The University Of Sheffield.

Data Provided:
None

DEPARTMENT OF COMPUTER SCIENCE AUTUMN SEMESTER 2019

Machine Learning and Adaptive Intelligence **2 hours**

Answer ALL the questions.

Figures in square brackets indicate the marks allocated to each part of a question, out of 100.

This page is blank.

1. Probability and Linear Algebra for Machine Learning [Total: 25 marks]

   a) The joint probability of two random variables X and Y is $P(X, Y) = \frac{1}{36}$. The marginal probability of X is $P(X) = \frac{5}{36}$ and the marginal probability of Y is $P(Y) = \frac{6}{36}$. Determine if the random variables are independent. Explain your answer. [5 marks]

   b) We have three independent random variables X, Y and Z that follow Gaussian distributions as follows $X \sim \mathcal{N}(2, 1)$, $Y \sim \mathcal{N}(-3, 1)$ and $Z \sim \mathcal{N}(0, 2)$. What is the distribution of the random variable $W = X + 3Y + Z$? [10 marks]

   c) Let us define a matrix $\mathbf{W}$ of dimensions $n \times m$, a vector $\mathbf{x}$ of dimensions $m \times 1$ and a vector $\mathbf{y}$ of dimensions $n \times 1$. Write the following expression in matrix form

$$\sum_{i=1}^{n}\sum_{j=1}^{m} w_{i,j}x_j + \sum_{j=1}^{m}\sum_{i=1}^{n} y_i w_{i,j}.$$

   [HINT: if necessary define a vector of ones $\mathbf{1}_p = [1 \cdots 1]^\top$ of dimensions $p \times 1$, where p can be any number]. [10 marks]

2. Regression and Objective Functions [Total: 25 marks]

Consider a regression problem for which each observed output y_i has an associated weight factor $r_i > 0$, such that the sum of weighted squared errors is given as

$$E(m, c) = \sum_{i=1}^{n} r_i (y_i - f(x_i))^2 = \sum_{i=1}^{n} r_i (y_i - mx_i - c)^2,$$

with $f(x_i) = mx_i + c$, where m is the slope and c the intercept.

a) We want to use a **coordinate descent** procedure to optimise the values of m and c. Find the fixed-point updates for m and c. [15 marks]

b) When n increases considerably, we would rather use **stochastic gradient descent** (SGD) to optimise m and c. Assume that we have access to the optimal value for $c^* = 1$. Derive the SGD update equation for m. Once you get the expression, assume that you initialise your parameter at time zero with $m = 0$ and that the first two random observations that you get are $(r_1, x_1, y_1) = (1, \frac{1}{3}, \frac{1}{2})$ and $(r_2, x_2, y_2) = (\frac{1}{2}, -3, \frac{5}{2})$. If you use a learning rate of $\eta = \frac{1}{2}$, what is the updated value of m after observing these two datapoints? [10 marks]

3. Bayesian Regression and Naive Bayes [Total: 25 marks]

a) Given training data D, model training for probabilistic models often involves taking point estimates for the model parameters θ, such as the *maximum a posteriori* (MAP) estimate. MAP aims to maximise the posterior probability.

Make use of Bayes' rule and log probability to define the objective function optimised to obtain the MAP estimate. Make sure that your expression has the fewest possible terms (i.e., remove any term that does not contribute to the maximisation). [10 marks]

b) Consider a dataset with 14 entries in the following table. Four weather condition variables *O*, *T*, *H*, and *W* help us predict the decision *C* (yes/no) on whether to play tennis.

Entry	**Outlook (O)**	**Temperature (T)**	**Humidity (H)**	**Wind (W)**	**PlayTennis (C)**
1	Sunny	Hot	High	Weak	no
2	Sunny	Hot	High	Strong	no
3	Cloudy	Hot	High	Weak	yes
4	Rain	Mild	High	Weak	no
5	Rain	Cool	Normal	Weak	yes
6	Rain	Cool	Normal	Strong	no
7	Cloudy	Cool	Normal	Strong	yes
8	Sunny	Mild	High	Weak	yes
9	Sunny	Cool	Normal	Weak	yes
10	Rain	Mild	Normal	Weak	no
11	Sunny	Mild	Normal	Strong	yes
12	Cloudy	Mild	High	Strong	yes
13	Cloudy	Hot	Normal	Weak	yes
14	Rain	Mild	High	Strong	no

Given a new entry *X=(O=Sunny, T=Cool, H=High, W=Strong)*, use a Naive Bayes classifier to determine whether the decision value of *C* (PlayTennis) is yes or no. Include all the necessary intermediate steps, all the computed probabilities, and the final results.

[15 marks]

4. Principal Component Analysis (PCA) and Logistic Regression [Total: 25 marks]

a) Using the method of Lagrange multipliers, show that for a scatter matrix (sample covariance matrix) $\mathbf{S}$, the expression $\mathbf{u}^\top \mathbf{S}\mathbf{u}$, subject to the constraint $\mathbf{u}^\top \mathbf{u} = 1$, is maximised when $\mathbf{u}$ is an eigenvector of $\mathbf{S}$. [8 marks]

b) For a 3×3 covariance matrix Σ, the corresponding eigenvectors are given in a matrix below (each column is an eigenvector):

$$\begin{bmatrix} 0 & 0.47 & -0.88 \\ 0 & -0.88 & -0.47 \\ 1 & 0 & 0 \end{bmatrix}$$

with corresponding eigenvalues $\{0, 0.065, 1.26\}$.

Apply PCA using the eigenvectors above to two data points $\mathbf{x}_1 = (2, 3, 3)^T$ and $\mathbf{x}_2 = (4, 1, 0)^T$ to reduce the dimension from 3 to 2. Report the computed low-dimensional representations for $\mathbf{x}_1$ and $\mathbf{x}_2$. Show the steps.

[7 marks]

c) The prediction function for logistic regression is the logistic function. Write down the logistic function that predicts the probability of positive outcome π of a binary classification problem as a function of the input $\mathbf{x}$, the parameter vector $\mathbf{w}$ and the basis function $\boldsymbol{\phi}(\cdot)$, and sketch a logistic function with the ranges of axes clearly labelled.

[6 marks]

d) Distinguish between generative models and discriminative models for probabilistic classification, and name one machine learning method for each.

[4 marks]

END OF QUESTION PAPER