CS9091

THE UNIVERSITY OF WARWICK

MSc Examinations: Summer 2016

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) Some concept learning algorithms exploit the fact that the hypothesis space is partially ordered. Define the *more general than or equal to* relation between hypotheses in the hypothesis space. [4]

   (b) Consider the set of instances described by the following attributes and values: [5]

   Animal – {Amphibian, Bird, Fish, Mammal, Reptile}
   Size – {Small, Medium, Large}
   Legs – {None, Two, Four, Six, Eight}
   Habitat – {Land, Sea, Air}

   If each hypothesis is described by a conjunction of constraints on the attributes, where each constraint may be "?" (any value is acceptable), $\emptyset$ (no value is acceptable) or a specific value, how many syntactically distinct and semantically distinct hypotheses are there in this hypothesis space?

   (c) Using the CANDIDATE-ELIMINATION algorithm, what are the most general members G and the most specific members S of the hypothesis space after all rows from the training set in Table 1 have been processed in the order given? [8]

Example	Status	Floor	Dept	Size	Bin?
1	Faculty	Four	CS	Medium	Yes
2	Faculty	Four	MS	Medium	Yes
3	Student	Four	CS	Small	No
4	Faculty	Five	CS	Medium	Yes

Table 1

   (d) Two well-known algorithms for concept learning are FIND-S and CANDIDATE-ELIMINATION. Briefly explain how each one searches the hypothesis space and compare their strengths and weaknesses. [8]

Continued

2. (a) Outline Hunt's algorithm for building decision trees from a set of training instances. [4]

   (b) i. Illustrating your answer with an example, define the hypothesis space described by the decision tree hypothesis representation. [2]

   ii. Explain how way the hypothesis space is searched by Hunt's algorithm biases the result. [2]

   (c) i. Give the definition of the Information Gain attribute selection method. [4]

   ii. Under what circumstances might Gain Ratio be a better a choice than Information Gain? Give an example. [3]

   (d) i. A decision tree is to be constructed from the dataset in Table 2, where the root of the tree will be the attribute with the lowest entropy. Which attribute is used for the root? [5]

   ii. Use Information Gain to determine which attribute will be selected as the node for the left-most branch of the root attribute you identified in d) i. [5]

Age	**Astigmatism**	**Tear rate**	**Contact lenses**
Young	No	Normal	Soft
Young	Yes	Reduced	None
Young	Yes	Normal	Hard
Pre-presbyopic	No	Reduced	None
Pre-presbyopic	No	Normal	Soft
Pre-presbyopic	Yes	Normal	Hard
Pre-presbyopic	Yes	Normal	None
Pre-presbyopic	Yes	Normal	None
Presbyopic	No	Reduced	None
Presbyopic	No	Normal	None
Presbyopic	Yes	Reduced	None
Presbyopic	Yes	Normal	Hard

Table 2

Continued

3. (a) Give the definition of the Sigmoid perceptron activation function and explain why it is a popular choice when building artificial neural networks. [4]

   (b) Derive the gradient descent algorithm when using the delta rule to train an artificial neural network using linear units and updating the weights at the end of each epoch. [5]

   (c) i. Explain how the backpropagation algorithm can be used to train artificial neural networks that have one or more hidden units, defining the weight increment terms for output and hidden layers. Assume that each unit is of the sigmoid variety. [7]

   ii. Use the backpropagation algorithm and the stochastic method to calculate the adjusted weights after the training instance $x_1 = 0$, $x_2 = 0$, for the XOR artificial neural network shown in Figure 1. The initial weights and threshold values are: $w_{13} = 0.5$, $w_{14} = 0.9$, $w_{23} = 0.4$, $w_{24} = 1.0$, $w_{35} = -1.2$, $w_{45} = 1.1$; $\theta_3 = 0.8$, $\theta_4 = -0.1$, $\theta_5 = 0.3$ [9]

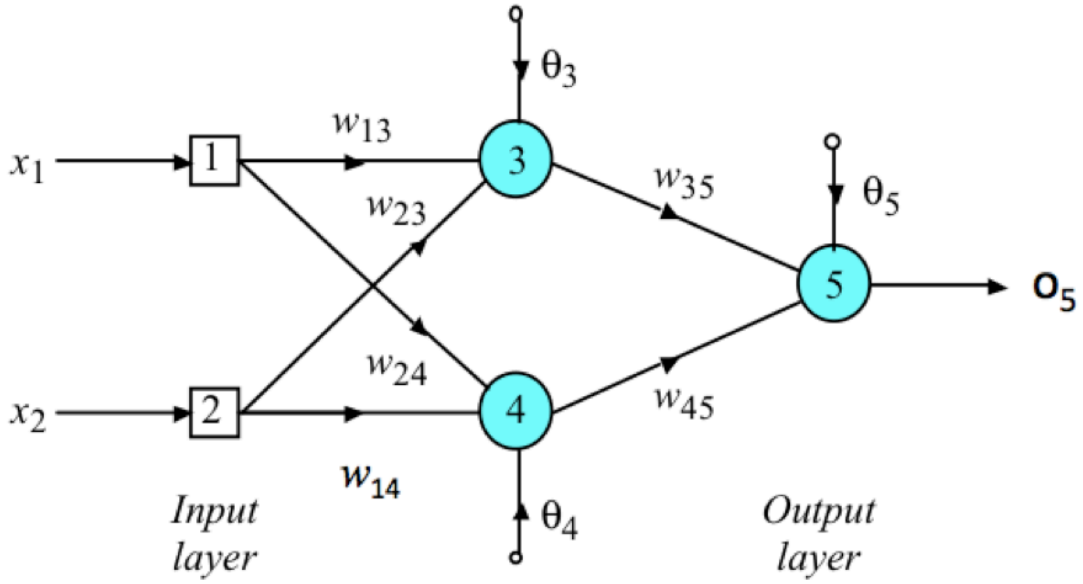


Figure 1

Continued

4. (a) Describe the Expectation-Maximization (E-M) algorithm as used to obtain an estimate of the probability density function of a population from a data sample. You may assume that the underlying probability distribution is a mixture of k Gaussians. How can E-M be used for clustering? [7]

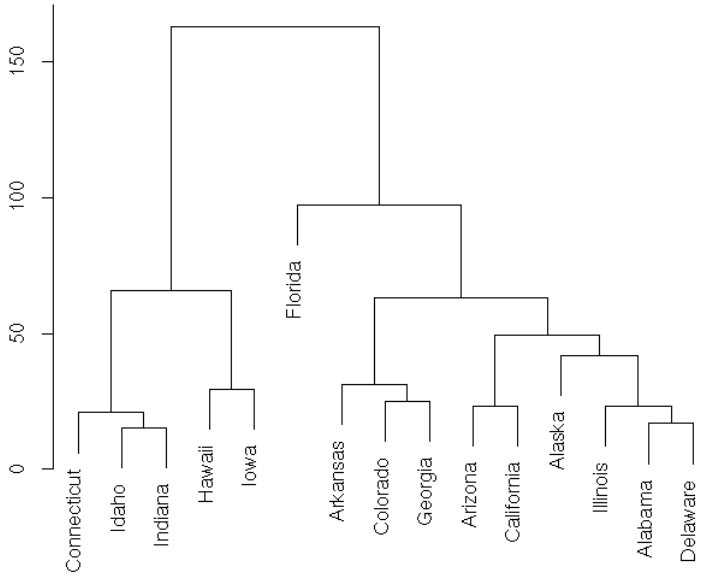
   (b) Using the basic principles underlying E-M and the table below, compute the maximum likelihood estimate of μ after two iterations, given $h = a + b$ and the formulas to compute the expected values of a and b and μ below. Use up to four decimal places in your calculations. The estimates to a, b, μ are given as $\hat{a}$, $\hat{b}$, $\hat{\mu}$ respectively. [5]
   $\hat{a} = \frac{1/2}{1/2+\mu}h$, $\hat{b} = \frac{\mu}{1/2+\mu}h$, $\hat{\mu} = \frac{b+c}{6(b+c+d)}$

Iteration	a	b	c	d	$\hat{\mu}$	$\hat{a}$	$\hat{b}$
1	55	24	40	7			
2							

   (c) Describe how hierarchical agglomerative clustering works. [3]

   (d) Define the single linkage function. [2]

   (e) Given the cluster dendrogram below, what would be a meaningful cutoff in terms of the resulting number of clusters and why? Give your answer considering the possibility of 2, 3, 4, 5 or 6 clusters. [3]



   (f) Define the notions of Support of an item set and Confidence of an association rule. [1]

   (g) Given the transactions in Table 3, which item set from the ones below covers the most transactions? What is the support of this item set? [2]

   (h) According to the transactions in Table 3 what is the confidence of the rule "if beer then nappies" ? [2]

Continued

Transaction	*Items*
1	nappies
2	beer, crisps
3	apples, nappies
4	beer, crisps, nappies
5	apples
6	apples, beer, crisps, nappies
7	apples, crisps
8	crisps

Table 3

Continued

5. (a) Give the definition of TF*IDF and what it is used for [2]

   (b) Below we have a term frequency table for a corpus consisting of only two documents. What is the TF*IDF value of ('example',d_2)? [3]

Document	*this*	*is*	*a*	*sample*	*example*	*another*
d_1	1	1	2	1	0	0
d_2	1	1	0	0	3	2

The 10 base logarithm of selected values are shown below:

x	0.50	0.667	1.00	1.25	1.50	1.75	2.00
$log_{10}(x)$	-0.301	-0.176	0	0.097	0.176	0.243	0.301

   (c) What is a bag-of-words approach to text classification and what are its advantages and disadvantages? [3]

   (d) Describe the main principles underlying topic models. [4]

   (e) Define Precision, Recall and F-measure. What are these measures used for and how do they relate to each other? [4]

   (f) Consider the following confusion matrix which shows the classification of 300 wines by grape variety. The true grape variety is compared against the predicted grape variety by a new experimental chemical analysis technique. [5]

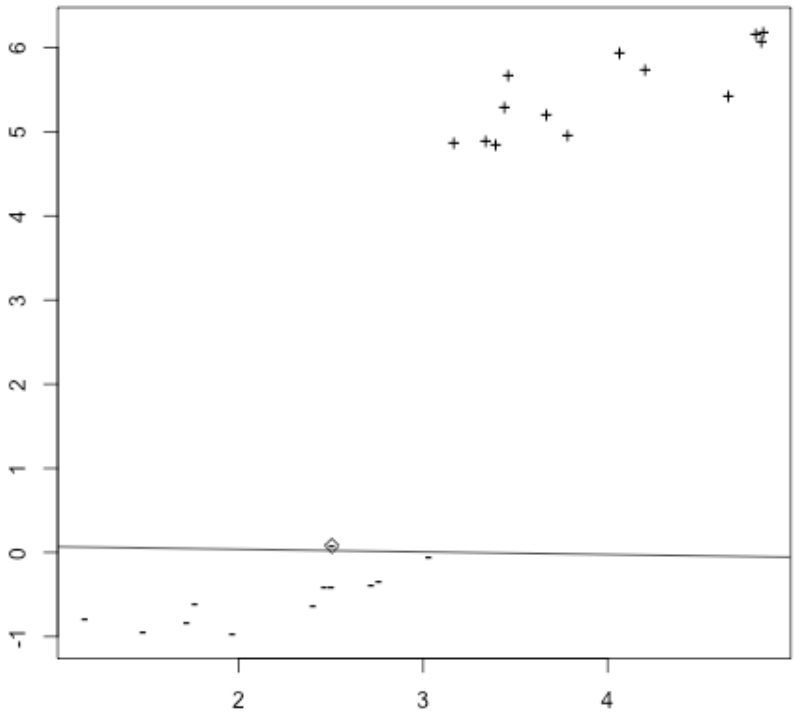
		Predicted		
		Cabernet	Syrah	Pinot
Actual	Cabernet	30	50	20
	Syrah	20	60	20
	Pinot	10	10	80

What are the values for macro-averaged Precision and Recall? What about the micro-averaged values for precision and recall?

   (g) Why and how is k-fold cross validation employed in machine learning? [4]

Continued

6. (a) Consider the perceptron in the image below, which has just misclassified the circled negative instance. [3]



Which TWO of the four states below will NOT be observed as the next step and why?

Continued

Figure 2: State A

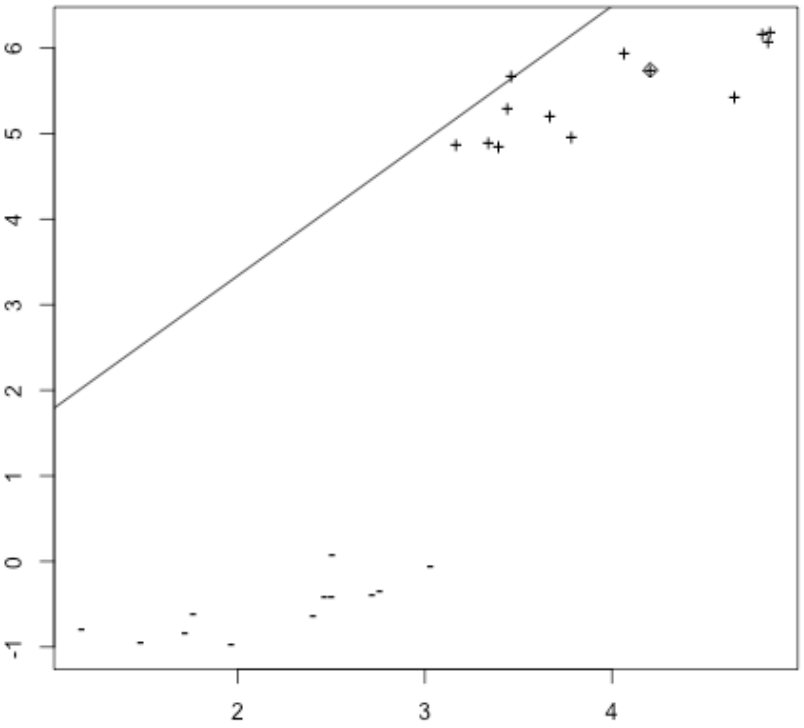
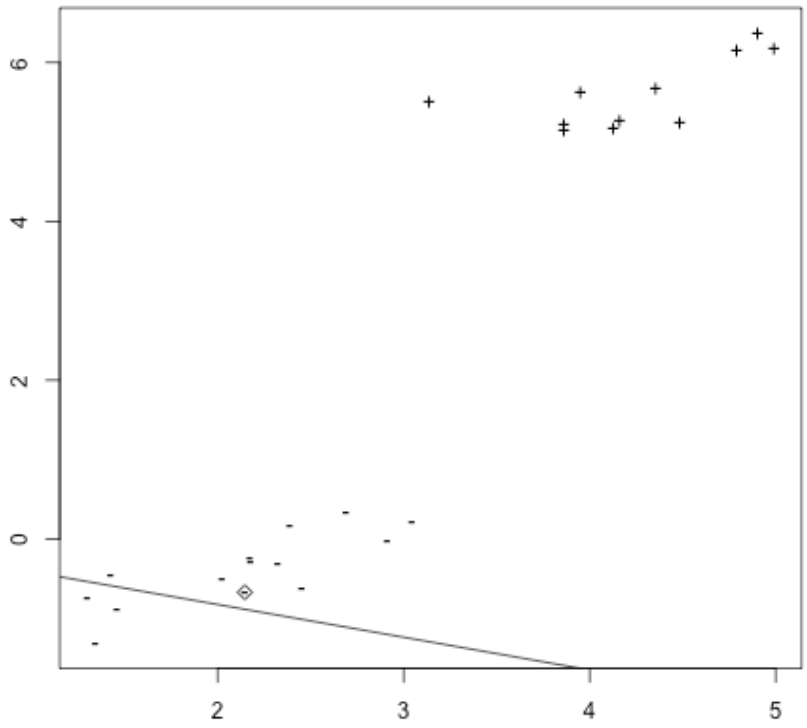


Figure 3: State B



Continued

Figure 4: State C

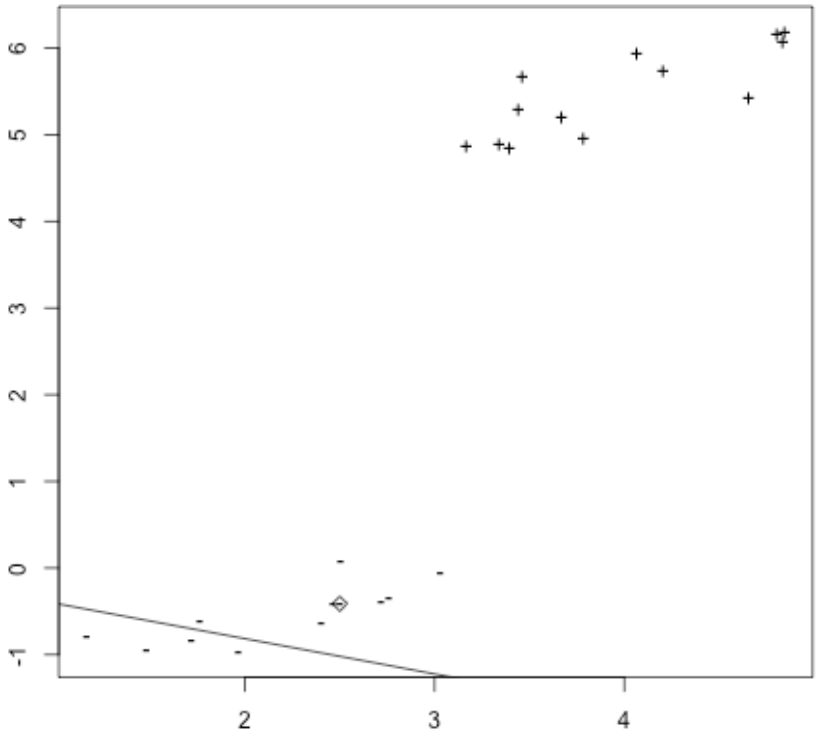
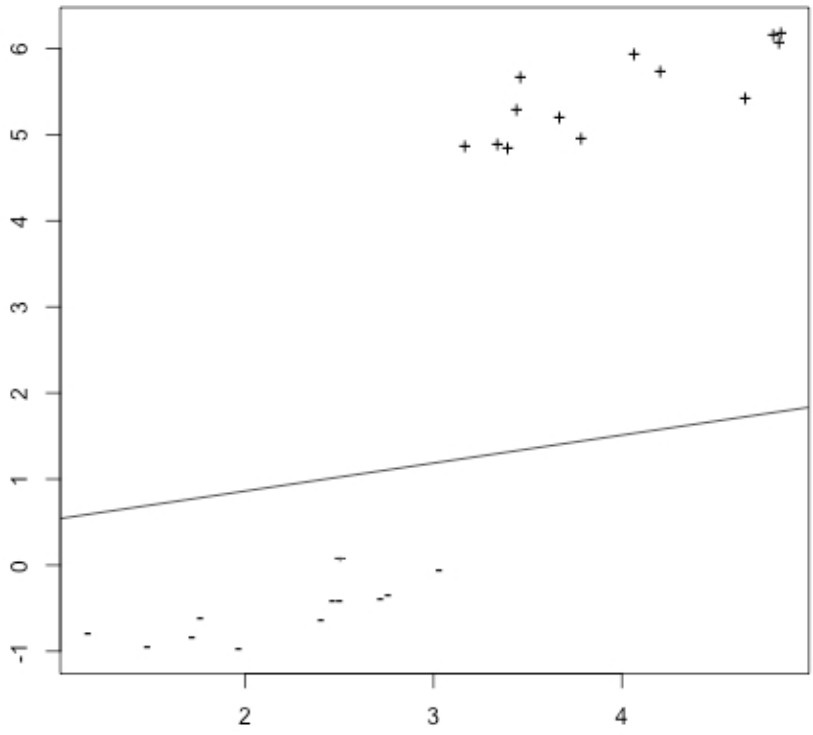


Figure 5: State D



(b) Give the equations for the functional and geometric margin of a perceptron and what they denote. When can the two margins be the same? When are they positive? [4]

(c) What do support vector machines learn? Give the quadratic constrained optimisation [6]

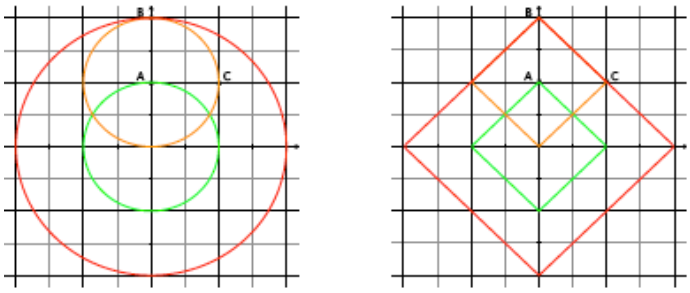
Continued

problem for both the hard margin and soft margin support vector machines. In the case of the soft margin SVM what is the behaviour of the width of the margin as the parameter C decreases?

(d) Explain why Ensemble models in general achieve higher accuracy than the individual base classifier. [2]

(e) Comment on the susceptibility of Ensemble models to overfit the data. [3]

(f) Consider the image below, where you should assume Euclidean distance is used in the image on the left and Manhattan distance is used in the image on the right: What is [3]



the distance between A,B and A,C in the two cases? What is the distance between the origin and each of B and C respectively in the two images? Explain why.

(g) Assume you have used some machine learning algorithm to classify 150 instances with accuracy 92.3%. Calculate the 95% interval for the expected accuracy in the underlying population that the instances are drawn from. [4]

Appendix

$N\%$:	50%	68%	80%	90%	95%	98%	99%
z_N:	0.67	1.00	1.28	1.64	1.96	2.33	2.58

Table 1: N% of area (probability) lies in $\mu \pm z_N \sigma$

The logarithm in base 2 of selected values are shown below.															
x	1/2	1/3	1/4	3/4	1/5	1/8	2/3	2/5	3/5	1/6	3/7	3/8	4/7	7/12	1
$\log_2(x)$	-1.0	-1.58	-2.0	-0.41	-2.32	-3.0	-0.58	-1.32	-0.74	-2.58	-1.22	-1.41	-0.81	-0.78	0