ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE
EIDGENÖSSISCHE TECHNISCHE HOCHSCHULE – LAUSANNE
POLITECNICO FEDERALE – LOSANNA
SWISS FEDERAL INSTITUTE OF TECHNOLOGY – LAUSANNE

Faculté Informatique et Communication
Introduction to Natural Language Processing (Ms; CS-431)
Chappelier, J.-C. & Rajman, M.

CS-431 Hands On Text Classification

J.-C. Chappelier M. Rajman

v. 2021118 – 1

QUESTION I [3 pt]

(from Fall 2018 quiz 4)

The Naïve Bayes algorithm is used in the framework of a sentiment analysis application to determine, for any input tweet, which, among a predefined set of sentiments, best corresponds to the mood expressed in the tweet.

Does the performed tweet classification task have to be supervised in this case?

[✓] yes [ ] no [ ] it depends on the implementation

Let us assume that only two sentiments are considered (“joyful” and “sad”) and that typically 70% of the tweets are “joyful”.

↳ $P(\text{joyful})$ ⟶ $P(\text{Sad}) = 30\%$

To which sentiment would the Naïve Bayes algorithm associate a tweet indexed by only two terms w_1 and w_2 , if:

$P(w_1|\text{joyful})$ $P(w_1|\text{Sad})$

- 10% of the occurrences of indexing terms in “joyful” tweets and 20% of the occurrences of indexing terms in “sad” tweets are w_1 ; while
- 30% of the occurrences of indexing terms in “joyful” tweets and 25% of the occurrences of indexing terms in “sad” tweets are w_2?

$P(w_2|\text{Sad})$

[ ] sad [ ] joyful [ ] undecidable

↳ $70 \cdot 10 \cdot 30 = 210$

$P(\text{sad}) \cdot P(w_1|\text{Sad}) \cdot P(w_2|\text{Sad}) = 30 \cdot 20 \cdot 25 = 150$

$$\sum_x P(X=x|Y=y) = 1$$

QUESTION II [2 pt]

(from Fall 2017 quiz 4)

Consider the following matrix of measures over a set of three items:

0	5	2
5	0	2
2	2	0

$$m(1,3) = 5 > \underset{2}{m(1,2)} + \underset{2}{m(2,3)}$$

What type(s) of measure is this matrix compatible with?

[ ] A dissimilarity only.

[ ] A dissimilarity and a distance/metric.

[ ] None of the two

QUESTION III [4 pt]

(from Exam 2019)

You're working on an email classification software (and have some corpus).

In order to better understand your corpus, you plan to cluster it using dendrograms. To do so:

- you represent each email body by the empirical probability distribution over the tokens it contains (simply estimated by their relative frequencies);
- and make use of the Hellinger distance.

What is the distance between the following two email bodies:

$$d(x,y) = N(x-y)$$

email 1: *ski sun money sun*

Hellinger = Euclidian($\sqrt{\ }$)

email 2: *sun ibm sun apple money sun money sun*

	ski	sun	money	ibm	apple	
e1	1	2	1	0	0	→ 4
e2	0	4	2	1	1	→ 8
Δ:	1/4	0	0	1/8	1/8	

$$\sqrt{\frac{1}{4} + \frac{2}{8}} = \frac{1}{\sqrt{2}}$$

QUESTION IV [3 pt]

(from Exam 2019)

You run the dendrogram clustering algorithm using complete linkage. At some point, it reaches a state where what remains to be clustered are the two clusters, G_1 and G_2, that have already been build so far, and two email bodies, B_1 and B_2. Here are the distances between each of them:

	B_1	B_2	G_1	G_2
B_1	0	0.7	0.6	0.2
B_2	0.7	0	0.5	0.3
G_1	0.6	0.5	0	0.4
G_2	0.2	0.3	0.4	0

first

→

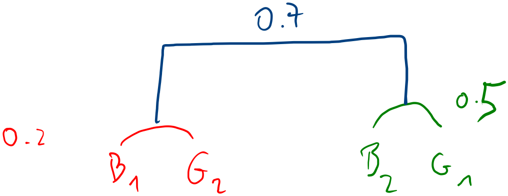
	new B_1G_2
B_2	0.7
G_1	0.6

0.5 ←

second

Draw the dendrogram corresponding to the final clustering.

third



① min $d(\text{element}_1, \text{element}_2)$

② distance between groups? d

complete (single, average)

↓ ↓ ↓

max min avg