# Machine Learning and Intelligent Data Analysis
# Solutions

Main Summer 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

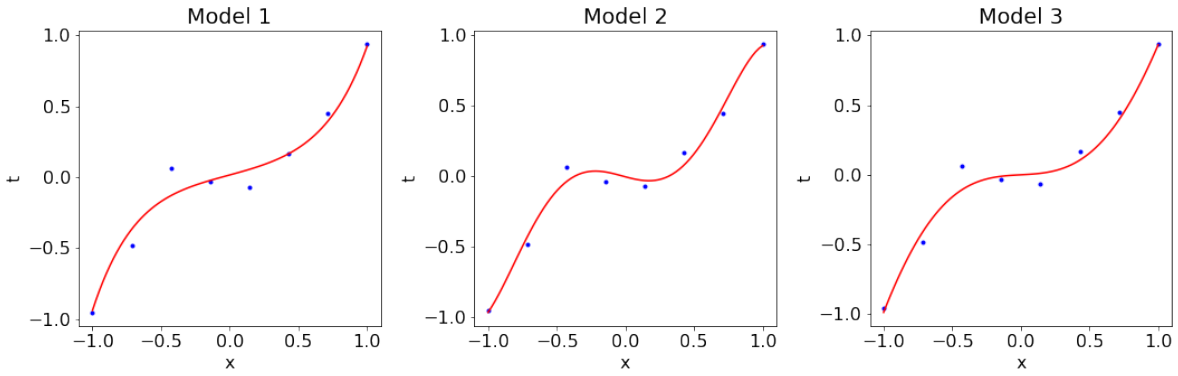
(c) Demonstrate understanding of the principles of efficient web-mining algorithms

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

# Exam paper

## Question 1 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]**

**Model answer / LOs / Creativity:**
Learning outcomes a, d. Parts a and b are unseen and creative.

(a)
- Model 1: $L_2$–regularised linear regression. The clue to this is in the shrinkage of the parameter vector compare to model 2. [3 marks]
- Model 2: unregularised linear regression. The clues to this are the presence of high order terms of similar magnitude to the low order terms. [3 marks]
- Model 3: $L_1$–regularised linear regression. The clue to this is in the sparsification of the parameter vector. [3 marks]

(b) It is very definitely not a good idea. Setting the prior to zero for certain parameter values means that no matter how strong the evidence, the likelihood, which is multiplicative, can never update the posterior at those parameter values: they will stay zero even if all the evidence points to them being the correct value. [5 marks]

(c) Since the learner can correctly classify its training set that it is a consistent learner [2 marks].

The hardest part of this question is computation of the size of the hypothesis space. It is **not** $|H| = 3^{20}$ (20 bits, 3 states: 0, 1, ignore) as the question asks for "all possible binary conjunctions of three or fewer variables". How of these are there?

- All conjuctions of no variable: 1 hypothesis
- All conjuctions of one variable: 20 variables, 2 choices for each ($A$, $\bar{A}$) — 40 hypotheses
- All conjuctions of two variables: $\binom{20}{2} = 190$ pairs, 4 choices for each ($AB$, $A\bar{B}$, etc) — 760 hypotheses
- All conjuctions of three variables: $\binom{20}{3} = 1140$ pairs, 8 choices for each ($ABC$, $AB\bar{C}$, etc) — 9120 hypotheses

There are therefore $|H| = 1 + 40 + 760 + 9120 = 9921$ hypotheses

[2 marks].

The upper bound is then given by $m \geq \frac{1}{\epsilon}\left(\ln|H| + \ln\frac{1}{\delta}\right)$ with $\epsilon = 0.1$ and $\delta = 0.2$, which gives $m = 108$ [2 marks]

[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/co-variance 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]**

**Model answer / LOs / Creativity:**

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

(a) [1 mark] For this question, we need to identify that we can take two alternating positions with both of them having merits. For full credits, the answer should identify both positions and highlight the issues.

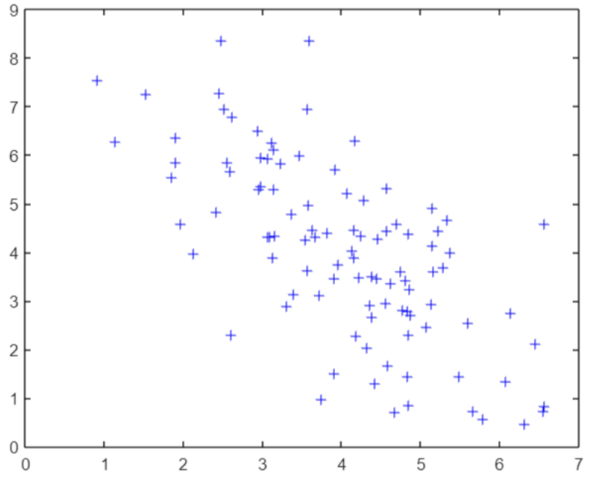
[3 marks] Position 1: we can find the distance of new point to both the cluster centres and assign it to the nearest cluster. However, this would necessitate recalculation of means after object assignment which could in turn change the cluster. First iteration is shown below:

As a first step, the means of cluster 1 and cluster 2 should be found which are: (2.33,0) and (-3,-0.5). Using squared Euclidean distance, the distance of new point (0,0) to both of these cluster centres is determined. The distance is: (0,0) to cluster

1 centre is 5.4289 and (0,0) distance to cluster 2 centre is 9.25. Hence, the point will be assigned to cluster 1. We will then recalculate the new centres which will be (1.75, 0) and (-3,-0.5), but this wouldn't alter the object assignment.

[2 marks] Position 2: we can take the position that since the clusters are known as per given information and thus assigning a new point to known groups becomes a classification problem rather than clustering problem. Hence, the answer could propose (and/or show) an algorithm like k-nn for the assignment of new point to one of the two 'known classes.

(b) [2 marks] From the mean and covariance, the rough drawing should look similar to the one below:




[2 mark] Since there is negative covariance (-2) between the two dimensions, it's showing in the point cloud that as x-axis value increases the y-axis value decreases. Further, the spread on the x-axis is narrow than the spread on the vertical axis since the x-axis variance is lower (2) compared to the y-axis variance (4).

(c) This is an open-ended and creative question with no single correct solution, but we covered TF-IDF similarity and PageRank in the lectures and I would expect these to be prominent in the submissions. The basic idea is that TF-IDF can be used to filter the documents for relevance which allows an initial filtering by content, then PR can be used to order by authority. However, we did not discuss this explicitly in lectures and so students will have to be creative and I would expect the following potential solutions to come up:

- Order by the sum of the sim and PR
- Order by the product of the sim and PR
- Threshold the sim and then order by PR so as only to return docs that match the query well. This is the best option here.

- Threshold by PR and then order by sim. (downside: could discard best results if the query is very specialised)

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

## 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]**

**Model answer / LOs / Creativity:**
Learning outcome a. Part *b* is creative.

(a) The boundaries $L$ and $H$ are used to ensure that not only the box constraints are satisfied, but also the constraint $\sum_{n=1}^{N} a^{(n)} y^{(n)} = 0$. If one forgets to clip these values, these constraints may not be satisfied.

[5 marks] for noting the potential violation of the box constraint and [5 marks] for noting the potential violation of the other constraint.

(b) The latter above-mentioned constraint can be re-written as follows:

$a^{(i)} y^{(i)} + a^{(j)} y^{(j)} = \zeta$, where $\zeta = -\sum_{n \neq i,j} a^{(n)} y^{(n)}$.

When $y^{(i)} \neq y^{(j)}$, we have the two possible scenarios [2 marks] :

- $y^{(j)} = -1$ and $y^{(i)} = 1$, which leads to $a^{(i)} - a^{(j)} = \zeta$.
- $y^{(j)} = 1$ and $y^{(i)} = -1$, which leads to $-a^{(i)} + a^{(j)} = \zeta$.

If we need to clip the value of $a^{(j)}$ to a highest possible value as specified in this question, this means that the adjustment in $a^{(j)}$ is increasing the value of this Lagrange multiplier.[2 marks]

In the first aforementioned case, an increase in $a^{(j)}$ will require an increase of equal amount in $a^{(i)}$ so that $a^{(i)} - a^{(j)}$ remains equal to $\zeta$. Similarly, in the second aforementioned case, an increase in $a^{(j)}$ will require an increase of equal amount in $a^{(i)}$ so that $-a^{(i)} + a^{(j)}$ remains equal to $\zeta$. [2 marks].

This means that highest possible value of $a^{(j)}$ in this iteration is constrained not only by the hyperparameter $C$ but also by the maximum possible amount by which $a^{(i)}$ can increase. Specifically, the highest possible value of $a^{(j)}$ can exceed neither $C$ nor $a^{(j)} + D$, where $D$ is the amount by which $a^{(i)}$ can increase. [2 marks]

The maximum amount by which $a^{(i)}$ can increase is $D = C - a^{(i)}$, as any increase beyond that would mean that $a^{(i)}$ becomes larger than $C$, violating the box constraints. [2 marks]

Therefore, the maximum value for $a^{(j)}$ in this iteration can exceed neither $C$ nor $a^{(j)} + (C - a^{(i)})$, which leads to $H = \min(C,\ C + a^{(j)} - a^{(i)})$.