UNIVERSITY OF BIRMINGHAM

**School of Computer Science**

**Machine Learning and Intelligent Data Analysis**

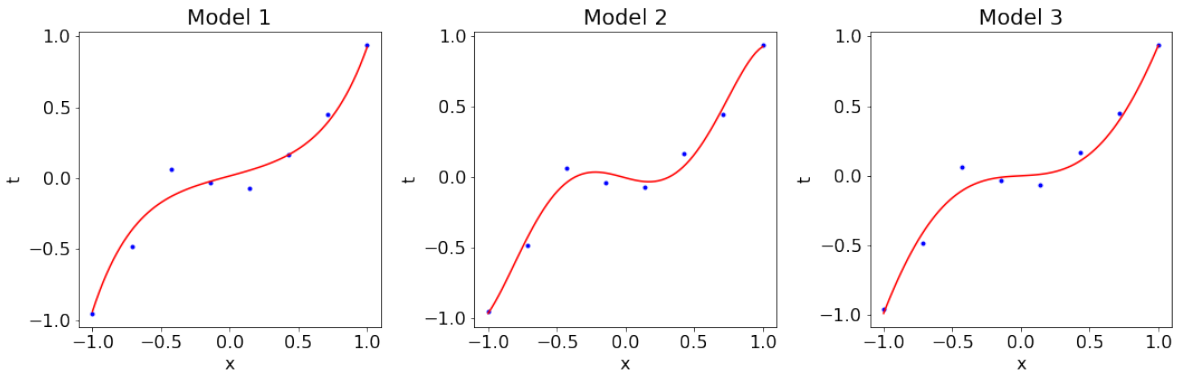
Main Summer Examinations 2022

Machine Learning and Intelligent Data Analysis

# Exam paper

## Question 1 Linear Regression and Learning Theory

(a) The images below show the results of fitting a dataset of $N = 8$ points (blue points) with a polynomial $f(x, \mathbf{w}) = \sum_{i=0}^{5} w_i x^i$. The line of best fit in each case is shown as a solid red line. Each fit was generated by minimising a least squares loss function with different regularisation applied.




The model parameters for each of the fits are.

| | $w_0$ | $w_1$ | $w_2$ | $w_3$ | $w_4$ | $w_5$ |
|---|---|---|---|---|---|---|
| Model 1 | 0.015 | 0.265 | 0.018 | 0.378 | -0.042 | 0.287 |
| Model 2 | -0.003 | -0.261 | 0.170 | 2.389 | -0.186 | -1.184 |
| Model 3 | 0.000 | 0.052 | 0.000 | 1.094 | -0.027 | -0.187 |

State what regularisation you think was used for each fit and **explain your reasoning**. **[9 marks]**

(b) You are working on a problem involving online Bayesian Regression on data that arrives in a stream. Beginning with a Gaussian prior with mean zero and variance one, you update the posterior distribution by multiplying by the likelihood of each event in the stream as it arrives. A colleague suggests that you could improve the efficiency of this by setting a threshold below which all values of the prior are set to zero. Do you think this is a good idea? **Explain your reasoning.** **[5 marks]**

(c) An instance class $X$ contains instances $\mathbf{x} \in X$ with twenty binary variables. A classifer $L$ is trained to return a hypothesis from the hypothesis class containing all possible binary conjunctions of three or fewer variables. Calculate an upper bound on the number of training samples needed to guarantee that the hypothesis returned by $L$ will have a true error of no more than 10% with 80% confidence. **[6 marks]**

## Question 2 Clustering, Dimensionality Reduction, and Text Analysis

(a) Consider the following two-dimensional data which has been divided into two clusters as shown:

- Cluster 1: (2,-3), (2,1), (3,2)
- Cluster 2: (-3,1), (-3,-2)

Given the above information, how can we determine which cluster a new point (0,0) belongs to by using k-means clustering algorithm? Show all the working/reasoning to support your answer. Describe any assumptions you may consider. **[6 marks]**

(b) Consider that we have estimated the below mean and covariance matrix of a 2-dimensional data set during the principal component analysis process.

$$\mu = \begin{pmatrix} 4 \\ 4 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 2 & -2 \\ -2 & 4 \end{pmatrix}$$

Make a rough drawing of the point cloud for this data set, assuming that the data is generated from a multivariate Gaussian probability density function. Briefly describe your observations about this data set by discussing the influence of the variance/covariance parameters on the point cloud shape. **[4 marks]**

(c) You are designing a document retrieval system in which a document $q$ should be queried against a corpus of $N$ *linked* documents $\{d_1, \ldots d_N\}$. Explain how you would design such a system and in particular how you would determine:

(i) Which documents should be returned in response to a query.

(ii) In what order the returned documents should be presented.

**Explain your reasoning and justify your choices.** **[10 marks]**

## Question 3 Sequential Minimal Optimisation

Sequential Minimal Optimisation (SMO) is a popular algorithm used with Support Vector Machines. Different variants of this algorithm have been proposed. In the version learned in this module, whenever updating the value of a given Lagrange multiplier $a^{(j)}$ in a given iteration, the new value of $a^{(j)}$ is "clipped" based on the lowest ($L$) and highest ($H$) possible values below:

- If $y^{(i)} = y^{(j)}$

  $L = \max(0,\ a^{(j)} + a^{(i)} - C)$

  $H = \min(C,\ a^{(j)} + a^{(i)})$

- If $y^{(i)} \neq y^{(j)}$

  $L = \max(0,\ a^{(j)} - a^{(i)})$

  $H = \min(C,\ C + a^{(j)} - a^{(i)})$

where $a^{(i)}$ is another Lagrange multiplier being updated and $C$ is a hyperparameter to control the strength of the penalty incurred by the Slack variables.

(a) **Explain briefly** what could happen if one was adopting SMO and forgot to "clip" the values of $a^{(j)}$. **[10 marks]**

(b) **Explain in detail** why $H = \min(C,\ C + a^{(j)} - a^{(i)})$ is an appropriate highest possible value for $a^{(j)}$ when $y^{(i)} \neq y^{(j)}$. **[10 marks]**