CS9090_B MSc
UNIVERSITY OF WARWICK
Standard Examination: Summer 2022
DATA MINING

Time Allowed: 2 hours

Calculators are allowed.

Answer FOUR questions.

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

1) Linear Discriminants and kNN Classification [25 marks]

a) Find the values of the four weights and bias parameters in the linear discriminant function $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + b = w_1x_1 + w_2x_2 + w_3x_3 + w_4x_4 + b = 0$ that allow it to separate the two classes in the dataset below. Verify your answer by calculating the discriminant function score of each example. **[10]**

Index	**Feature-1**	**Feature-2**	**Feature-2**	**Feature-4**	**Label**
1	1	-3	-5	-1	**-1**
2	3	0	2	1	**1**
3	3	0	-5	1	**-1**
4	3	4	-4	0	**1**
5	1	4	0	-5	**-1**

b) Given the following 8 data points (with 4 features each) along with their class labels as training data:

i. Calculate the label that would be generated by a k-Nearest Neighbour Classifier with k=3 for an input example [0.0, 0.0, 0.0, 0.0] with Euclidean distance. **[3] -1**

ii. k=1 1+ 0-
k=2 1+ 1-
k=3 1+ 2-

ii. How many of the "k" nearest neighbours of this example belong to each class? Show all steps involved in the determination of the label for this example. **[4]**

iii. Would the classification boundary generated by this classifier be linear or non-linear? Justify your answer with appropriate reasoning or a proof. **[3]**

Index	Feature-1	Feature-2	Feature-3	Feature-4	Label	
1	-2	4	-5	-4	-1	61
2	0	-1	3	4	-1	26
3	3	1	3	-3	-1	28
4	0	-1	-5	3	1	35
5	0	3	-4	4	1	41
6	3	4	-5	0	1	50
7	-2	4	0	2	1	24
8	-5	2	1	-3	-1	39

c) What is the impact of the choice of distance metric and the value of k in k-nearest neighbour classification on its regularization? **[5]**

k=1 过拟合 k=10欠拟合

2) Structural Risk Minimization and SVM **[25 marks]**

a) What is the relationship between regularisation, margin, empirical error and generalisation of a classification model? Define each term and explain their inter-relationship. **[5]**

b) Which of the following kernels $k(a,b)$ between examples a and b can you choose for use with a support vector machine for solving a classification problem that is not linearly separable? Give your reasoning. **[5]**

i. $k(a,b) = a^T b$

ii. $k(a,b) = (a^T b) + 1$

iii. $k(a,b) = (a^T b)^2 + 1$

iv. $k(a,b) = exp(-\lambda \, \|a - b\|^2)$

c) Calculate the margin for the two linear discriminants each operating over two-dimensional examples $\boldsymbol{x} = [x_1 \quad x_2]^{\boldsymbol{T}}$. Show all steps involved in the calculation. **[5]**

i. $f_1(\boldsymbol{x}) = x_1 + x_2 + 1 = 0$

ii. $f_2(\boldsymbol{x}) = -x_1 - x_2 + 1 = 0$

d) If y and z denote the target and output from a machine learning algorithm for a given example, respectively, which of the following loss functions can be utilised for classification by minimisation of the cumulative loss over a training set? Give reasoning for each case outlining clearly the problem(s) (if any) associated with each. **[10 (2.5 each)]**

i. $l_1(z,y) = \begin{cases} 0 & yz > 0 \\ 1 & yz \leq 0 \end{cases}$

ii. $l_2(z,y) = -y \log(z) \; + (1-y)\log(1-z)$

iii. $l_3(z,y) = (y-z)^3$

iv. $l_4(z,y) = -(y-z)^2$

3) Performance Evaluation [25 marks]

a) Assume a discriminant function $f(\boldsymbol{x})$ which is used to generate a label y for a given example $\boldsymbol{x}$ based on a threshold θ for a binary classification problem such that $y = \begin{cases} +1 & if\ f(\boldsymbol{x}) > \theta \\ -1 & else \end{cases}$. You are given the prediction scores generated by the model for examples in a test set and the corresponding labels in the table below. Use this table to answer the following questions.

Index	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
$\boldsymbol{f(x)}$	1	5	-9	-7	-11	-5	-9	-1	-11	-1
Label	1	-1	-1	-1	1	-1	1	-1	1	-1

i. Calculate the sensitivity, specificity and positive predictive value of the model at a threshold of $\theta = 0$. **[4]**

ii. What is the expected impact (in terms of increase or decrease) of decreasing the threshold to $\theta = -2$ on sensitivity, specificity and precision? Give reasoning for each metric. **[4]**

iii. Give reasoning as to why this predictor is or is not good for classification over this test set. Is it any better than a random predictor? **[4] not a good threshold to divide f(x) into + and -.**

iv. What metric would you use for comparing this predictor to another one over the same test set? Give reasoning behind your choice. **[4]**

v. Would the area under the receiver operating characteristic curve for this classifier be greater than 0.5 or less than 0.5? Give your reasoning. **[4]**

b) Assume that you want to develop a predictor for diagnosing a rare disease using imaging data of the general population. What performance metrics would be best suited for comparing two different predictors for this problem and why? **[5]**

4) PCA and Regression [25 marks]

a) For the two dimensional data points in the plot given below:

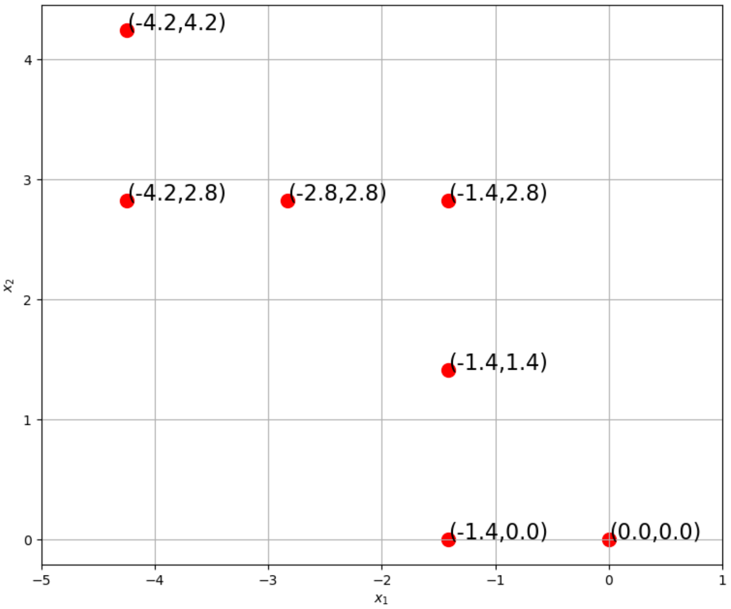
i. Write the first principal component as a unit vector. **[3]**

ii. Write the second principal component as a unit vector. **[3]**

iii. Project the data onto the first principal component and write the corresponding projection values for each point. **[3]**

iv. What is the percentage of variance captured along the first principal component for this data? Show all steps used in this calculation. **[3]**

v. Would it be possible to reconstruct the 2-D data points from their projection onto the first principal component without any loss of information? Give reasoning for all answers. **[3]**



b) For a given feature vector x with associated target value y, write the regression function $f(x)$ and loss function $l(f(x), y)$ used in least squares regression. What optimizer would you use for minimization of this loss? **[5]**

c) Suppose you want to regress between image data of a hurricane captured from a satellite and a hurricane's intensity measured by an aircraft flying through the hurricane. Given that the sensors mounted on the plane may experience a lot of vibration and noise and the satellite imagery may also contain noise, what would be your choice of the regressor for solving this problem out of the options given below? Give appropriate reasoning. **[5]**

i. Linear Least Squares Regression

ii. Support Vector Regression with epsilon-insensitive loss with epsilon = 0

iii. Support Vector Regression with epsilon-insensitive loss with epsilon > 0

5) Neural Networks **[25 marks]**

a) Which of the following activation functions $u(z)$ over net input z would you pick for solving a classification problem that is not linearly separable? Give reasoning along with any possible limitations of other options. **[5]**

i. $u(z) = z$

ii. $u(z) = 2z + 1$

iii. $u(z) = \frac{2tan^{-1}(x)}{\pi}$

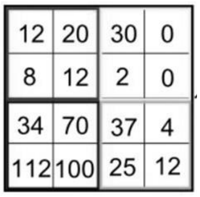
iv. $u(z) = z^2$

b) Which of the two activation functions $u(z)$ for net input z would result in faster convergence and why? **[5]**

i. $u(z) = \frac{1}{1+e^{-z}}$

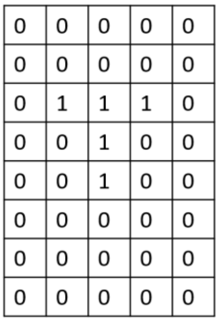
ii. $u(z) = max(\alpha(e^z - 1), z), \alpha \ll 1$

c) What would the effect of applying max pooling with a stride of 2 on the following output? **[5]**



12	20	30	0
8	12	2	0
34	70	37	4
112	100	25	12

d) Assume you are using a convolutional neural network with a single 3 × 3 sized filter to detect T-shaped objects in images as shown below (marked by 1). What would you expect the filter to look like after training? Write the 3 × 3 filter and give appropriate reasoning/proof. **[5]**



0	0	0	0	0
0	0	0	0	0
0	1	1	1	0
0	0	1	0	0
0	0	1	0	0
0	0	0	0	0
0	0	0	0	0
0	0	0	0	0

e) Which one of the following networks would you expect to converge faster for a given image classification task and why? Assume that the number of trainable parameters are the same for all networks. **[5]**

i. Fully connected multi-layered perceptrons

ii. Convolutional neural networks with residual connections

iii. Convolutional neural networks without residual connections

-END-