Data Mining Homework 1

Fabio Fraschetti 1834942

October 2023

[Fabio Fraschetti]

1 Exercise

1.1

Ω = all 52! possible outcomes that can occur when shuffling the deck.
The probability of each element is $\frac{1}{52!}$, which ensures that the sum of the probabilities of all possible outcomes is equals to 1.

1.2

- **a)** The probability of the event "The first two cards include at least one ace" is the probability of finding an ace at the first extraction plus the probability of not finding an ace in the first extraction and find an ace in the second extraction:
  $P = \frac{4}{52} + \frac{48}{52}\frac{4}{51} = \frac{4}{52} + \frac{48}{663} = 0.076 + 0.072 = 0.149$
- **b)** The probability of the event "The first five cards include at least one ace" is calculated as before but iterated 5 times:

  $P = \frac{4}{52} + \frac{48}{52}\frac{4}{51} + \frac{48}{52}\frac{47}{51}\frac{4}{50} + \frac{48}{52}\frac{47}{51}\frac{46}{50}\frac{4}{49} + \frac{48}{52}\frac{47}{51}\frac{46}{50}\frac{45}{49}\frac{4}{48} = 0.341$
- **c)** The probability of the event "The first two cards are a pair of the same rank". Once the first card is extracted the probability that the second card as the same value of the first one is:
  $P = \frac{3}{51} = 0.0588$ because only 3 cards with the same rank of the first card are left and the total number of cards now are 52-1
- **d)** The probability of the event "The first five cards are all diamonds":
  $P = \frac{13}{52}\frac{12}{51}\frac{11}{50}\frac{10}{49}\frac{9}{48} = 0.00049$ because the probability of choosing the first diamond card is $\frac{13}{52}$, the probability of choosing the second diamond card is $\frac{12}{51}$ and so on until the 5th diamond card that has probability $\frac{9}{48}$
- **e)** The probability of the event "The first five cards form a full house":
  $P = \binom{5}{3}(\frac{52}{52}\frac{3}{51}\frac{2}{50}\frac{48}{49}\frac{3}{48}) = 0.00144$
  Because the probability of choosing the first rank card is $\frac{52}{52}$, the probability of choosing the second card with the same rank is $\frac{3}{51}$ and the probability of choosing the third card with the same rank is $\frac{2}{50}$. Then we can choose the 4th card (with another rank) from a deck of 49 cards with probability $\frac{48}{49}$ and the probability of choosing the 5th card with the same rank of the 4th card is $\frac{3}{48}$. Then we need to calculate all the combinations that can happen of this sequence with $\binom{5}{3}$

1.3

I write a python code, it's in the folder and its name is: HW1FraschettiEX1.py

2 Exercise

2.1

The sample space is $\Omega = \{B_i \mid i \in \{1, 2, \ldots, n+4+1\}\}$ where B_i is the boy or girl in the i_{th} position. I placed $n+4+1$ because n is the number of girls 4 is the number of boys and 1 is the newborn child. With this sample space we got 2^{n+4+1} dispositions. Now we

define the event that is a subset in which we take all the dispositions where there are 4 boys in (n+1)+4 children or 5 boys in n+5 children. Now we can calculate the probability of take a boy conditioning that the baby born at midnight was a girl as: $P(B \mid N_F) = \frac{4}{n+4+1}$. Then the probability to take a boy conditioning that the baby born at midnight was a boy: $P(B \mid N_B) = \frac{5}{n+4+1}$. So the probability that a data miner take a boy is the sum of the above probabilities times $\frac{1}{2}$ (the probability that born a boy or a girl). The final probability of the event is: $P(B) = \frac{1}{2}\frac{5}{n+4+1} + \frac{1}{2}\frac{4}{n+4+1} = \frac{9}{2n+10}$

2.2

The probability that the baby born at midnight was a boy conditioning the data miner pick up a boy is: $P(N_B \mid B) = \frac{P(N_B)\cdot P(B|N_B)}{P(B)} = \frac{\frac{1}{2}\frac{5}{n+5}}{\frac{9}{2n+10}} = \frac{5}{9}$ here I use bayes formula, I apply all the probabilities in the exercise above. $P(N_B)$ is the probability that the new born is a boy $\frac{1}{2}$.

3 Exercise

3.1

I define a set $\Omega_i = (a_1, b_1, c_1 \ldots, a_i, b_i, c_i)$ *s.t.* $a, b, c \in \{1, 2, 3, 4, 5, 6\}$ that is the set that contain all the possible results of throwing three dice i times. Then we define the sample space: $\Omega = \cup_{i=1}^{\infty}\Omega_i$. Then we define a generic event $\omega_n = (a_1, b_1, c_1 \ldots, a_n, b_n, c_n)$ and

$$P(\omega_n) = \begin{cases} \frac{1}{6^{3n}} & \forall l < n : (a_l + b_l + c_l \neq 11 \textit{ and } 16) \textit{ and } (a_n + b_n + c_n = 11 \textit{ or } 16) \\ 0 & \textit{otherwise} \end{cases}$$

The probability of this event is $\frac{1}{6^{3n}}$ because the probability of each throw is $\frac{1}{6^3}$ times the number of throws n and the probability is set to 0 otherwise because we are not interested in that event, so it can't happen.

3.2

The probability that we stop because we got a 16 is:

$$P(16) = \sum_{i=0}^{\infty}(P_n)^i \cdot P_{16} = \frac{P_{16}}{1 - P_n} = \frac{0.027}{0.149} = 0.181$$

where P_n is the probability that neither 11 nor 16 appear in a single triple, that is $\frac{183}{216}$. P_{16} is the probability that 16 appear in a single triple and is equal to $\frac{6}{216}$. I can do the first step of the equation because $\sum_{i=0}^{\infty}(P_n)^i \cdot P_{16}$ is in the form of a geometric series where we know that the result is $\frac{1}{1+P_n}$.

4 Exercise

I approximate the number of presence of the word "mining" to a binomial distribution

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

where k is the number of presence of the word, n is the total number of possible trials that is 99.999.999.995(is $100.000.000.000 - 5$ because i can't put the word mining in the last 5 position of the string) and p is the probability that the word appear that is $\frac{1}{26^6}$. Now I can use the expected value of the binomial as:

$$E[X] = n \cdot p = 99.999.999.995 \cdot \frac{1}{26^6} = 323,712$$

5 Exercise

I use as time unit the quarter, and I consider the probability of not see a bicycle in a quarter so ($q = 1 - p$). Then the probability of not see bicycle in 45 minutes is $q^3 = 0.03$ I can find $q = 0.3107$ that is the probability of not see a bicycle in a quarter. So the probability of see a bicycle in 15 minutes is $1 - q = 0.6893$

6 Exercise

6.1

The sample space is $\Omega = \{m_{ij}\ \forall i \in \{1, 2, \ldots, n-1\},\ \forall j \in \{2, \ldots, n\}, i < j\}$ where mij is a graph that contain the edge between the nodes i and j. The size Ω is $\binom{n}{2}$. The event space is the set of the subset of Ω this set has dimension $2^{\binom{n}{2}}$. If I have set m=$(m_{12}..., m_{ij})$ I consider this set of graphs as a unique graph with n edges where n is the length of m.

6.2

The probability of each element in the event space E_k is :

$$P(E_k) = p^k \cdot (1-p)^{\binom{n}{2}-k}$$

this probability depend from parameter k that is the size of the event. I obtained it because the probability of the presence of the edge in the graph is p and the probability that it's not presence is 1-p. Adequately composing the probabilities presences or the not of an edge p and (1-p) I obtain the formula.

6.3

The probability that the graph contain only a triangle is:

$$P(3) = \binom{n}{3} \cdot p^3 \cdot (1-p)^{\binom{n}{2}-3}$$

where I set the parameter k at 3 because we need to calculate the probability of a triangle (3 edges) times the combinations of n nodes 3 by 3.

6.4

The probability that the graph contain only a triangle is:

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

where I set the parameter k to n-1 because a complete line is composed by n-1 edge times all the permutations of the nodes.

6.5

The expected value of edges in a graph can be calculated as:

$$E[k] = \sum_{k=0}^{\binom{n}{2}} k \cdot \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{2} \cdot p$$

because we can see that the distribution is binomial and so we can apply the property of expected value for a binomial that is $\binom{n}{2} \cdot p$.

6.6

The expected value of 3-stars in a random graph is:

$$E[X] = 4\binom{n}{4}p^3(1-p)^3$$

because $4\binom{n}{4}$ represent the possible 3-stars generated in a graph, $\binom{n}{4}$ represent the possible combinations of taking 4 vertices and 4 because each edge can be the central one. The probability of a graph with 4 nodes and 3 edges is $p^3(1-p)^3$ because the number of possible edges is 6 of which 3 are set into p and other 3 into (1-p).

6.7

We can define a k-star similarly as the point above, in particular:

$$E[X] = k\binom{n}{k}p^{k-1}(1-p)^{\binom{k}{2}-k+1}$$

here we have k vertices so the possible edges are $\binom{k}{2}$ and we need to subtract the number of edges in p that are k-1.

7 Exercise

I write a python code, it's in the folder and its name is HW1EX7Fraschetti.py