UNIVERSITY OF BIRMINGHAM

School of Computer Science

Machine Learning and Intelligent Data Analysis

Resit Examinations 2022

Machine Learning and Intelligent Data Analysis

Exam paper

Question 1 Clustering, Dimensionality Reduction, and Text Analysis

(a) Consider the following objects in 2-dimensions: $V(2, 10)$, $W(2, 5)$, $X(8, 4)$, $Y(5, 1)$, $Z(8, 5)$. Using min/single link and Manhattan distance, cluster these objects using hierarchical agglomerative clustering method. Show all the working (but no need to show the dendrogram). In addition, describe the cluster formation at height/distance 3. **[6 marks]**

(b) Consider the following three objects in 2-dimensions:

$$\mathbf{X} = \begin{pmatrix} 1 & 6 \\ 3 & 4 \\ 5 & 2 \end{pmatrix}$$

By following part of the principal component analysis process, estimate the covariance matrix for this data. **[4 marks]**

(c) The PageRank algorithm is well-known to be the basis of Google's search engine. However, PageRank is based only on the connectivity of documents and does not take their content into account at all, and therefore cannot provide results based on a specific search term. Suggest how this could be addressed. **[10 marks]**

Question 2 Linear Regression and Learning Theory

(a) The regularised form of the least square loss is $\mathcal{L}(\mathbf{w}, \lambda) = \mathcal{L}_{\text{err}}(\mathbf{w}) + \lambda R(\mathbf{w})$ where $\mathcal{L}_{\text{err}}$ is the least squares loss. The regularisation term $R(\mathbf{w}) = \alpha \|\mathbf{w}\|_2^2 + \beta \|\mathbf{w}\|_1$ is sometimes used in practical applications of regression.

Explain what effect this term will have on the characteristics of a model fitted using this loss, and suggest when this might be useful. **[5 marks]**

(b) Consider a regression problem in which we aim to predict a single dependent variable t from a single independent variable x.

It is known that the true data generating function is $t = h(x) + \epsilon$, where $h(x) = c$, a constant, and ϵ is normally distributed with mean 0 and variance $\sigma^2 = 1/2$.

We would like to estimate the value of c by fitting a model $f(x, w) = w$ using Bayesian regression. Our estimate for w provides an estimate for c.

The prior distribution of w is assumed to be $p(w) \propto \exp(-w^2)$.

A single data point $X = (x, t) = (3, 10)$ is known.

(i) In the absence of data, what is $\mathbb{E}[w]$ (the expected value of w)?

(ii) Write down the likelihood of the data point X.

(iii) Write down the posterior distribution of w given data point X.

(iv) Compute the posterior estimate of w by minimising the negative log of the posterior distribution. **Explain your answer**.

You may use the result that a quadratic $ax^2 + bx + c$ is minimised by $x = \frac{-b}{2a}$.

[10 marks]

(c) The following five pairs of numbers were sampled from a two-dimensional normal distribution with mean $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariance $\Sigma = \begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}$

x_1	2.02	-0.21	1.55	-0.05	0.81
x_2	-2.08	-1.18	-0.77	-1.15	1.32

Compute the sample mean and sample covariance, and explain the implications for learning of the results of your calculations. **[5 marks]**

Question 3 Classification

(a) Logistic regression is based on the cross entropy loss function shown below (to be minimised):

$$E(\mathbf{w}) = -1 \times \sum_{i=1}^{N} y^{(i)} \ln p_1(\mathbf{x}^{(i)}, \mathbf{w}) + (1 - y^{(i)}) \ln (1 - p_1(\mathbf{x}^{(i)}, \mathbf{w})) \qquad (1)$$

where $\mathbf{w}$ is the vector of parameters of the Logistic Regression model, $\mathbf{x}^{(i)}$ is the vector of input variables of example i, $y^{(i)}$ is the output variable of example i, N is the number of training examples, $p_1(\mathbf{x}^{(i)}, \mathbf{w}) = \exp(\mathbf{w}^T\mathbf{x})/(1 + \exp(\mathbf{w}^T\mathbf{x}))$ and exp is the exponential function.

Answer the following questions regarding the components shown in red of this loss function:

(i) What is the effect of multiplying this equation by -1 on the training process? **Justify** your answer in detail. **[6 marks]**

(ii) Why are the left and right terms of the summation multiplied by $y^{(i)}$ and $(1 - y^{(i)})$, respectively? **Justify** your answer in detail. **[4 marks]**

(b) The Gaussian Kernel is a very popular kernel that is frequently used with Support Vector Machines. It is defined based on a Gaussian function, which is associated to a hyperparameter σ:

$$k(\mathbf{x}, \mathbf{x}^{(n)}) = e^{-\frac{\|x - x^{(n)}\|^2}{2\sigma^2}}$$

Explain the effects that **increasing** and **reducing** the value of σ would have on the function below, which is used to predict the output value of an example described by the input vector $\mathbf{x}$:

$$f(\mathbf{x}) = \sum_{n \in S} a^{(n)} y^{(n)} k(\mathbf{x}, \mathbf{x}^{(n)}) \; + b$$

where $a^{(n)}$ is the Lagrange multiplier associated to the support vector n, $y^{(n)}$ is the output value of the support vector n, $\mathbf{x}^{(n)}$ is the vector of input values of the support vector n, S is the set of indexes of the support vectors and

$$b = \frac{1}{N_S} \sum_{n \in S} \left(y^{(n)} - \sum_{m \in S} a^{(m)} y^{(m)} k(\mathbf{x}^{(n)}, \mathbf{x}^{(m)}) \right)$$

where N_S is the number of support vectors.

Instructions: assume that the Lagrange multipliers associated to all support vectors always have the same value, i.e., assume that the kernel and output values are the only factors influencing $f(\mathbf{x})$.

[10 marks]