ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ & ΤΗΛΕΠΙΚΟΙΝΩΝΙΩΝ
ΕΛΛΗΝΙΚΗ ΔΗΜΟΚΡΑΤΙΑ
Εθνικόν και Καποδιστριακόν Πανεπιστήμιον Αθηνών
ΙΔΡΥΘΕΝ ΤΟ 1837
ATHENA Research & Innovation Information Technologies
ΙΕΛ LSP

M902
Βασικές Μαθηματικές Έννοιες στη Γλωσσική Τεχνολογία

Project 4

Κυλάφη Χριστίνα-Θεανώ

LT1200012

January, 2021

TABLE OF CONTENTS

Question **1**

i . $n = 3$ independent experiments (coin flips)

$$\Omega = \left\{ \begin{array}{c} \text{ККК, ККГ, КГК, КГГ} \\ \text{ГКК, ГКГ, ГГК, ГГГ} \end{array} \right\}$$

ii . $A_1 = \{ \text{ГКК, КГК, ККГ, ККК} \}$
$A_2 = \{ \text{ГКК, КГК, ККГ} \}$
$A_3 = \{ \text{ГКК, КГК, ККГ, ККК} \} = A_1$
$A_4 = \{ \text{ККК, ГГГ} \}$
$A_5 = \{ \text{ККК, ККГ, КГК, КГГ} \}$

Let X a random variable expressing the number of successes (coin flip result → **К**), following **Binomial Distribution** $Bin(\ k;\ N,\ p)$ (spoilers for iii below), $X \sim Bin(k;\ 3,\ 0.5)$. Then:

$$P(X = k) = \left(\begin{array}{c} n \\ k \end{array} \right) p^k\ (1-p)^{n-k}\ ,$$

where
p is the **probability** of “success” outcome,
k is the number of **successes**,
n the total number of independent **experiments** performed.

$$P(A_1) = P(X = 2) + P(X = 3) = \binom{3}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right) + \binom{3}{3}\left(\frac{1}{2}\right)^3 = \frac{3!}{2!1!}\ \frac{1}{8} + \frac{3!}{3!0!}\ \frac{1}{8} = = \frac{3}{8} + \frac{1}{8}$$

$$= \frac{N(A_1)}{N(\Omega)} = \frac{4}{8} = 0.5$$

$$P(A_2) = P(X = 2) = \binom{3}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right) = \frac{3!}{2!1!}\ \frac{1}{8} = \frac{3}{8} = \frac{N(A_2)}{N(\Omega)} = 0.375$$

$$P(A_3) = P(A_1) = 0.5$$

$$P(A_4) = P(X = 0) + P(X = 3) = \binom{3}{0}\left(\frac{1}{2}\right)^3 + \binom{3}{3}\left(\frac{1}{2}\right)^3 = \frac{3!}{0!3!}\ \frac{1}{8} + \frac{3!}{3!0!}\ \frac{1}{8} = = \frac{1}{8} + \frac{1}{8}$$

$$= \frac{N(A_1)}{N(\Omega)} = \frac{2}{8} = 0.25$$

Event A_5 concerns only the first coin flip, which is independent of the overall number of experiments. Therefore, the probability of a sole coin flip (the first one) resulting in K, is always $P(K) = \frac{1}{2} = 0.5$.

iii . n independent experiments (coin flips)

Here, for event A_2 we apply the same formula as in ii, with $X \sim Bin(k;\ n,\ 0.5)$:

$$P(A_2) = P(X = 2) = \binom{n}{2}\left(\frac{1}{2}\right)^2\left(\frac{1}{2}\right)^{n-2} = \frac{n!}{2!(n-2)!}\left(\frac{1}{2}\right)^n = \frac{n\ (n-1)}{2}\left(\frac{1}{2}\right)^n$$

As also mentioned in ii , the probability of A_5 is always the same and equals the probability of a single coin flip resulting in K, $P(K) = \frac{1}{2} = 0.5$.

Question **2**

Let X a random variable following Normal Distribution $X \sim N(\ 60,\ 5^2\ )$ expressing the student weights. Then:

$$\alpha)\ \ P(X > 70) = P\left(\frac{X-\mu}{\sigma} > \frac{70-\mu}{\sigma}\right) = P\left(Z > \frac{70-60}{5}\right) = P(Z > 2) = 1 - P(Z < 2)$$

$$= 1 - \Phi(2) = 1 - 0.9772 = 0.0228$$

$$\beta)\ \ P(55 < X < 65) = P(X < 65) - P(X < 55) = P\left(\frac{X-\mu}{\sigma} < \frac{65-\mu}{\sigma}\right) - P\left(\frac{X-\mu}{\sigma} < \frac{55-\mu}{\sigma}\right)$$

$$= P\left(\frac{X-\mu}{\sigma} < \frac{65-60}{5}\right) - P\left(\frac{X-\mu}{\sigma} < \frac{55-60}{5}\right) = P(Z < 1) - P(Z < -1)$$

$$= \Phi(1) - \Phi(-1) = \Phi(1) - (1 - \Phi(1)) = 2\ \Phi(1) - 1 = 2 * 0.8413 - 1 = 0.6826$$

Question **3**

$P(\text{ αvoσ }) = 0.7 = p$

The problem can be modelled as a binary outcome (rabbit immunised or not) experiment, executed n times (selecting n rabbits). Then, X is a random variable expressing the number of immunised rabbits picked, with $X \sim Bin(k;\ n,\ 0.7)$, where $n = 5$:

$i.\ \ P(X = 3) = \binom{5}{3} * 0.7^3 * (1 - 0.7)^{5-3} = \frac{5!}{3!2!} * 0.7^3 * 0.3^2 = 10 * 0.7^3 * 0.3^2 = 0.3087$

$ii.$ Here, two explanations of the question are going to be followed. However, the resulted probabilities are equal.

1. The probability of picking 3 non-immunised (failure) rabbits and then 1 immunised (success). The task can be modelled as the calculation of the probability that the first success (immunised rabbit) requires k independent trials, thus we calculate the probability of $k - 1$ failures and 1 success (k_{th} trial). In this particular case, X is following the **Geometric Distribution**, $X \sim Geo(0.7)$:

$$P(X = k) = (1 - p)^{k-1} p$$

Then:

$$P(X = 4) = (1 - 0.7)^{4-1} 0.7 = 0.3^3 * 0.7 = 0.0189$$

2. The probability of the first rabbit to be the only immunised one, out of 4 rabbits picked in total.

$$P(1_{st} \text{ rabbit immunised }) = 0.7 * (1 - 0.7)^{4-1} = 0.7 * 0.3^3 = 0.0189$$

Question **4**

The solution of this problem, was calculated through code developed in Python. The results are presented below:

		Class	Sentence
Training	**1**	**-**	μη χάσετε το χρόνο σας
	2	**+**	καταπληκτικές ερμηνείες σε ένα δύσκολο έργο
	3	**+**	η καλύτερη θεατρική παράσταση του χειμώνα
	4	**-**	δεν ήταν ευχάριστη
	5	**+**	μια ευχάριστη έκπληξη
Test	**1**	**?**	πέρασα μια ευχάριστη θεατρική βραδιά
	2	**?**	δεν πέρασα μια ευχάριστη θεατρική βραδιά

Word list of concatenated sentences of positive class “ + ” :
{ 'καταπληκτικές', 'ερμηνείες', 'σε', 'ένα', 'δύσκολο', 'έργο', 'η', 'καλύτερη', 'θεατρική', 'παράσταση', 'του', 'χειμώνα', 'μια', 'ευχάριστη', 'έκπληξη' }
Count: 15

Word list of concatenated sentences of negative class “ - ” :
{ 'μη', 'χάσετε', 'το', 'χρόνο', 'σας', 'δεν', 'ήταν', ‘ευχάριστη' }
Count: 8

Word set (union) of the above (all sentences):
{ 'παράσταση', 'σε', 'ερμηνείες', 'καταπληκτικές', 'το', 'δύσκολο', 'ήταν', 'καλύτερη', 'έκπληξη', 'ευχάριστη', 'έργο', 'ένα', 'μη', 'του', 'η', 'μια', 'χρόνο', 'χάσετε', 'δεν', 'θεατρική', 'σας', ‘χειμώνα' }
Count: 22

$$P(-) = \frac{N_{sentences\ of\ the\ class}}{N_{total\ sentences}} = \frac{2}{5} = 0.4$$

$$P(+) = \frac{N_{sentences\ of\ the\ class}}{N_{total\ sentences}} = \frac{3}{5} = 0.6$$

We then calculate the conditional probability of all the possible classes (negative / positive), given each test sentence. The maximum probability dictates the predicted class of the respective sentence, by the Naive Bayes classifier :

$$P(c \mid S_N) = P(c) \prod_{w \in W_{S_N}} P(w \mid c)\,,$$

where
W_{S_N} the words comprising test sentence S_N being examined, which are also included in the training set, otherwise they are omitted
$c \in C$, with C being the set of classes

i. $P(-\mid S_1) = P(-) \prod_{w \in W_{S_1}} P(w \mid -) = P(-)\ P(\text{μια} \mid -)\ P(\text{ευχάριστη} \mid -)\ P(\text{θεατρική} \mid -)$

$$= 0.4 * \frac{0+1}{8+22} * \frac{1+1}{8+22} * \frac{0+1}{8+22} = 0.4 * \frac{2}{30^3} = 2.\overline{962} * 10^{-5}$$

$$P(+\mid S_1) = P(+) \prod_{w \in W_{S_1}} P(w \mid +) = P(+)\ P(\text{μια} \mid +)\ P(\text{ευχάριστη} \mid +)\ P(\text{θεατρική} \mid +)$$

$$= 0.6 * \frac{1+1}{15+22} * \frac{1+1}{15+22} * \frac{1+1}{15+22} = 0.6 * \frac{8}{37^3} = 9.476 * 10^{-5} > P(-\mid S_1)$$

Therefore, test sentence **1** is **classified as " + "**, which is **correct** !

ii. The second sentence differs from the first one only on one word, "δεν", so we can calculate the respective probabilities by multiplying each of the previous probabilities with the term $P(\text{δεν} \mid c)$, where c the class for which we examine the sentence.

$$P(-\mid S_2) = P(-) \prod_{w \in W_{S_2}} P(w \mid -) = P(-\mid S_1)\ P(\text{δεν} \mid -) = 0.4 * \frac{2}{30^3} * \frac{2}{30}$$

$$= 0.4 * \frac{4}{30^4} = 1.975 * 10^{-6}$$

$$P(+\mid S_2) = P(+) \prod_{w \in W_{S_2}} P(w \mid +) = P(+\mid S_1)\ P(\text{δεν} \mid +) = 0.6 * \frac{8}{37^3} * \frac{1}{37}$$

$$= 0.6 * \frac{8}{37^4} = 2.561 * 10^{-6} > P(-\mid S_2)$$

Test sentence **2** is also **classified as “ + ”**, which is **incorrect** !

However, it can be explained, mostly due to the nature of Naive Bayes classifier.

For instance, the aforementioned classifier, is largely **affected** by the amount of **training data** available. That means, less training data for a class may result in a bias towards the opposite class in a binary classification task like the one above (weight shrinking for classes with fewer examples). In fact, in our case, the training data for the true class of **sentence 2** (“ - ”) are less than the opposite class (“ + ”), in which it was falsely classified from NB classifier.

In this particular case, word “δεν”, which is the only difference between the two sentences that are tested, seems to have little contribution to the decision, as it is assigned a weight not capable of changing the classification result from “ + ” (**sentence 1**) to “ - ” (**sentence 2**) .

What is more, the classifier **assumes feature independency** in a manner that ignores possible relations between the words, such as “δεν”, which gives a completely different meaning to verbs when added before them. As a result, even as an important feature that changes the semantic frame of the sentence thus its class, it is not given the proper weight, leading to misclassification.

Question **5**

$P(X) = 0.1$,
therefore $P(X') = 1 - 0.1 = 0.9$

$P(+|X) = 0.95$

$P(+|X') = 0.95$

$P(X'|+) = ??$

Applying Naive Bayes rule, we have:

$$P(X'|+) = \frac{P(+|X')\ P(X')}{P(+)} \quad (1)$$

The only unknown value is $P(+)$ which can be calculated as follows (applying the law of total probability):

$$P(+) = P(+|X)\ P(X) + P(+|X')\ P(X') = 0.95 * 0.1 + 0.07 * 0.9$$

$$= 0.095 + 0.063 = 0.158 \quad (2)$$

$$\overset{(1),\ (2)}{\Longrightarrow} P(X'|+) = \frac{P(+|X')\ P(X')}{P(+)} = \frac{0.07 * 0.9}{0.158} = \frac{0.063}{0.158} = 0.399$$

Question **6**

The general formula to make a transition from odds ($x : y$) in favour of an event to the probability of the event, is :

$$\frac{x}{x+y}$$

i . The odds for rain in Helsinki are 206:159

$$206 : 159 \rightarrow \frac{206}{206+159} = \frac{206}{365} = 0.564 = 56.4\,\%$$

ii . The odds for getting three of a kind in poker are about 1:46

$$1 : 46 \rightarrow \frac{1}{1+46} = \frac{1}{47} = 0.021 = 2.1\,\%$$

Question **7**

Let random variable X which expresses the daily product demand, following Normal Distribution, $X \sim N(\ 5000,\ 300^2)$.

Then:

$$\alpha)\ \ P(X < 5300) = P\left(\frac{X-\mu}{\sigma} < \frac{5300-5000}{300}\right) = P(Z < 1) = \Phi(1) = 0.8413$$

$$\beta)\ \ P(X < w) = P\left(\frac{X-\mu}{\sigma} < \frac{w-5000}{300}\right) = P\left(Z < \frac{w-5000}{300}\right)$$

$$= \Phi\left(\frac{w-5000}{300}\right) = 0.9 \approx \Phi(1.28)$$

$$\Longrightarrow\ \Phi\left(\frac{w-5000}{300}\right) = \Phi(1.28)$$

$$\Longrightarrow\ \frac{w-5000}{300} = 1.28$$

$$\Longrightarrow\ w = 300 * 1.28 + 5000 = 5384 \text{ products}$$

Question **8**

Let X random variable expressing student performance, following Normal Distribution, $N(\ \mu,\ \sigma^2\ )$.

Then:

$P(X < 67.2) = 0.0808 \quad (1)$

$P(X > 72.2) = 0.1375 \quad (2)$

In order to find the values of μ and σ, we perform the following calculations:

$$\overset{(1)}{\Longrightarrow} P(X < 67.2) = P\left(\frac{X-\mu}{\sigma} < \frac{67.2-\mu}{\sigma}\right) = \Phi\left(\frac{67.2-\mu}{\sigma}\right) = 0.0808$$

$$= 1 - 0.9192 = 1 - \Phi(1.4) = \Phi(-1.4) \quad (3)$$

$$\overset{(3)}{\Longrightarrow} \Phi\left(\frac{67.2-\mu}{\sigma}\right) = \Phi(-1.4) \implies \frac{67.2-\mu}{\sigma} = -1.4 \implies \mu = 67.2 + 1.4 * \sigma \quad (4)$$

$$\overset{(2)}{\Longrightarrow} P(X > 72.2) = P\left(\frac{X-\mu}{\sigma} > \frac{72.2-\mu}{\sigma}\right) = \Phi\left(-\frac{72.2-\mu}{\sigma}\right) = 0.1375$$

$$= 1 - 0.8625 = 1 - \Phi(1.1) = \Phi(-1.1) \quad (5)$$

$$\overset{(5)}{\Longrightarrow} \Phi\left(-\frac{72.2-\mu}{\sigma}\right) = \Phi(-1.1) \implies -\frac{72.2-\mu}{\sigma} = -1.1 \implies \mu = 72.2 - 1.1 * \sigma \quad (6)$$

$$\overset{(5),\ (6)}{\Longrightarrow} 67.2 + 1.4 * \sigma = 72.2 - 1.1 * \sigma \implies \sigma = \frac{5}{2.5} = 2 \quad (7)$$

$$\overset{(6),\ (7)}{\Longrightarrow} \mu = 72.2 - 1.1 * \sigma = 72.2 - 1.1 * 2 = 72.2 - 2.2 = 70$$

Hence:

$X \sim N(\ 70,\ 2^2\ )$

Question **9**

The probability density function (pdf) of normal distribution with mean **50** and variance **9** ($N(\ 50,\ 3^2\ )$) is presented below in **Figure 1**. The areas between the points with distance equal to integral multiples (up to 3) of standard deviation are also demonstrated.

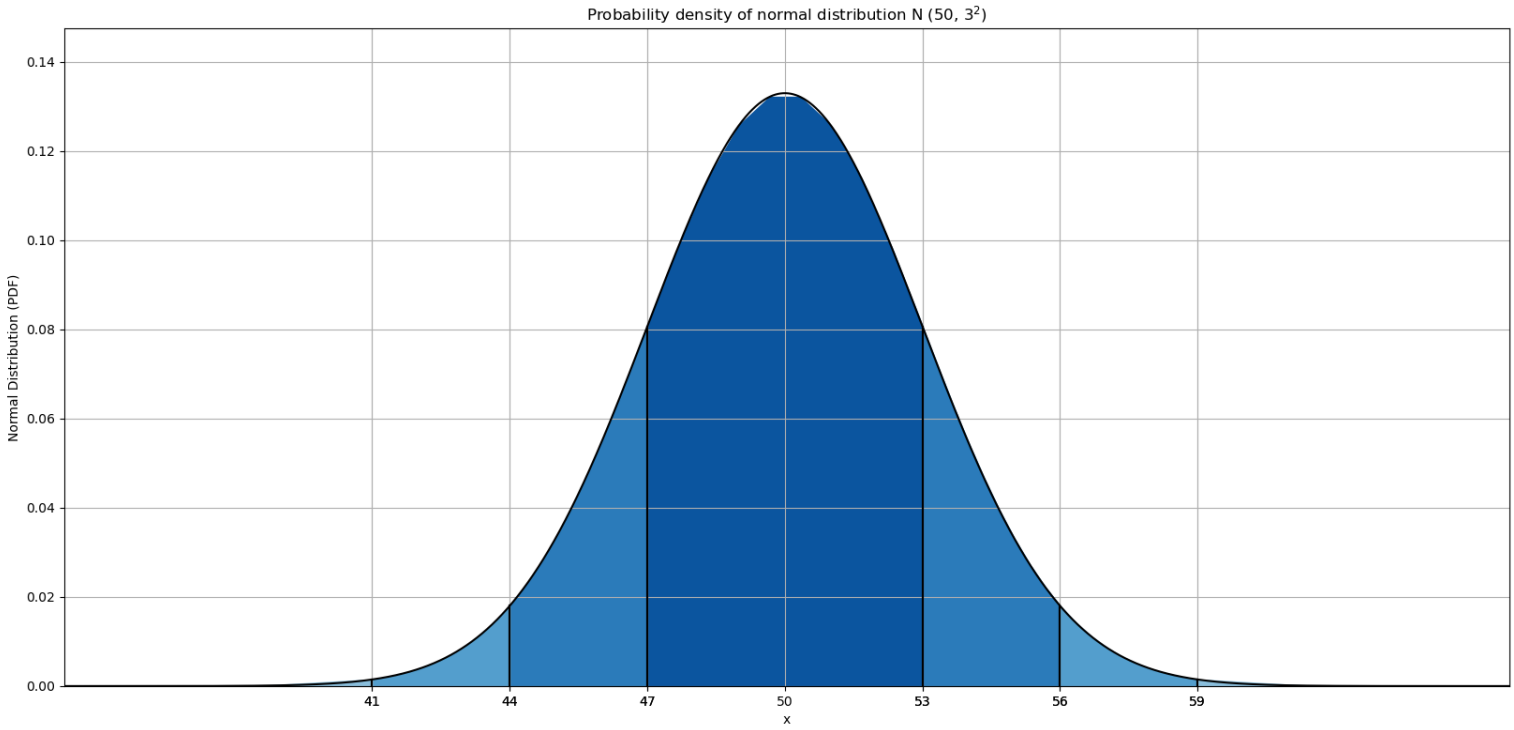

Figure 1

Question **10**

In order to check the documents for high correlation between any of their features, we could employ multivariate normal distribution technique, representing them as matrices, comprising document features. Then, we could perform the necessary calculations for the purpose of demonstrating the respective relationships. A visual representation of those relationships, could also be given by the deployment of scatter plots.

Let $f_1, f_2, \ldots, f_N$ each d_i document's vector of 5 features.

The matrix of all the documents' features will be:

$$A = \begin{bmatrix} f_{1_1} & \cdots & f_{1_5} \\ f_{2_1} & \cdots & f_{2_5} \\ & \vdots & \\ f_{N_1} & \cdots & f_{N_5} \end{bmatrix}$$

After that, we calculate the respective mean, covariance and correlation values of the variables. The most important attribute in our case is the correlation matrix C, consisting of the correlation factors, in which the relationship of all the possible variable pairs is indicated.

The existence of highly correlated features allows or even dictates, depending on the occasion, feature selection. When two features demonstrate a high level of correlation, one of them can be dropped, resulting in diminishing the computational complexity of the process and poor system performance, thus optimising the task been working on. However, the choice of the feature to delete, is of high importance. In some cases, even highly correlated with another, a feature could carry important information of high impact (large weight) on the prediction to be made (high correlation to the target). So, there is always the risk of a cost due to information loss when deleting a feature, which could be greatly considerable.

Therefore, another approach with computational complexity consideration could be followed, such as dimensionality reduction. When applying the aforementioned method, no loss of information is occurring, as the data is mapped to a lower dimension through the projection of all the features available in the dataset.