the number of edges

$$E(\mathrm{N}) = \binom{n}{2} p$$

6.1 Answer to 5.7

In this case, the number of Bernoulli trial will be $\binom{n}{5}$ which is all the possible ways of grouping 5 nodes for the papillon subgraph and the probability is $p^6(1-$

$p)^4$ which is the probability that we have exactly 6 edges, those to form the papillon subgraph and no other edge. So, consider PS the random variable for the number of papillon subgraph

$$E(\text{PS}) = \binom{n}{5} p^6 (1-p)4$$

7 Problem 6

I used the following single line command

```
cut -f1 beers.txt | sort | uniq -c | sort -nr | head -10
```

and got that the top-10 beers with the highest number of reviews is

```
3696 Guinness Draught
3662 Pabst Blue Ribbon
3230 Dogfish Head 90 Minute Imperial IPA
3126 Budweiser
3119 Sierra Nevada Pale Ale &#40;Bottle&#41;
3110 Samuel Adams Boston Lager
3056 Chimay Bleue &#40;Blue&#41; / Grande Rserve
2904 North Coast Old Rasputin Russian Imperial Stout
2872 Stone Arrogant Bastard Ale
2813 Orval
```

8 Problem 7

See the code provided.