CS3420

THE UNIVERSITY OF WARWICK

Third Year Examinations: Summer 2016

Machine Learning

Time allowed: 2 hours.

Answer **FOUR** questions, **TWO** questions from Section A and **TWO** questions from Section B.

Use one answer book for your Section A answers and a separate answer book for your Section B answers.

Read carefully the instructions on the answer book and make sure that the particulars required are entered on **each** answer book.

Approved calculators are allowed.

Section A Answer **TWO** questions

---

1. (a) Describe the three main subfields of Machine Learning; give examples for each one and describe their differences and similarities. [6]

   (b) Describe two different Loss functions for regression and/or classification given training data $\{t_n, \mathbf{x}_n\}_{n=1}^N$ and a model estimate $\hat{t}_n$. [7]

   (c) Explain what is the Naive Bayes assumption of independence. Give an example where that assumption is violated and explain why. [6]

   (d) Give the mathematical description of the $F1$ score, *Sensitivity* and *Specificity*. What do they measure? What does a value of $F1 = 1$ say for our model? [6]

Continued

2. (a) Describe PCA and give its mathematical relation to an eigenvalue decomposition problem. When is it useful to apply PCA and how would you choose the number of principal components to use? [7]

   (b) What is overfitting and how can we prevent overfitting in decision tree models? [6]

   (c) Give the mathematical description of 3 different distance metrics. What distance is suitable for binary-valued vectors? [6]

   (d) Starting from first principles derive the likelihood function for linear regression assuming white noise. Describe the maximisation problem associated with both a Maximum Likelihood and a Maximum-a-Posteriori approach for this setting. [6]

3. (a) Explain the concept of Leave-One-Out-Cross-Validation (LOOCV). When would you use LOOCV? [4]

   (b) Describe Multi-dimensional Scaling (MDS) and provide a mathematical and intuitive description of *Stress*. Explain the difference between Metric and Non-Metric MDS. [6]

   (c) Give the mathematical description of the Manhattan distance between two vectors $\mathbf{x}_i$ and $\mathbf{x}_j$. What is the relation of the $L1$ norm to the Manhattan distance? [4]

   (d) For a specific probabilistic model the probability for an unseen observation $\mathbf{x}^* \in R^D$ to belong to class k is given by:
   $$P(t^* = k|\mathbf{X}, \mathbf{t}, \mathbf{x}^*) = \frac{p(\mathbf{x}^*|t^* = k, \mathbf{X}, \mathbf{t})P(t^* = k)}{\sum_{j=1}^{K} p(\mathbf{x}^*|t^* = j, \mathbf{X}, \mathbf{t})P(t^* = j)}$$

   i) Is this a generative or a discriminative model? Why? [2]
   ii) Re-write the likelihood of this model under the Naive Bayes assumption. [2]
   iii) What should $P(t^* = 1)$ be equal to if we have a binary classification problem and 90% of the training observations belong to class 0? [2]

   (e) In the context of Bayesian Linear Regression (BLR) the predictive density of interest is given by $p(\mathbf{t}^*|\mathbf{x}^*, \mathbf{X}, \mathbf{t}) = \int p(\mathbf{t}^*|\mathbf{x}^*, \mathbf{w})p(\mathbf{w}|\mathbf{X}, \mathbf{t})d\mathbf{w}$. Apply Bayes rule in the posterior density and describe each resulting term within the context of BLR. What probability density function we typically use for i) the likelihood, and ii) the prior? What type of pdf is the resulting posterior density? [5]

Continued

Section B Answer **TWO** questions

1. (a) The probabilistic output of a binary classifier on a test set of 10 observations is depicted in the left column of Table 1 together with the true labels in the right column. Calculate the Precision and Recall for a decision threshold of 0.5. Show your working. [9]

   (b) What is the range of values the decision threshold can take in order for the classifier to get 0 False Positives on this test set? [7]

   (c) Calculate the point on the ROC curve for the probabilistic classifier using its output in Table 1 and a decision threshold of 0.6. What is the AUC and what information it conveys? [9]

$P(t^* = +1)$	True Class
0.4	-1
0.4	+1
0.6	+1
0.2	-1
0.6	-1
0.1	-1
0.6	+1
0.4	+1
0.5	+1
0.5	-1

Table 1: Your model's probabilities for class 1 assignment versus the true class of the test set.

Continued

2. (a) Given the example set in Table 2, answer the following questions and show your calculations.
   i) What is the entropy of the training set, H(S)? [4]
   ii) What is the information gain of each of the three attributes? [12]
   iii) Which attribute would ID3 choose as the root node? Explain your choice. [2]

Gender	Age Group	Blood Pressure	**Class Variable**
F	A	Low	Yes
M	A	High	Yes
M	C	High	No
F	B	Low	No
F	B	High	Yes
F	A	Low	No
M	B	High	No
M	C	High	Yes
F	C	Low	No
M	A	High	Yes

Table 2: A toy dataset with 10 observations (rows), 3 attributes (Gender, Age, Blood Pressure) and one target class (Class Variable)

   (b) Explain how the Gaussian Mixture Model algorithm works. [7]

Continued

3. (a) Give the pseudocode for stochastic gradient descent with a *linear unit* perceptron. Explain the differences between batch-mode and stochastic gradient descent. [8]

   (b) Give the primal optimisation problem for both the Hard Margin and Soft Margin SVM. Explain the role of the slack variables, and parameter C. How do we set C? [7]

   (c) Given the training examples in Table 3, a fixed learning rate of $\eta = 0.1$, and initial values for the parameters $w_0 = 0.1$, $w_1 = 0.1$, $w_2 = -0.1$, perform the first 3 parameter updates for the standard perceptron with a Heaviside step activation function. Show your working. [10]

x_1 (first attribute)	x_2 (second attribute)	**Class Variable**
0.5	0.5	-1
2	2	1
2	-1	1
1	-1.5	-1
-1	-1	-1

Table 3: A toy dataset with 5 observations (rows), 2 attributes (x_1, x_2) and the target class (Class Variable)

End