*Sample paper*

*ECS708P  Machine Learning        Duration:  2 hours*

This is an open-book exam, which should be completed in approximately 2 hours.

You can refer to textbooks, notes and online materials to facilitate your working, but normal referencing and plagiarism rules apply and you must cite any sources used.

Calculators are permitted in this examination.

---

## Answer FOUR questions

**You MUST adhere to the word limits, where specified in the questions. Failing to do so will lead to those answers not being marked.**

---

**YOU MUST COMPLETE THE EXAM ON YOUR OWN, WITHOUT CONSULTING OTHERS**

**Examiners: Dr. Jesus Requena Carrion, Dr. Lin Wang**

**Question 1**

a) In this question we explore regression in a problem involving two attributes $x$ and $y$, where $x$ is the predictor feature and $y$ is the prediction. The dataset that we will use for training contains four samples and is shown in Table 1.

| $x$ | $y$ |
|---|---|
| 2 | $1 + 0.1 \times D_1$ |
| 4 | $5 + 0.1 \times D_2$ |
| 1 | $2 + 0.1 \times D_3$ |
| 3 | $2 + 0.1 \times D_4$ |

**Table 1**

In Table 1, $D_1$, $D_2$, $D_3$ and $D_4$ represent the last four digits of your student ID ($D_1$ being the last, $D_2$ the second last, etc). Before continuing, calculate the numerical value of the predictor $y$ for each sample (for instance, if $D_1 = 1$, then $1 + 0.1 \times D_1 = 1.1$).

The coefficients $w$ of the Minimum Mean Square Error (MMSE) solution of a simple linear model can be obtained as

$$w = (X^T X)^{-1} X^T y$$

where $X$ is the design matrix and $y$ is the prediction vector.

   i) Obtain the MMSE coefficients of the simple linear model $y = w_0 + w_1 x$. You can use the following intermediate result:

$$(X^T X)^{-1} = \begin{bmatrix} 1.5 & -0.5 \\ -0.5 & 0.2 \end{bmatrix}$$

   ii) Calculate the training Mean Square Error (MSE) of the MMSE solution (*Use $w_0 = 0$ and $w_1 = 1$ if you did not obtain the MMSE solution*).

**[15 marks]**

b) Consider the cubic model $y = w_0 + w_1 x + w_2 x^2 + w_3 x^3$ for the dataset in Table 1.

   i) What would you expect the training MSE of this model to be?
   ii) Assuming that the true model is $y = x + n$, where $n$ is zero-mean Gaussian noise, identify the main sources of error in the prediction of this cubic model during deployment.

**[10 marks]**

**Question 2**

a) In this question we explore classification in a problem involving two predictor features $x_A$ and $x_B$, and two classes, namely $\bigcirc$ (positive class) and $\times$ (negative class). The dataset that will be used in this question is shown in Figure 1.
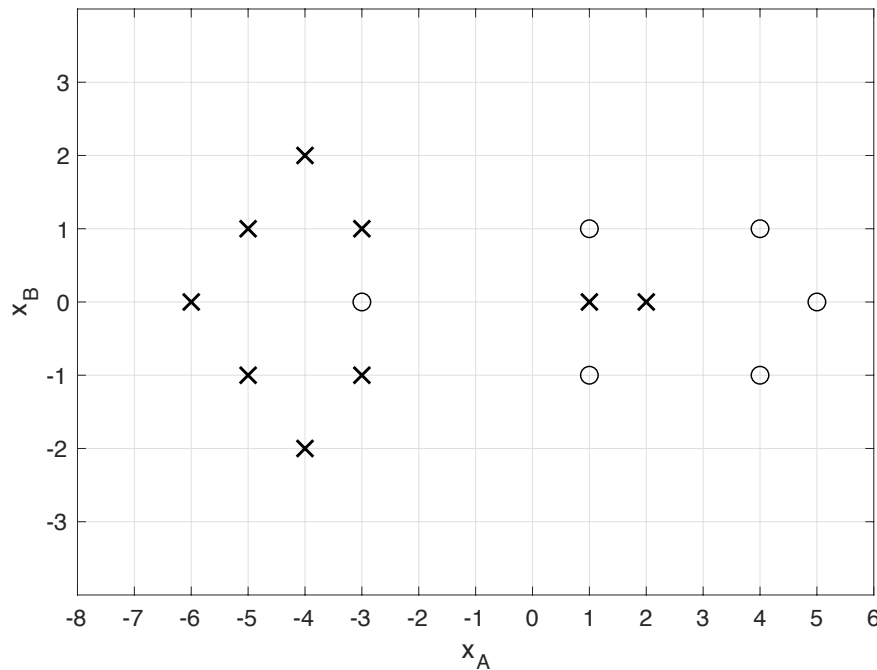


**Figure 1**

Consider a family of linear classifiers defined by $w^T x = 0$, where $w = [w_0, w_A, w_B]^T$ and $x = [1, x_A, x_B]^T$. A sample $x_i$ such that $w^T x_i > 0$ will be labelled as $\bigcirc$, otherwise it will be labelled as $\times$. Given the linear classifier defined by the coefficients $w_B = 0$, $w_A = 1$ and $w_0 = 0.25 \times D$, where $D$ is the last digit of your student ID:

   i)   Obtain the classifier's decision regions.
   ii)  Obtain the classifier's confusion matrix for the dataset shown in Figure 1 and identify its sensitivity and specificity.

**[10 marks]**

b) We now want to build a Linear Discriminant Analysis classifier that uses as predictor feature $x_A$.
   i)   Obtain the priors for each class, namely P($\bigcirc$) and P($\times$), and the means of the distributions P($x_A$ | $\bigcirc$) and P($x_A$ | $\times$).
   ii)  Describe the corresponding Bayes classifier.
   iii) If the standard deviations of P($x_A$ | $\bigcirc$) and P($x_A$ | $\times$) are equal, how would a sample such that $x_A$ = -0.5 be classified?

**[15 marks]**

**Question 3**

a) This question concerns model optimisation in machine learning. The k-means algorithm is said to converge to local minima, rather than to the global minimum.

   (i)   Use the notion of error function to explain the concepts of local minimum and global minimum.
   (ii)  Explain what is meant by the statement *the k-means algorithm converges to a local minimum*.
   (iii) Considering the risk of converging to a local minimum, design a strategy that would improve the solution provided by the k-means algorithm.

**[15 marks]**

b) The validation-set approach allows one to evaluate the performance of different models during model selection.

   (i)   After applying a validation-set approach, the validation errors of two models $f_1$ and $f_2$ are found to be respectively $E_1 = 10$ and $E_2 = 12$. How would you use this result to inform your selection of either $f_1$ or $f_2$?
   (ii)  Due to the low number of samples in the available dataset, it is suggested that the whole dataset should be used for training models $f_1$ and $f_2$ and both models should be compared based on their training errors. What is your view on this suggestion?

**[10 marks]**

**Question 4**

a) This question concerns neural networks.
  i) Which layers offer greater flexibility, fully-connected layers or convolutional layers?
  ii) Why are convolutional networks suitable for time series and image data?
  iii) The number of feature maps in convolutional architectures usually increase as we move closer to the output layer. What is the idea behind this design?

**[15 marks]**

b) Consider a dataset consisting of grayscale images of size 100 x 100 pixels and a binary label. A deep neural network combining convolutional, pooling and fully-connected layers is chosen for building a classifier for this dataset.

  i) The first hidden layer is a convolutional one and consists of two 100 x 100 feature maps. Each map is obtained by applying a different filter of dimensions 3 x 3. How many parameters need to be trained in the first layer?
  ii) The second layer is a 2x2 max-pooling layer. How many feature maps does this layer have and what are their dimensions?
  iii) The third hidden layer is also convolutional and consists of 8 feature maps defined by filters of dimensions 3 x 3 x $D$. What is the value of $D$?

**[10 marks]**

**End of Paper**