Machine Learning and Intelligent Data Analysis
Solutions

Resit Examinations 2022

Machine Learning and Intelligent Data Analysis

Learning Outcomes

(a) Demonstrate knowledge and understanding of core ideas and foundations of unsupervised and supervised learning on vectorial data

(b) Explain principles and techniques for mining textual data

(c) Demonstrate understanding of the principles of efficient web-mining algorithms

(d) Demonstrate understanding of broader issues of learning and generalisation in machine learning and data analysis systems

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]**

Model answer / LOs / Creativity:

Learning outcomes a, b, c. Part a is creative.

2 marks As a first step, we will need to compute the distance matrix using the Manhattan distance which will look as below:

	X	W	X	Y	Z
V		5	12	12	11
W			7	7	6
X				6	1
Y					7
Z					

[3 marks] Next step is to successively merge objects/clusters using the min/single link. The merge will happen at heights as below:

- @1: (X,Z), V,W,Y
- @5: (X,Z), (V,W), Y
- @6: ((X,Z), (V,W), Y)

[1 mark] At height 3, the cluster formation will be same as at distance 1 i.e. (X,Z), V,W,Y as the next distance after 1 at which clusters merge is distance 5.

1 mark As a first step, we need to estimate mean for mean subtraction: $\bar{\mathbf{X}} = \begin{pmatrix} 3 \\ 4 \end{pmatrix}$

[1 mark] Then we need to conduct mean subtraction: $\tilde{\mathbf{X}} = \begin{pmatrix} -2 & 2 \\ 0 & 0 \\ 2 & -2 \end{pmatrix}$

[2 marks] Next step is to estimate the covariance matrix using the mean subtracted data: $\mathbf{\Sigma} = \frac{1}{3}\tilde{\mathbf{X}}^\top\tilde{\mathbf{X}} = \frac{1}{3}\begin{pmatrix} 8 & -8 \\ -8 & 8 \end{pmatrix}$

(a)

- What's needed is a method for assessing the relevance of a document to the query, eg the TF-IDF similarity.
- The goal will then be to first find the documents that are relevant to the query, then to return those documents that PageRank deems authoritative.
- Searching can therefore proceed in two stages. i) use TF-IDF similarity to compute which documents match the query (using the inverse index to speed this up), possibly using some threshold below which documents are deemed irrelevant. ii) Rank the remaining documents according to their PageRank.

[10 marks to be awarded, with 1 mark for each relevant point raised.]

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]**

Model answer / LOs / Creativity:
Learning outcomes a, d. Part b is creative.

(a) This is the ElasticNet loss (but is not named as such to avoid students simply looking it up). It combines the L_2 and L_1 regularisers and therefore simultaneously shrinks and sparsifies the model. The resulting model should therefore "select" the relevant terms by the sparsification property of the L_1 loss, and then shrink over the other

terms. The extent to which it does each of these is determined by $\alpha/beta$. Because the L_2 term makes the loss convex, and the L_1 term sparsifies the solution, this is particularly useful for high dimensional problems.

(b) (i) $p(w) = \exp(-x^2)$ is a normal distribution with $\mathbb{E}[x] = \mu = 0$ [1 mark]

(ii) The likelihood (with $\sigma = 1/\sqrt{2}$) is $p(t|w) = \exp[-(t-f(x,w)^2)] = \exp[-(10-w)^2]$ [2 marks]

(iii) The posterior is $p(w|t) = p(w)p(t|w) = \exp[-w^2-(10-w)^2] = \exp[-(2w^2-20w+100)]$ [2 marks]

(iv) $-\log p(w|t) = 2w^2 - 20w + 100$ which is minimised by $w = 5$. [2 marks]
This is not the answer we would necessarily expect: the data point implies that $c \approx 10$. However, we have to include the effect of the prior. Since the prior and the likelihood have the same variance, with a single data point we end up with the average of the prior and the max likelihood estimate. [3 marks]

(c) The data were were drawn from a two-dimensional normal distribution (the population distribution) with mean $\mu = \begin{pmatrix} 0 \\ 0 \end{pmatrix}$ and covariance $\Sigma = \begin{pmatrix} 2 & -1 \\ -1 & 3 \end{pmatrix}$ and therefore a training sample drawn from this distribution should ideally have the same distribution as the population [2 marks].

The mean of the sample is $(0.82, -0.77)$ and the covariance is $\begin{pmatrix} 0.95 & -0.21 \\ -0.21 & 1.60 \end{pmatrix}$. [2 marks]

These are not the same as the population statistics and this illustrate the different between the true risk and the empirical risk. It is likely that this data, were it to be used as training data for some learning algorithm, would not yield an accurate hypothesis with high probability. [2 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:

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

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]

Model answer / LOs / Creativity:
Learning outcomes a, d. Part *b* is creative.

(a) (i) The value of ln $p_1(\mathbf{x}^{(i)}, \mathbf{w})$ and ln $(1 - p_1(\mathbf{x}^{(i)}, \mathbf{w}))$ will always be either zero or negative. The value of zero happens when the probability p_1 associated to the training example is equal to the target probability. A negative value means that it is different from the target probability, meaning that this training example is incurring some loss (2 marks). Multiplying by -1 will mean that any incurred loss becomes positive, i.e., it leads to a worse (larger) value for the loss function (2 marks). As the loss function is to be minimised, multiplying it by -1 will guide the learning process towards learning weights that assign probabilities as close as possible to the target probabilities (2 marks).

(ii) This is necessary so that the left (right) term will contribute towards the summation only when the training example i belongs to class 1 (0) (2 marks). If that was not the case, then the value of ln $p_1(\mathbf{x}^{(i)}, \mathbf{w})$ would be summed for examples of class 0, meaning that a probability p_1 of zero for examples of class 0 would be considered a bad value, leading to an increase in the loss. However, such probability value should be considered as a very good value when the example belongs to class 0. A similar issue would happen with the right term being used for examples of class 1 (2 marks).

(b) Larger values of σ will result in a wider Gaussian function where the similarity values retrieved by the kernel would always be very similar to each other (2 marks). This means that the prediction given to a given example will receive a lot of influence from support vectors that are not so similar to it (2 marks). In particular, if we use an extremely large value for σ, then the similarities between different examples will always have a very similar value, meaning that all support vectors contribute almost equally to the predictions (1 mark).

In contrast, smaller values of σ will result in a narrower Gaussian function with a higher peak, where the similarity values retrieved by the kernel would only be high when the examples given as arguments to the kernel are very similar to each other (2 marks). Therefore, only the support vectors that are very similar to the example being predicted would provide a considerable contribution to the predictions (2 marks). In particular, an extremely small value for σ would mean that only the closest support vector would have any meagninful effect on the predictions (1 mark).