UNIVERSITY OF SOUTHAMPTON COMP6245W1

SEMESTER 1 FINAL ASSESSMENT 2020 - 2021

MACHINE LEARNING

DURATION 240 MINS (4 Hours)

This paper contains 20 questions

Provide answers to **all** questions in a single page, neatly numbered in order. You may attach two further pages with any workings where useful, clearly numbered and in order of the questions. These additional pages will be looked at if the question requires any derivation and the answer you provided is incorrect.
Note the questions will have at least one correct answer. Where the question has more than one correct answer, you must select all the correct ones. For these questions, partial credit will usually not be given.
You should upload a maximum of three pages as a single pdf file.
Each question is worth *ive marks*.

Question 1.

In a far away island, a highly infectious disease is spreading across the population. A third of those infected appear to suffer long term illness, whereas two thirds recover. The precise reasons as to who suffers adverse conditions is unknown. Scientists claim to have discovered two proteins, concentrations of which in blood could be implicated in the adverse conditions. Measurements of concentrations of these proteins (P_1 and P_2) were carried out in samples of patients who suffered long term conditions (denoted A) and those who made full recovery (B). Bivariate Gaussian models were fitted to the data ($\boldsymbol{x} = [P_1\ P_2]^T$), and estimated means and covariances of A and B were as follows:

$$\boldsymbol{m}_A = \begin{bmatrix} 1.3 \\ 4.3 \end{bmatrix},\ \boldsymbol{m}_B = \begin{bmatrix} 8.5 \\ 4.7 \end{bmatrix},$$

$$\boldsymbol{\Sigma}_A = \begin{bmatrix} 3.0 & 0.001 \\ 0.001 & 1.5 \end{bmatrix} \text{ and } \boldsymbol{\Sigma}_B = \begin{bmatrix} 3.0 & 0.001 \\ 0.001 & 1.5 \end{bmatrix}.$$

A multi-national company contracted by the government of the island recommends the use a *linear classifier*: $f(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0$ to predict adverse outcomes. You are asked to comment on the proposed classifier.

Which of the following statements is/are true?

1. The proposed linear classifier is optimal and should be deployed.
2. A distance-to-mean classifier based on Euclidean distance will be the optimal solution.
3. Inspired by the brain, we should train an artificial neural network.
4. A distance-to-mean classifier based on the Mahalanobis distance will be the optimal solution.
5. The implication of protein P_1 with this condition is suspect.
6. The implication of protein P_2 with this condition is suspect.

[5 marks]

Question 2.

Consider the scenario described in Question 1. The study was repeated with refined measurements of the two proteins, producing the following results:

$$\boldsymbol{m}_A = \begin{bmatrix} 2.35 \\ 4.76 \end{bmatrix}, \ \boldsymbol{m}_B = \begin{bmatrix} 2.42 \\ 4.82 \end{bmatrix},$$

$$\boldsymbol{\Sigma}_A = \begin{bmatrix} 2.0 & 1.0 \\ 1.0 & 2.0 \end{bmatrix} \text{ and } \boldsymbol{\Sigma}_B = \begin{bmatrix} 2.0 & 0.001 \\ 0.001 & 2.0 \end{bmatrix}.$$

What might you suspect?

1. A third protein might be involved in causing long term illness.
2. We could still consider the use of an artificial neural network.
3. A linear support vector machine that maximizes the margin is a better solution.
4. Measuring protein P_1 is sufficient for accurate prediction.
5. The linear classifier recommended by the company contracted by the government is the optimal solution to this problem.

[5 marks]

TURN OVER

Question 3.

Consider again the scenario described in Question 1. After realising their error, the government of the island decides to consult expert clinicians and data scientists of Dolphin University who base their study on x-ray imaging data of affected organs. Abnormal regions of the images were annotated by the clinical experts and prediction systems were designed by the data scientists. Interpretation of the images being time consuming, and the clinical experts paid much higher salaries than the data scientists, only part of the data (set $\mathcal{A}$) was annotated. We denote the remaining set $\mathcal{B}$.

Six features were extracted from each image by the data scientist, formulating a regression problem in $\boldsymbol{x} \in \mathcal{R}^6$, and predicting how long a seriously infected patient might survive in intensive care conditions.

A radial basis function model $f(\boldsymbol{x}) = \sum_{j=1}^{M} \lambda_j \, \phi\left(\alpha||\boldsymbol{x} - \boldsymbol{m}_j||\right)$ was proposed by the data scientist who suggested that the data in set $\mathcal{B}$ be clustered using $K-$means clustering to set $\boldsymbol{m}_j, \;\; j = 1, 2, ..., M$ and data in set $\mathcal{A}$ be used to solve a regression problem to estimate the $\lambda_j, j = 1, 2, ..., M$.

The approach used by the data scientist is best described as:

1. Supervised learning
2. Unsupervised learning
3. Semi supervised learning
4. Transfer learning
5. Deep learning
6. Online learning
7. Self supervised learning

[5 marks]

Question 4.

The textbook Pattern Recognition and Machine Learning gives expressions for the Bayesian estimation of the mean of a univariate Gaussian density in Equations 2.141 and 2.142. Please refer to these equations before answering the question below.

Which of the following statements is/are true?

1. When $N \rightarrow \infty$, the Bayesian and maximum likelihood estimates are the same.
2. A high confidence prior has large σ_0.
3. With a high confidence prior, the Bayesian and maximum likelihood estimates, using the same amount of data, will be identical.
4. Uncertainty in the Bayesian estimates reduces with sample size.

[5 marks]

TURN OVER

Question 5.

Consider the derivation of equation (2.126) for the maximum likelihood estimation of the mean in the textbook Pattern Recognition and Machine Learning:

$$\boldsymbol{\mu}_{\mathrm{ML}}^{(N)} = \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)} + \frac{1}{N}(\boldsymbol{x}_N - \boldsymbol{\mu}_{\mathrm{ML}}^{(N-1)}).$$

Which of the following statements is/are true?

1. This formula is useful for accurate estimation of the mean of a multivariate Gaussian distribution.
2. This formula is useful for solving an online learning problem.
3. We might use this formula in a situation where the size of a given dataset is very large.
4. The quantity computed by this formula can sometimes not converge to the true mean unless the learning rate is set very low.

[5 marks]

Question 6.

Consider the univariate function

$$y = \exp\left\{-\frac{1}{2}(x - 0.2)^2\right\}$$

What is $\int_{-\infty}^{+\infty} y\, dx$?

1. $\sqrt{\pi}$
2. $\sqrt{2\pi}$
3. $0.2\sqrt{2\pi}$
4. 9.8
5. 6.0×10^{23}

[5 marks]

TURN OVER

Question 7.

You are tasked with predicting the market price of an asset using its past values and several variables relating to the underlying economy within which the business operates. The dataset given to you spans one year of daily trading (252 items) of 180 variables. You are required to split the data into a training set and an evaluation set of equal sizes and use a linear model as predictor. You attempt to solve the problem of estimating regression coefficients by

$$\boldsymbol{w} = \left(X^T X\right)^{-1} X^T \boldsymbol{t}$$

Which of the following is/are true?

1. The attempt above will not work without suitable regularization.
2. I would advocate the use of a *Lasso* regularizer to solve the problem.
3. If the problem is solved using a regularizer

   $$\min_{\boldsymbol{w}} ||\boldsymbol{t} - X\,\boldsymbol{w}|| + \gamma||\boldsymbol{w}||^2,$$

   setting γ to very small values will produce sparse solutions.
4. We cannot use more data acquired by taking longer windows (say several years of trading instead of just one) because the underlying statistical relationships might have changed over time.

[5 marks]

Question 8.

With usual notation, Fisher Linear Discriminant Analysis (FLDA) maximizes the objective function

$$J(\boldsymbol{w}) = \frac{\boldsymbol{w}^T S_B \boldsymbol{w}}{\boldsymbol{w}^T S_W \boldsymbol{w}}$$

to arrive at the discriminant direction $\boldsymbol{w}_F = \beta S_W^{-1}(\boldsymbol{\mu}_2 - \boldsymbol{\mu}_1)$. Which of the following statements is/are true?

1. If the features are uncorrelated, $\boldsymbol{w}_F$ is the same as the line joining the means.
2. It is not necessary to compute the term β in the solution accurately.
3. Computing $\boldsymbol{w}_F$ is a necessary step in deriving the Receiver Operating characteristic (ROC) curve for any pattern classification problem.
4. If the class conditional densities are multi-modal, using FLDA is not recommended.
5. The factor β could be tuned to improve regularization.

[5 marks]

TURN OVER

Question 9.

In solving a two class pattern classification problem, it is thought Fisher LDA could be improved by accounting for prior probabilities of classes, $p(C_1)$ and $p(C_2)$. The corresponding objective function to maximize is:

$$J(\boldsymbol{w}) = \frac{(\mu_1 - \mu_2)^2}{p(C_1)s_1^2 + p(C_2)s_2^2},$$

where μ_1 and μ_2 are projected mean and s_1 and s_2 are scatters of projected data.

Derive the direction that maximize $J(\boldsymbol{w})$.

Your answer is:

1. $\boldsymbol{w} = \beta\,(\boldsymbol{\Sigma_1} + \boldsymbol{\Sigma_2})^{-1}(\boldsymbol{\mu_2} - \boldsymbol{\mu_1})$
2. $\boldsymbol{w} = \beta\,\frac{p(C_1)}{p(C_2)}(\boldsymbol{\Sigma_1} + \boldsymbol{\Sigma_2})^{-1}(\boldsymbol{\mu_2} - \boldsymbol{\mu_1})$
3. $\boldsymbol{w} = \beta\,(p(C_1)\boldsymbol{\Sigma}_1 + p(C_2)\boldsymbol{\Sigma_2})^{-1}(\boldsymbol{\mu_2} - \boldsymbol{\mu_1})$
4. $\boldsymbol{w} = \sqrt{2\pi}\,\beta\,(\exp(p(C_1))\boldsymbol{\Sigma}_1 + \exp(p(C_2))\boldsymbol{\Sigma_2})^{-1}(\boldsymbol{\mu_2} - \boldsymbol{\mu_1})$

$\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma_1}, \boldsymbol{\Sigma_2}$ are the means and covariance matrices of the two classes.

[5 marks]

Question 10.

Dolphin University scientists have developed a novel method to predict coronavirus infection based on traces of mobile phone usage. A continuous valued score is computed from the duration of contact with persons known to have tested positive. A threshold is set and if the score exceeds this threshold, the person concerned is requested to self-isolate.

In the above setting, which of the following is/are true?

1. There is an economic cost associated with *False Positive* predictions.
2. High *False Negatives* lead to infection risk in the community.
3. *True Positives* of the test are caused by the test themselves.
4. Inspired by how the brain works, I will input the score to an artificial neural network for accurately predicting coronavirus infection.

[5 marks]

TURN OVER

Question 11.

Which of the following statements is/are true about a Receiver Operating Characteristic (ROC) curve?

1. The area under the curve can sometimes be negative.
2. The probability of correct ranking is given by the area under the ROC curve.
3. Every operating point on the ROC curve yields the same misclassification error.
4. It is not advisable to use area under the ROC curve as a performance measure if we can estimate the different costs of misclassification.
5. Increasing the learning rate when training a neural network always increases the area under the corresponding ROC curve.

[5 marks]

Question 12.

In a two-class pattern classification problem involving a positive-valued univariate feature x, the class conditional densities are both uniformly distributed as follows:

$$p(C_1|x) = \begin{cases} \alpha & a \leq x \leq b \\ 0 & \text{otherwise} \end{cases} \text{ and } p(C_2|x) = \begin{cases} \beta & c \leq x \leq d \\ 0 & \text{otherwise} \end{cases},$$

where $a \leq c \leq b$.

Compute the area under the Receiver Operating Characteristics (ROC) curve for this problem, assuming the prior probabilities of the classes, $p(C_1)$ and $p(C_2)$ are equal.

Your answer is:

1. $1 - \frac{(b-c)^2}{2(b-a)(d-c))}$
2. $1 - (b-a)(d-c)$
3. $1 - (d-b)[1 - \frac{1}{2}\frac{(d-a)}{(b-c)}]$
4. $0.5[1 + (c-a)(b-c)^2(d-b)]$

[5 marks]

TURN OVER

Question 13.

A dataset consists of shopping habits of $N = 300$ individuals. The number of times any individual purchased any of $p = 600$ items in the two weeks prior to Christmas has been recorded in the dataset. The data is contained in a matrix X of dimensions $N \times p$. The purchasing power of the individuals was also acquired from their annual tax returns and is contained in an $N-$dimensional vector $\boldsymbol{y}$.

The following analysis was performed on this data:

$$\min_{W,H} ||X - W\,H||^2 \quad \text{subject to } w_{ij} \geq 0,\ h_{ij} \geq 0,$$

where matrices W and H are of dimension $N \times r$ and $r \times p$ respectively, and w_{ij} and h_{ij} denote their elements. We also chose r such that $r < N$.

A linear prediction of the purchasing power from the items purchased was also attempted.

Which of the following statements is/are true?

1. The rank of the reconstruction $W\,H$ is p.
2. The rank of the reconstruction $W\,H$ is at most r.
3. W defines a dimensionality reduction of features that preserves the variance in the original features.
4. W defines a dimensionality reduction of features leading to a sparse set of features.
5. Predicting the purchasing power from features given in W is preferable to predicting it directly from the original features given in X.

[5 marks]

Question 14.

The distribution of two two-dimensional variables $\boldsymbol{x}$ and $\boldsymbol{y}$ are shown as scatter plots in Fig. 1.

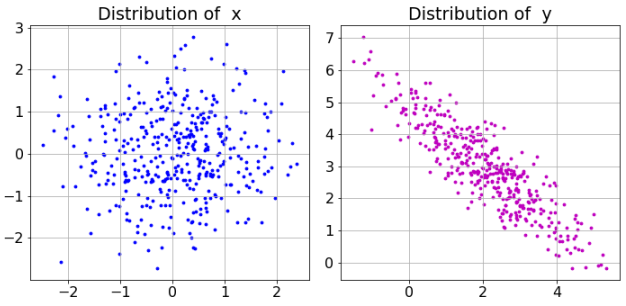

FIGURE 1: Distribution of two two-dimensional variables

Which of the following statements is/are true?

1. Variable $\boldsymbol{y}$ could have been derived from variable $\boldsymbol{x}$ by a linear transformation of the form $\boldsymbol{y} = A\,\boldsymbol{x} + \boldsymbol{b}$.
2. Variable $\boldsymbol{y}$ coud have been derived from variable $\boldsymbol{x}$ by a linear transformation of the form $\boldsymbol{y} = A\,\boldsymbol{x}$.
3. Variable $\boldsymbol{x}$ is likely to have a covariance matrix $\begin{bmatrix} 2.0 & 0.0 \\ 0.0 & 0.0 \end{bmatrix}$.
4. Variable $\boldsymbol{y}$ is likely to have a covariance matrix $\begin{bmatrix} 2.0 & -1.8 \\ -1.8 & 2.0 \end{bmatrix}$
5. Normalization has been applied to variable $\boldsymbol{y}$ so that it has zero mean.

[5 marks]

TURN OVER

Question 15.

A multi-layer perceptron (MLP) is usually trained using gradient descent, with the gradient computed using the error backpropagation algorithm. Which of the following statements is/are true?

1. Given the data in the form of an $N \times p$ matrix X, where N is the number of data items and p, the input dimensions, and the targets in vector $\boldsymbol{t}$, the weights $\boldsymbol{w}$ could be solved by the formula:
$$\boldsymbol{w} = (X^T X)^{-1} X^T \boldsymbol{t}$$
2. The error function of the MLP is quadratic.
3. The speed of convergence of a gradient descent algorithm of the form
$$\boldsymbol{w} \leftarrow \boldsymbol{w} - \alpha \nabla_{\boldsymbol{w}} E.$$
could be increased by cross validation.
4. The use of a *momentum* term usually helps improve speed of convergence.

[5 marks]

Question 16.

Given a classification problem $\{\boldsymbol{x}_n,\, t_n\}_{n=1}^{N}$ the perceptron learning algorithm updates weights using the formula

$$\boldsymbol{w}^{(\tau)} \;=\; \boldsymbol{w}^{(\tau-1)} \;+\; \eta\, t_n\, \boldsymbol{x}_n.$$

Which of the following statements is/are true?

1. $\boldsymbol{x}_n$ is an item of data correctly classified by the current estimate of weights $\boldsymbol{w}^{(n-1)}$.
2. $\boldsymbol{x}_n$ is an item of data misclassified by the current estimate of weights $\boldsymbol{w}^{(n-1)}$.
3. The solution to which the algorithm converges could be written as $\sum_{n=1}^{N}\; \alpha_n\, \boldsymbol{x}_n$, *i.e.* a weighted combination over all the data.
4. If the data is linearly separable, the iterative algorithm is guaranteed to terminate.
5. The learning rate η should be set by cross validation.
6. The above algorithm minimizes the following cost function:

$$E(\boldsymbol{w}) \;=\; \sum_{n=1}^{N} (t_n \;-\; \boldsymbol{w}^T\, \boldsymbol{x}_n)^2$$

[5 marks]

TURN OVER

Question 17.

Two groups of people are sitting in a park. Group A consists of 4 members, whereas Group B consists of 10 members. The positions of all group members are shown in Fig. 2. The k-means algorithm is applied to the data for clustering. A Gaussian Mixture Model (GMM) is also fitted to the data. The initial centroids of k-means and of the GMM are randomly selected from the samples. Both algorithms are run for up to 100 iterations. Since the algorithms are randomly initialised, k-means and the GMM are evaluated over 50 independent trials.

Which of the following statements is/are true?

1. For each trial, the cluster centers and the cluster assignments are identical between k-means and the GMM.
2. For each trial, k-means assigns all points to the same cluster.
3. For some trials, the cluster centers obtained using the GMM may be skewed towards the mean of all samples.

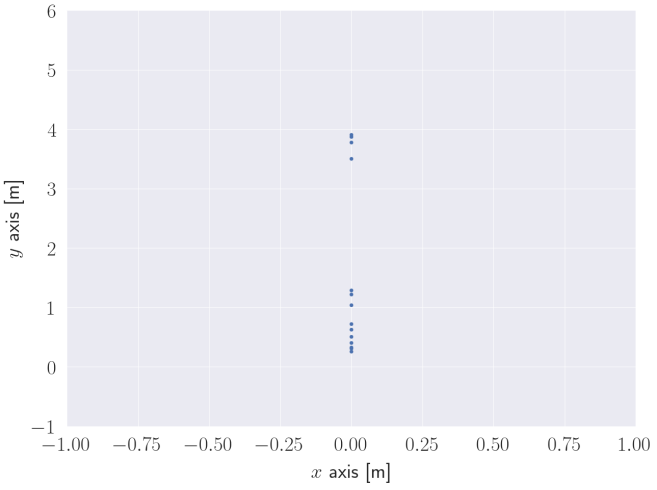

FIGURE 2: Distribution of two two-dimensional variables

[5 marks]

Question 18.

A researcher is deriving the maximisation of the log-likelihood function for Gaussian Mixture Models. The log-likelihood function is given by:

$$\ln p_{\boldsymbol{\theta}}\left(\mathbf{X}\right) = \sum_{n=1}^{N} \ln \left[\sum_{k=1}^{K} \pi_k \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right) \right], \tag{1}$$

where $\mathcal{N}(\cdot)$ denotes the probability density function of a Gaussian; the dataset is given by $\mathbf{X} = \left[\mathbf{x}_1, \ldots, \mathbf{x}_N\right]$ and sample, $n \in \{1, \ldots, N\}$, is denoted by $\mathbf{x}_n$; the number of samples is N; the number of Gaussian mixture components is K; π_k, $\boldsymbol{\mu}_k$ and $\boldsymbol{\Sigma}_k$ correspond, respectively, to the weight, mean, and covariance of component, $k \in \{1, \ldots, K\}$; and $\theta = \{(\boldsymbol{\mu}_1, \boldsymbol{\Sigma}_1, \pi_1), \ldots, (\boldsymbol{\mu}_K, \boldsymbol{\Sigma}_K, \pi_K)\}$.

Which of the following statements is/are true?

1. $\frac{\partial}{\partial \boldsymbol{\mu}_k} \ln p_{\boldsymbol{\theta}}\left(\mathbf{X}\right) = \sum_{n=1}^{N} \frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \cdot \frac{\partial}{\partial \boldsymbol{\mu}_k} \left[\ln \mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right]$
2. $\frac{\partial}{\partial \boldsymbol{\mu}_k} \ln p_{\boldsymbol{\theta}}\left(\mathbf{X}\right) = \sum_{n=1}^{N} \ln \left[\frac{\pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)} \cdot \frac{\partial}{\partial \boldsymbol{\mu}_k} \left[\mathcal{N}\left(\mathbf{x}_n \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\right)\right] \right]$
3. The partial derivative cannot be solved.

[5 marks]

TURN OVER

Question 19.

As a data scientist, you are given the following data matrix:

$$\mathbf{X} = \begin{bmatrix} 0.85 & -1.12 & 1.14 \\ 0.67 & -0.46 & 1.06 \\ 1.43 & -0.67 & -0.98 \\ -1.04 & -0.58 & 0.09 \\ 1.12 & -0.38 & -1.08 \\ 0.06 & 1.24 & -0.33 \end{bmatrix}, \tag{2}$$

where the rows correspond to features and the columns correspond to samples. The number of samples is $N = 3$, and the number of features is $D = 6$. Which of the following statements is/are true?

1. The problem is overdetermined.
2. The problem is underdetermined.
3. $P = 3$ principal components are required to explain 100% of the variance in the data.
4. $P = 5$ principal components are required to explain 95% of the variance in the data.

[5 marks]

Question 20.

Which of the following statements is/are true? Principal Component Analysis (PCA)...

1. ... is not a supervised learning algorithm.
2. ... can be used for dimensionality reduction.
3. ... can be used for data analysis.
4. ... minimises the variance of the projected data.

[5 marks]

END OF PAPER