INTRODUCTION TO ARTIFICIAL INTELLIGENCE – LECTURE 2 – CLASSIFICATION

Name and surname: Sofya Aksenyuk Index no.: 150284

I. Let us consider the following documents **D1-D3** and a set of terms **T** = {t_1=artificial, t_2=intelligence, t_3=introduction}:

- **D1** = {artificial intelligence artificial intelligence}
- **D2** = {artificial artificial introduction}
- **D3** = {artificial intelligence introduction artificial}

a) Represent D1-D3 using the binary setting, Bag-Of-Words (BOW), and Term-Frequency (TF):

Binary

	t_1	t_2	t_3
D1	1	1	0
D2	1	0	1
D3	1	1	1

BOW

	t_1	t_2	t_3
D1	2	2	0
D2	2	0	1
D3	2	1	1

TF

	t_1	t_2	t_3
D1	2/2	2/2	0/2
D2	2/2	0/2	1/2
D3	2/2	1/2	1/2

b) Assume the documents are represented by means of a binary setting. What is the Jaccard coefficient for D1 and D2 or D1 and D3?

Answer: *Jac*(D1,D2)= $\frac{1}{3}$ *Jac*(D1,D3)= $\frac{2}{3}$

c) What is the cosine similarity for D1 and D3, when using the TF representation?

Answer: for TF: *cos*(D1,D3)= $\frac{1 + 1/2 + 0}{\sqrt{1 + 1} * \sqrt{1 + 1/4 + 1/4}} = \frac{1 + 1/2}{1{,}41 * 1{,}22} = \frac{1{,}5}{1{,}72} = 0{,}87$

II, Let us consider the following similarities of Y with D1-D10 and classes A, B, and C.

	D1	D2	D3	D4	D5	D6	D7	D8	D9	D10		**Y**
Class	A	C	B	C	B	A	A	C	A	B		**?**
Similarity with Y	0.9	0.8	0.6	0.6	0.5	0.4	0.4	0.2	0.1	0.0		**1.0**

a) Use 4-NN to provide a classification for Y using the majority voting rule:

Answer: scores for A = 1(D1) = 1 B = 1(D3) = 1 C = 1(D2) + 1(D4) = 2; recommended class = C

b) Use 4-NN to provide a classification for Y using the weighted voting rule:

Answer: scores for A = 0.9(D1) = 0.9 B = 0.6(D3) = 0.6 C = 0.8(D2) + 0.6(D4) = 1.4; recommended class = C

c) Would the results for the majority voting rule be different in case of using either k=7?

Answer: for k=7 - recommended class = A

d) What should be the minimal k so that A is recommended using the majority voting rule?

Answer: k = 7

III. Let us consider the following collection of documents involving five training documents (D1-D5) and one test document D6. These documents contain terms T1, T2, and T3 (see table below).

	Doc	Content	Class
training	D1	T1 T2	A
training	D2	T3 T3	A
training	D3	T1 T3	A
training	D4	T1 T2 T2	B
training	D5	T1	B
test	**D6**	T1 T3 T3	?

a) Assume the documents are represented with the binary vector of terms (see table below to the left). Use the Naive Bayes classifier to classify D6, and answer the following questions by filling in the respective empty spaces below: What is the binary representation of D5? What is P(A)? What is P(T3=1|A)? (in case of 0, assume 0.1 as a simple smoothing technique) What class would be recommended for D6 and why (compare P(A|D6) and P(B|D6))?

Doc	T1	T2	T3	Class
D1	1	1	0	A
D2	0	0	1	A
D3	1	0	1	A
D4	1	1	0	B
D5	1	0	0	B
D6	1	0	1	?

P(A) = 3/5

P(T1=1|A) = 2/3

P(T2=0|A) = 2/3

P(T3=1|A) = 2/3

P(A|D6) ≈ 0.17

P(B) = 2/5

P(T1=1|B) = 1

P(T2=0|B) = 1/2

P(T3=1|B) = 0 ≈ **0.1**

P(B|D6) ≈ 0.02

Answer: D6 would be assigned to class A, because P(A | D6) > P(B | D6)

b) Assume the documents are represented in terms of Bag-Of-Words.

What is the size of the dictionary? **Answer**: 3

What is P(B)? What is P(T3|B) while using add-one smoothing technique? What class would be recommended for Y and why (compare P(A|D6) and P(B|D6))?

P(A) = **3/5**

P(T1|A) = (2+1)/(**6**+3) = 3/9

P(T3|A) = (3+1)/(**6**+3) = 4/9

P(A|D6) ≈ $3/5 \cdot 3/9 \cdot (4/9)^2 = 0.039$

P(B) = 2/5

P(T1|B) = (2+1)/(4+3) = 3/7

P(T3|B) = (0+1)/(4+3) = 1/7

P(B|D6) ≈ 2/5 * 3/7 * (1/7)^2 = 0.003

Answer: D6 would be assigned to class A, because P(A | D6) > P(B | D6)

IV. Let us consider the following confusion matrix for the classification problem involving four classes $C_1 - C_4$ and 100 documents.

confusion matrix		predicted class			
		C_1	C_2	C_3	C_4
original (actual) class	C_1	20	0	0	0
	C_2	5	15	0	5
	C_3	0	10	15	0
	C_4	5	5	0	20

Compute the following measure values:

a) Classification accuracy = $\frac{70}{100} = 0,7$

b) Misclassification error = $1 - 0,7 = 0,3$

c) Recall for C_1 = $\frac{20}{20}$ Recall for C_4 = $\frac{20}{30}$

d) Precision for C_2 = $\frac{15}{30}$ Precision for C_3 = $\frac{15}{15}$

e) Can you immediately say which class has the highest recall / precision?

Answer: The class with the highest recall is: C1 , and the class with the highest precision is: C3 .

V. Considering the above confusion matrix and the below cost matrix, compute the misclassification cost.

cost	C_1	C_2	C_3	C_4
C_1	0	1	1	2
C_2	1	0	3	3
C_3	1	3	0	1
C_4	2	5	1	0

Answer: Misclassification cost = $1 + 1 + 2 + 3 + 1 + 1 + 1 = 10$