COMP 3010J - Graded Assignment-1
17-October-2019

1. Consider the following dataset of movies, with a count of scenes labelled as **Romantic** or **Action**

Movie Name	Number of fight scenes	Number of romantic scenes	Class
California Man	3	104	Romantic movie
Beautiful Woman	1	81	Romantic movie
Kevin Longblade	101	10	Action movie
Amped II	98	2	Action movie

An unknown movie that has not been classified has the following count of scenes:

Number of fight scenes – 18
Number of romantic scenes – 90

a. Use the *Minkowski* distance function with $p = 2$, to calculate the distance between the unknown movie and each movie in the dataset. [**4 marks**]

Marking Scheme: The minkowski distance formula is given, so no marks for writing that. The distance for each movie from the unknown movie is as follows:

1. California Man – 20.51
2. Beautiful Woman – 19.23
3. Kevin Longblade – 115.27
4. Amped – 118.92

1 mark for each correct distance calculation. Be flexible about the last decimal digit. As long as the first decimal is ok, give marks.

b. What class would a 3-NN classifier assign to the unknown movie? [**2 marks**]

Marking Scheme: 3-NN would classify the movie as **Romantic Movie**. All-or-nothing. 2 marks for the correct answer. No marks for any other answer.

2. Consider the following dataset

ID	color	height	class
01	red	tall	good
02	blue	short	bad
03	blue	medium	none
04	yellow	medium	none
05	red	medium	bad
06	red	short	bad
07	yellow	short	good
08	yellow	tall	bad

a) Calculate the entropy of the dataset using Shannon's entropy [**2 marks**]

Marking Scheme: There is nothing complicated here. A very simple application of the formula. The formula was available to them. Hence, this is an all-or-nothing answer.

The correct answer is: 1.5 bits. The student **must** show full working to get full marks.

b) Calculate the impurity of the dataset using Gini index [**2 marks**]

Marking Scheme: Same as above. The formula was available to the students. Hence, this is an all-or-nothing answer. **The correct answer is: 0.625 [up to two decimal digits is okay]**.

The student must show full working to get full marks.

3. In a dataset of 1000 fruits, the following counts have been obtained for 3 features

Type	**Long**	**Not Long**	**Sweet**	**Not Sweet**	**Yellow**	**Not Yellow**	**Total**
Banana	400	100	350	150	450	50	500
Orange	0	300	150	150	300	0	300
Mango	100	100	150	50	50	150	200
							1000

a) Calculate the prior probabilities for Banana, Orange and Mango [**1 mark**]

Marking Scheme: The prior probabilities are extremely easy to calculate. No marks for any working. 1/3 mark for each correct prior probability calculation

P(Banana) = 500 / 1000 = 0.5

P(Orange) = 300 / 1000 = 0.3

P(Mango) = 200 / 1000 = 0.2

b) Create a Naïve Bayes' classifier that will classify a fruit that is 'Long', 'Yellow' and 'Sweet'. Show your working.

[**4 marks**]

First, calculate the probability of the evidence

P(Long) = 500 / 1000 = 0.5

P(Sweet) = 650 / 1000 = 0.65

P(Yellow) = 800 / 1000 = 0.8

Marking Scheme: 1/3 for each probability

Second, lay out the calculation for posterior probability

P (Banana | Long, Sweet and Yellow) = P(Long | Banana) * P(Sweet| Banana) * P(Yellow|Banana) * P(Banana) **/** P(Long) * P(Sweet) * P(Yellow)

P (Orange | Long, Sweet and Yellow) = P(Long | Orange) * P(Sweet| Orange) * P(Yellow|Orange) * P(Orange) **/** P(Long) * P(Sweet) * P(Yellow)

P (Mango | Long, Sweet and Yellow) = P(Long | Mango) * P(Sweet| Mango) * P(Yellow|Mango) * P(Mango) **/** P(Long) * P(Sweet) * P(Yellow)

P(Long) * P(Sweet) * P(Yellow) = 0.5 * 0.65 * 0.8 = 0.26

Marking Scheme: 1/3 for each posterior probability written down. No marks if there's no showing of the full formula

NOTE: The denominator is the same in each case, so even if they don't calculate the actual value in the next step, it doesn't matter.

Third, calculation of the likelihood of Banana, Orange and Mango

P(Long | Banana) = 400 / 500 = 0.8

P(Sweet | Banana) = 350 / 500 = 0.7

P(Yellow | Banana) = 450 / 500 = 0.9

P (Banana | Long, Sweet and Yellow) = 0.8 * 0.7 * 0.9 * 0.5 / 0.26 = **0.252/0.26** = **0.96**

P(Long | Orange) = 0 / 300 = 0

P(Sweet | Orange) = 150 / 300 = 0.5

P(Yellow | Orange) = 300 / 300 = 1

P (Orange | Long, Sweet and Yellow) = 0 * 0.5 * 0.1 * 0.3 / 0.26 = **0/0.26** = **0**

P(Long | Mango) = 100 / 200 = 0.5

P(Sweet | Mango) = 150 / 200 = 0.75

P(Yellow | Mango) = 50 / 200 = 0.25

P (Mango | Long, Sweet and Yellow) = 0.5 * 0.75 * 0.25 * 0.2 / 0.26 = **0.0187/0.26** = **0.07**

Marking Scheme: 2/3 for each correct posterior probability for each fruit. It does not matter if they ignore the denominator and calculate only the numerator.