THE UNIVERSITY OF WARWICK

MSc Examinations: Summer 2017

CS9091 Data Mining

Time allowed: 2 hours.

Answer ANY FOUR questions.
Read carefully the instructions on the answer book and make sure the particulars required are entered on each answer book.
Approved calculators are allowed

---

1. (a) List four statistics that can be calculated on a dataset consisting of only ordinal attributes. [2]

   (b) If the income range £5,000 to £85,000 is normalised to the range $[1.0, 10.0]$ using the min-max normalisation method, what new income value is £45,000 mapped to? [2]

   (c) A dataset contains the following values for one of its attributes: [4]

   34, 4, 21, 25, 28, 26, 8, 29, 21, 9, 24, 9

   What would be the result of smoothing the attribute into 4 bins of equal size by i) bin means and ii) bin boundaries?

   (d) Sampling is performed to reduce the size of a dataset. Explain the difference between simple random sampling and stratified sampling and why the latter may be preferred. [4]

   (e) Explain the similarities and differences between the ChiMerge and ChiSplit attribute discretization algorithms. [4]

   (f) The ChiMerge algorithm is being applied to a dataset and after several iterations produces the following frequency table (Table 1) for two intervals: Calculate the χ^2 statistic and use this to determine if the rows can be merged at a 90% significance level. χ^2 threshold values for selected degrees of freedom and significance levels are shown in Table 2 overleaf. [9]

Interval	**Yes**	**No**	**Total**
$3 \leq A < 10$	6	2	8
$11 \leq A < 15$	1	3	4
Total	**7**	**5**	**12**

Table 1

Degrees of freedom	90% Significance Level	95% Significance Level	99% Significance Level
1	2.71	3.84	6.64
2	4.61	5.99	9.21
3	6.25	7.82	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.65	12.59	16.81
7	12.02	14.07	18.48
8	13.36	15.51	20.09

Table 2

Continued

2. (a) State and explain the meaning of Bayes' Theorem. [2]

   (b) A screening test for stroke is known to provide a positive result 89% of the time when a patient suffering a stroke is tested, while it gives a negative result for 92% of patients tested who have not had a stroke. The risk of stroke within the general population is 4.5%.

   i. Given that a patient has tested positive, use Bayes' Theorem to decide on the probability that the patient has had a stroke. [4]

   ii. The doctor orders a second test, the result of which is positive. Determine how this would affect the probabilities of the patient having and not having a stroke. State any assumptions you make in calculating the probabilities in this case. [6]

   (c) Write down and explain the meaning of an expression defining how a Naïve Bayes classifier can be used to classify a new instance x. State and justify any assumptions you make. [4]

   (d) Table 3 contains 12 instances from a dataset for animals. Explain how you would apply Bayes' Theorem to determine the most likely classification for the new instance x = $< no, yes, yes, no >$. Show all your workings. [9]

Name	**Give Birth**	**Can Fly**	**Live in Water**	**Have Legs**	**Class**
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Table 3

3. (a) i. What is the kernel trick and why do we use it? [3]

   ii. Write the algorithm of a kernel perceptron. [5]

   iii. Which is more efficient, the kernel perceptron or the support vector machines algorithm? Explain why. [2]

   (b) What is the role of slack variables in the support vector machines algorithm? [2]

   (c) How do we regulate the effect of the slack variables in the support vector machines algorithm and how does this affect the complexity of the resulting model? [3]

   (d) Consider Figure 1 showing a soft margin support vector machine classifying four instances. The values for the Lagrange multipliers α_i are $\alpha_1 = C = 5/16, \alpha_2 = 0, \alpha_3 = 1/16$ and $\alpha_4 = 1/4$.

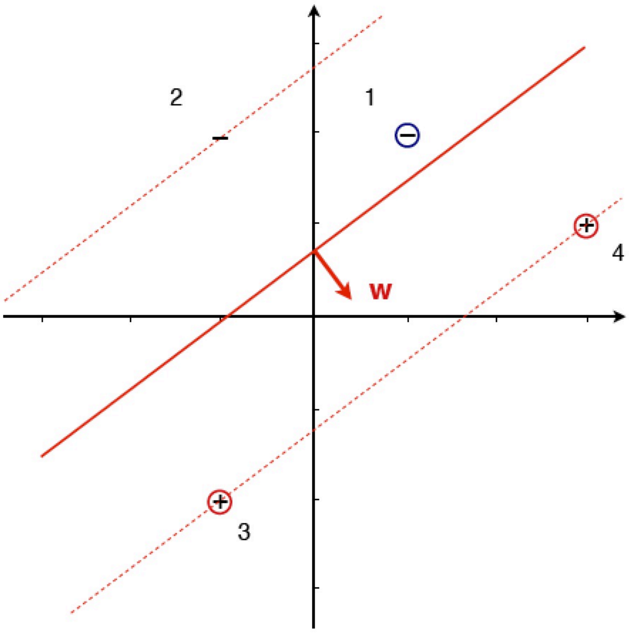


Figure 1

   i. What is happening with instance one? [2]

   ii. Is instance two a support vector? Justify your answer. [2]

   iii. What will happen if the value of the C parameter decreases to 3/16? [2]

   (e) Assume you want to train an SVM classifier on a dataset containing 150 instances described by 16 features each.

   i. Which parameters do you need to specify to the training algorithm? [3]

   ii. How do you obtain the values of these parameters? [1]

4. (a) Before any data mining algorithms can be applied to a document the text needs to be converted to matrix format. Describe the process of converting a document to a 'Bag of Words'. [4]

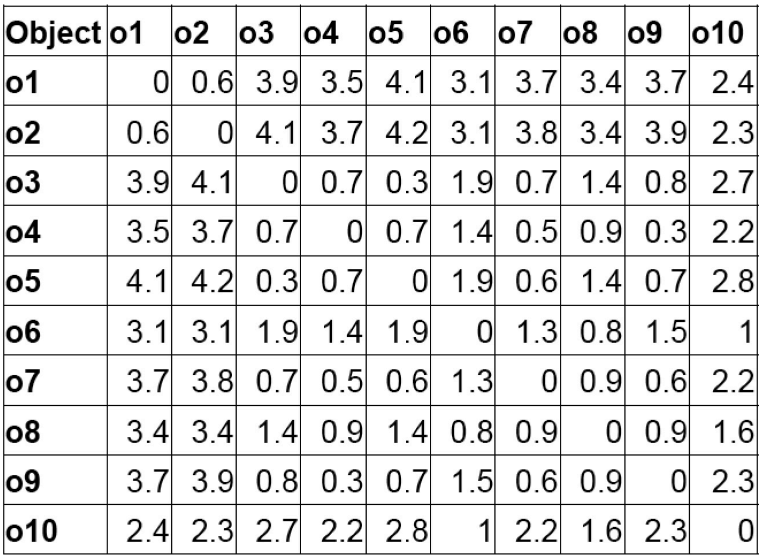
(b) In the context of text mining:

i. Describe a *Binary* weighting scheme. [2]

ii. Describe a *Normalised Term Frequency* weighting scheme. [2]

iii. Give the binary and normalised term frequency bag of words representation for the document collection consisting of the sentences "He saw the man", "The man wore the hat". [4]

(c) Consider the pairwise distance between 10 objects in Figure 2.



Object	o1	o2	o3	o4	o5	o6	o7	o8	o9	o10
o1	0	0.6	3.9	3.5	4.1	3.1	3.7	3.4	3.7	2.4
o2	0.6	0	4.1	3.7	4.2	3.1	3.8	3.4	3.9	2.3
o3	3.9	4.1	0	0.7	0.3	1.9	0.7	1.4	0.8	2.7
o4	3.5	3.7	0.7	0	0.7	1.4	0.5	0.9	0.3	2.2
o5	4.1	4.2	0.3	0.7	0	1.9	0.6	1.4	0.7	2.8
o6	3.1	3.1	1.9	1.4	1.9	0	1.3	0.8	1.5	1
o7	3.7	3.8	0.7	0.5	0.6	1.3	0	0.9	0.6	2.2
o8	3.4	3.4	1.4	0.9	1.4	0.8	0.9	0	0.9	1.6
o9	3.7	3.9	0.8	0.3	0.7	1.5	0.6	0.9	0	2.3
o10	2.4	2.3	2.7	2.2	2.8	1	2.2	1.6	2.3	0

Figure 2

i. Describe the single linkage Hierarchical Agglomerative Clustering (HAC). [4]

ii. Suggest what is a sensible number of clusters for clustering this dataset when using the single linkage Hierarchical Agglomerative Clustering (HAC). Justify your answer. [4]

iii. Express the groupings of objects into clusters when $K = 4$. [2]

iv. What is one of the disadvantages of using K-means for clustering? [1]

(d) Why do ensemble models have better performance than individual classifiers? [2]

5. (a) State the general backpropagation algorithm for training an artificial neural network with an arbitrary number of hidden layers using the stochastic method. [4]

(b) Given the XOR artificial neural network shown in Figure 3, with initial weights $w_{13} = 0.5$, $w_{14} = 0.9$, $w_{23} = 0.4$, $w_{24} = 1.0$, $w_{35} = -1.2$, $w_{45} = 1.1$; threshold values: $\theta_3 = 0.8$, $\theta_4 = -0.1$, $\theta_5 = 0.3$; learning rate of 0.5 and the instance $x_1 = 1.0$, $x_2 = 0.0$, use the backpropagation algorithm with the stochastic method to calculate:

i. The error gradient δ_5 for unit 5. [1]

ii. The weight corrections Δw_{35} and Δw_{45}. [2]

iii. The error gradients δ_3 and δ_4 for the hidden layer units 3 and 4. [2]

iv. The weight corrections Δw_{13}, Δw_{14}, Δw_{23} and Δw_{24} . [4]

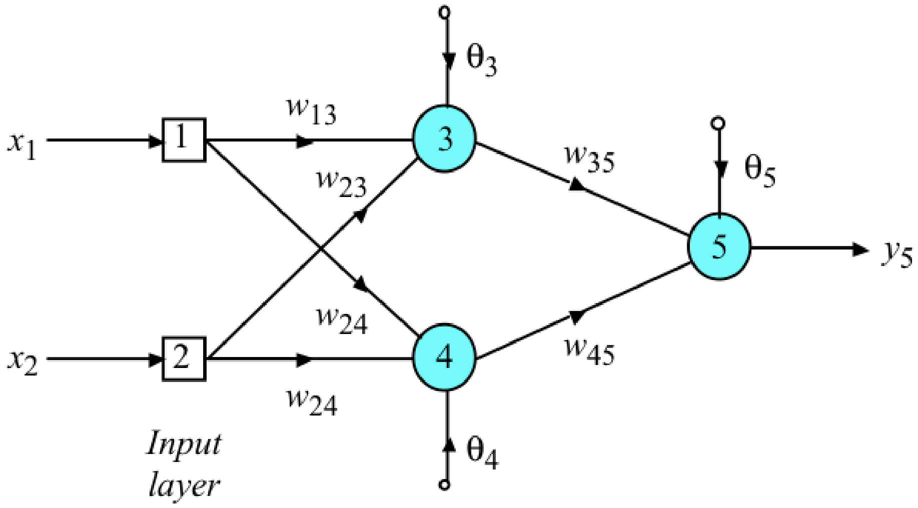


Figure 3

(c) Can one guarantee the convergence of the perceptron in linearly separable data? If yes, how? [2]

(d) A classifier is required to separate positive data (training data shown by circles) from negative data (training data shown by squares). Two candidate linear classifiers are presented in Figure 4 overleaf by the lines A and B respectively.

Continued

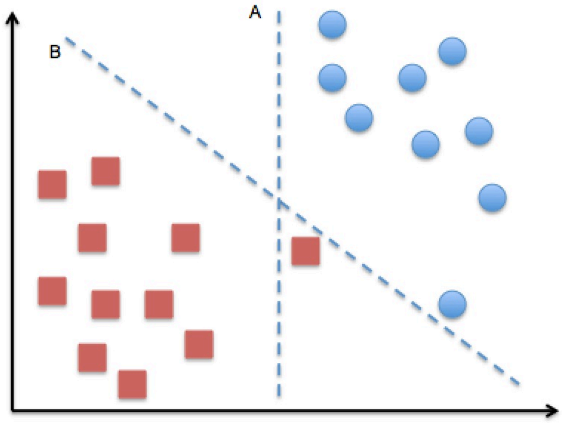


Figure 4

i. Which of the two classifiers (A or B) is most suitable for this purpose ? Justify your answer. [3]

ii. What technique could you use to further improve the classifier you have chosen in question d(i) and how would the improved classifier behave? [2]

iii. What type of classifier would this be? [1]

(e) Why and how is k-fold cross validation performed in data mining? [4]

6. (a) i. Explain what you understand by the term *pseudo attribute* and give an example of its use. [2]

   ii. Give an example of a situation where you would choose to use a pseudo attribute in a classifying task. [1]

   (b) i. Provide a definition of the Gini Index. [1]

   ii. Explain how you would use the Gini Index to decide on which attribute to split on when generating a decision tree. [2]

   (c) Given the set of instances shown in Table 4 for the Play Tennis dataset, and assuming that the attribute Outlook has been chosen as the root node, use the Gini Index to decide on which attribute to split on for the branch Outlook=Sunny. [6]

Outlook	**Temperature**	**Humidity**	**Wind**	**Play Tennis**
Sunny	Hot	High	Weak	No
Sunny	Hot	High	Strong	No
Overcast	Hot	High	Weak	Yes
Rain	Mild	High	Weak	Yes
Rain	Cool	Normal	Weak	Yes
Rain	Cool	Normal	Strong	No
Overcast	Cool	Normal	Strong	Yes
Sunny	Mild	High	Weak	No
Sunny	Cool	Normal	Weak	Yes
Rain	Mild	Normal	Weak	Yes
Sunny	Mild	Normal	Strong	Yes
Overcast	Mild	High	Strong	Yes
Overcast	Hot	Normal	Weak	Yes
Rain	Mild	High	Strong	No

Table 4

   (d) Describe the process of Bagging. [3]

   (e) Two search engines are created to search through 100 documents in a database. Twenty four of these documents relate to housing. When a user types the word 'house' into search engine A it returns fifteen documents that are all relevant. When a user types the word 'house' into a search engine B it returns 30 documents (all 24 of the housing documents and six about other subjects).

   i. What is the Precision and Recall of each search engine? [4]

   ii. Use the F_1 score to explain which search engine is more reliable. [3]

   iii. Derive the formula for a $F_{0.5}$ score. When would you use $F_{0.5}$ instead of F_1? [3]