Workshop Week 11 - COMP20008 2020SM1

1. Consider the following data set for a binary class problem:

Feature A	Feature B	Class Label
T	F	+
T	T	+
T	T	+
T	F	-
T	T	+
F	F	-
F	F	-
F	F	-
T	T	-
F	F	-

We wish to select the feature that best predicts the class label using the χ^2 method.

- Write down the observed and expected contingency tables for feature A
- Calculate the $\chi^2(A, Class)$ value.
- Using the table below, conclude whether feature A is independent of the class label for $p = 0.05$.

df	P = 0.05	P = 0.01	P = 0.001
1	3.84	6.64	10.83
2	5.99	9.21	13.82
3	7.82	11.35	16.27
4	9.49	13.28	18.47
5	11.07	15.09	20.52
6	12.59	16.81	22.46

- Repeat the process for feature B and decide which feature could be best used for predicting the class label.

2. Open Jupyter Notebook and implement a coded solution to Question 1. Ensure that your code gives the same answer as your calculation in Question 1.

3. Consider the following dataset:

User	Iron Man	Superman	Batman	Spiderman	Ant-Man	Wonder Woman
Anne	3		3	3.5	2.5	3
Bob	4	3.5	2.5	4	3	3
Chris	3	3	3			4
Dave		3.5	2	4	2.5	
Eve		3		3	2	5
Frank	2.5	4		5	3.5	5
Gary		4.5	3	4	2	

- Use the Item-based recommender systems approach discussed in lectures to predict Frank's rating for Batman.
- Use the User-based recommender systems approach to predict Frank's rating for Batman
- Identify the advantages and disadvantages of each approach

4. Recommender systems are challenged by the "cold start" problem - how to make recommendations to new users, about whom little is known, and how to make recommendations about new items. Suggest three strategies that might be used to address this.

5. Recommender systems are sometimes criticised for over-recommending popular items to users and under-recommending rarer items. Why do you think this happens? How might it be addressed?

Exam 2016 - Questions 1d 3a-c

Consider the following dataset D which describes 3 people:

Age	Weight
20	40
30	50
25	25

a) (2 marks) Using a Euclidean distance function, construct the dissimilarity matrix for D. Show all working. (You may leave any square root terms unsimplified).

b) (2 marks) A dissimilarity matrix can be visualised as a heat map. Explain why it can be useful to reorder a dissimilarity matrix, using a technique such as the VAT algorithm. What information can this provide and why?

c) (2 marks) Would there be a benefit of applying principal components analysis to D to assist in visualisation? Explain.

Exam 2016 - Question 3

d) Bob and Alice are having a debate about imputation methods for missing values. Bob argues that imputation using the mean of a feature works best. Alice argues that imputation using matrix factorization is better. To decide who is correct, they select a dataset with 10 features $F_1, \ldots, F_{10}$ and 100 instances $x_1, \ldots, x_{100}$. For each instance, they randomly make one of its feature values to be missing (yielding 100 missing values overall). They then separately apply each of the two imputation methods and assess which performs better in recovering the missing values.

i) (2 marks) What metric would be appropriate here for assessing performance of an imputation method? Provide a formula for your metric and justify it.

ii) (3 marks) Bob and Alice are joined by a third student, Barbara. Barbara says "You are both wrong. If an instance has any missing values, it is better to discard it entirely rather than trying to perform imputations." Describe three scenarios which support Barbara's reasoning.

Exam 2018 - Question 10

Consider the following steps of the k-NN algorithm to classify a single test instance

```
1. Compute distance of the test instance to each of the instances
in the training set and store these distances
2. Sort the calculated distances
3. Store the K nearest points
4. Calculate the proportions of each class
5. Assign the class with the highest proportion
```

Step 1 may be very slow if the training set is large.
Suggest three possible strategies that might be used to speed up this step and describe a disadvantage of each.