The University Of Sheffield.

Data Provided:
None

DEPARTMENT OF COMPUTER SCIENCE **AUTUMN SEMESTER 2017–2018**

MACHINE LEARNING AND ADAPTIVE INTELLIGENCE **2 hours**

Answer THREE questions.

All questions carry equal weight. Figures in square brackets indicate the percentage of available marks allocated to each part of a question.

THIS PAGE IS BLANK.

1. Consider an Olympic marathon dataset given by (x_i, y_i) for $i = 1, \ldots, n$ where x_i and y_i denote a year and the winning time of that year, respectively. Suppose that linear regression is used for prediction of a winning time, and that the prediction function is given by

$$f(x_i) = mx_i + c$$

with a slope m and an offset c.

a) Define the objective function in L^2-norm, given the actual value and the prediction function described as above. [10%]

b) Explain briefly why the objective function, defined in 1(a), is appropriate for the task. [6%]

c) Derive the gradients for the slope and the offset. [16%]

d) Show update equations for the slope and the offset that are used in the gradient descent algorithm. Update equations require an additional factor. What is it? How does it affect the gradient descent algorithm? [20%]

e) Given the set of n data points, write pseudocode for calculating the slope and the offset using the gradient descent algorithm. Your code should include a suitable stopping criterion. [16%]

f) Explain your stopping criterion in 1(e). Suppose that you are able to run the code in 1(e), how would you expect the objective function to change at each iteration? [8%]

g) Rewrite the gradient descent pseudocode in 1(e) using minimal changes to create pseudocode for the 'stochastic' gradient descent algorithm. [16%]

h) Discuss the benefit of using the stochastic gradient descent instead of the conventional gradient descent algorithm. [8%]

2. This question concerns a non-linear regression model. Suppose that, for $i = 1, \ldots, n$, an observed data point (x_i, y_i) is presented by

$$y_i = f(x_i) + \varepsilon$$

where Gaussian noise $\varepsilon \sim \mathcal{N}(\mathbf{0}, \sigma^2\mathbf{I})$ is assumed.

a) A prediction function can be written as a set of non-linear basis functions, Φ, linearly combined with a set of weights, $\mathbf{w}$. By defining $\mathbf{w}$ and Φ suitably, write the prediction function. [12%]

b) Give one example of a basis function along with its mathematical expression that may be used in 2(a). [8%]

c) For the non-linear regression model stated above, derive (i) a data point likelihood, (ii) a dataset likelihood and (iii) a log likelihood for the dataset. [20%]

d) State two reasons why we would use the *log* likelihood rather than use the likelihood directly. Justify why it is a valid thing to do when calculating the maximum likelihood estimate. [12%]

e) Explain the relation between the maximum likelihood estimate and the sum of squared errors for an objective function. [8%]

f) We would like to construct a Bayesian regression model by extending the above non-linear model. To that end we assume that the parameter set $\mathbf{w}$ is also sampled from the Gaussian distribution, *i.e.*, $\mathbf{w} \sim \mathcal{N}(\mathbf{0}, \alpha\mathbf{I})$. In this scenario, explain the roles of probabilities, $p(\mathbf{w})$ and $p(\mathbf{w} \mid \mathbf{y}, \Phi)$. Show a mathematical expression that relates these two probabilities. [10%]

g) Explain why the Gaussian distribution is a suitable prior density for the model in 2(f). What might happen if a distribution, that was not Gaussian (*e.g.*, a gamma distribution), was used as a prior density? [10%]

h) It is known that the posterior is sampled from the Gaussian $\mathbf{w} \sim \mathcal{N}(\mu, \mathbf{C})$ where

$$\mu = \mathbf{C}\sigma^{-2}\Phi^\top\mathbf{y}, \qquad \mathbf{C} = \left[\sigma^{-2}\Phi^\top\Phi + \alpha^{-1}\mathbf{I}\right]^{-1}$$

Defining $\mathbf{X}$, $\mathbf{w}$ and $\mathbf{y}$ suitably, show that the vector space solution for a linear regression $\mathbf{X}^\top\mathbf{X}\mathbf{w} = \mathbf{X}^\top\mathbf{y}$ is the special case of the Bayesian regression posterior update. [20%]

3. a) A tumor marker is a biomarker that can be measured in the human body. It helps detect cancer because the increased value of a marker often, but not always, indicates the presence of one or more types of cancer.

Consider a tumor marker of a certain type. It is known that this marker shows positive with 80% of cases if one has cancer, however, the down side is that it also returns a positive result with 10% of cases where one does not have cancer.

Many people went through a diagnostic test using this marker. Statistics indicates that 1% of these people do have cancer, and the rest do not. What is the chance that one has cancer if this tumor marker returns a positive result? Show your work in full. [20%]

b) Let $(\mathbf{X},\mathbf{y}) = (\mathbf{x}_1, y_1, \ldots, \mathbf{x}_n, y_n)$ denote a training dataset, consisting of p-dimensional inputs $\mathbf{x}_i = (x_{i1}, \ldots, x_{ip})$ and the label data y_i. The task is to find the test label y^*, given a test input $\mathbf{x}^* = (x_1, \ldots, x_p)$, by evaluating the conditional density $p(y^* \mid \mathbf{X}, \mathbf{y}, \mathbf{x}^*, \theta)$ where θ denotes a set of model parameters. To this end, we consider use of a naive Bayes classifier.

(i) Roughly outline the derivation of the following expression:

$$p(y^* \mid \mathbf{X}, \mathbf{y}, \mathbf{x}^*, \theta) = \frac{\prod_{j=1}^{p} p(x_j^* \mid y^*, \theta) p(y^*)}{\sum_{y^*} \prod_{j=1}^{p} p(x_j^* \mid y^*, \theta) p(y^*)}$$

The derivation need not be mathematically exact, as long as the ideas are clear and correct, however you must explain carefully what naive Bayes assumption(s) you have made at certain step(s) in order to reach the above expression. [30%]

(ii) The right side of the equation in 3(b)(i) does not explicitely involve $\mathbf{X}$, $\mathbf{y}$ (training data) or n (the number of data points). Instead it incorporates the maximum likelihood estimate for θ that needs to be derived from the training set. Show the joint density of the training data, then suggest a suitable objective function. What may be needed if you would like to minimise the above objective function? Note that this question does not require you to derive the maximum likelihood estimate for θ. [30%]

(iii) The Bernoulli distribution with a parameter, π, given by

$$p(y_i) = \pi^{y_i}(1-\pi)^{(1-y_i)}$$

may be a suitable prior when the task involves a binary decision (*i.e.*, $y_i = 0, 1$). Derive the maximum likelihood estimate for π, given that there are n data points. [20%]

4. There are no more than 10 balls in a box. Balls are identical except that each ball is labelled either 1, 2 or 3. We draw balls with replacement — that is, each time we draw a ball, we record the label and return the ball to the box before the next draw.

   This question concerns optimisation with constraints using the method of Lagrange multipliers. The aim is to apply the concept of entropy to find the most unbiased estimates of the numbers of balls labelled 1, 2 and 3.

   a) Explain the principle of maximum entropy modelling. [10%]

   b) What are probabilities of drawing balls labelled 1, 2 and 3 when the entropy is (i) maximum and (ii) minimum? [20%]

   c) After drawing balls with replacement many times, we find that the average of labels drawn is $\frac{11}{7}$. Let p_1, p_2, p_3 denote probabilities that we draw a ball labelled 1, 2, or 3. Write the mathematical expressions for (i) the entropy H and (ii) all constraints in this scenario. [20%]

   d) Define suitable multiplier(s) and write down the Lagrangian Λ for the problem in 4(c). [10%]

   e) Solve the Lagrangian in 4(d) and find the most unbiased estimates for the numbers of balls labelled 1, 2 and 3. [40%]

END OF QUESTION PAPER