LH Machine Learning and Intelligent Data Analysis
Solutions

Main Summer Examinations 2023

Note

Answer ALL questions. Each question will be marked out of 20. The paper will be marked out of 60, which will be rescaled to a mark out of 100.

Question 1 Regression

(a) The least squares error function is defined as

$$\mathcal{L}(\mathbf{w}) = \sum_{i=1}^{N} (t_i - f(x_i, \mathbf{w}))^2 \, .$$

This function is commonly used to measure how well a function $f(x, \mathbf{w})$ parameterised by $\mathbf{w}$ fits a set of N data points $\mathcal{D} = \{(x_i, t_i)\}_{i=1}^{N}$.

The likelihood of a data point t having been generated by given a model $f(x, \mathbf{w})$ can be written as $p(t|f(x, \mathbf{w}))$. Explain how, and under what assumptions, the least squares error is derived from the likelihood. You do not need to reproduce all of the mathematical steps of the derivation. **[6 marks]**

(b) Given some dataset, the expected value of the LSE $\mathcal{L}$ can be written as

$$\mathbb{E}[\mathcal{L}] = \sigma^2 + \mathrm{var}[f] + (h - \mathbb{E}[f])^2,$$

where σ^2 is the variance of the data, f is the estimated fit, and h is the true data generating function. Explain the terms in this expression and its relevance for learning. **[6 marks]**

(c) Given the data point $(2, 1)$, sketch a diagram of the likelihood in parameter space that this data point was generated by functions of the form $f(x, \mathbf{w}) = w_0 + w_1 x$. Your sketch should cover the domain $\{w_0, w_1\} \in [-1, 1]$. **[8 marks]**

Model answer / LOs / Creativity:

(a) Least squares error is the function (technically the functional) that, when minimised, maximises the likelihood of the data [2 marks] under the assumption of normally distributed iid data points. [2 marks]

The steps in the derivation are:

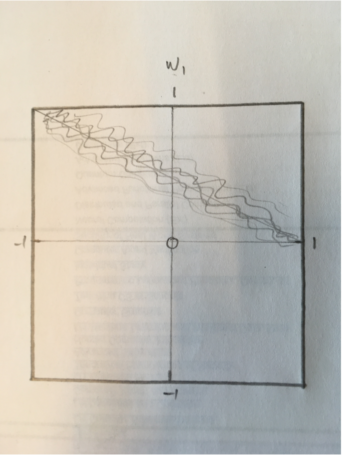
- Assume Gaussian likelihood function.
- Assume data points are independent and identically distributed. Then likelihood is a product of identical univariate Gaussians.
- Maxmimising the likelihood is equivalent to maximising it's log.

- The log of the product of Gaussians becomes the sum of the logs of the Gaussians.
- The sum of the logs of the Gaussians is the negative LSE.
- So maximising the likelihood is equivalent to minimising the LSE

[2 marks]

(b) This is the bias-variance decomposition. It is the expected value of the least squares error expressed as the sum of three terms: i) the noise in the data [1]; ii) the variance of the estimator [1]; iii) the difference (bias) between the true function and the expected value of the estimator [1]. It implies that for a given goodness-of-fit, there is a trade-off between the bias and variance of the estimator, with lower bias models necessarily having a higher variance for the same expected LSE and vice-versa. This places a fundamental limit on the generalisability of a learned model, because simple models will tend to have high bias, and complex models high variance. [2].

(c) We are looking here for the set of lines that pass near to the point having a higher probability, and those that do not having a lower probability. Those lines that pass through the point satisfy $w_0 + 2w_1 = 1$ or $w_1 = (1 - w_0)/2$ and so this defines the "ridge" of maximum likelihood. The sketch should therefore look something like the following:



Learning outcomes: Demonstrate knowledge and understanding of core ideas and foundations of unsupervised and supervised learning on vectorial data; Demonstrate understanding of broader issues of learning and generalisation in machine learning and data analysis systems. The creative part is part (c).

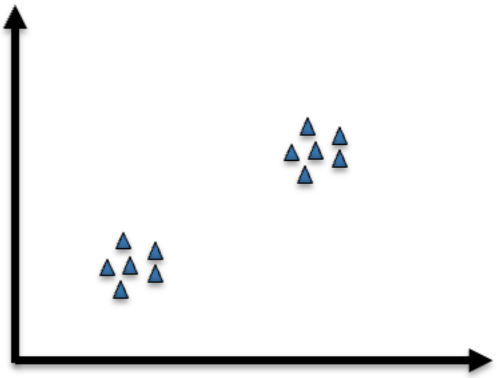
Question 2 Clustering, Dimensionality and Text Analysis

(a) Cluster the data in the table below using hierarchical clustering with Euclidean distance and single linkage, and draw the dendrogram.

Label	A	B	C	D	E
Coordinates	(4,2)	(7,8)	(3,2)	(3,4)	(8,7)

[6 marks]

(b) The graph below shows a two dimensional dataset.



(i) Reproduce the plot and draw the first and second principal components. Your drawing does not need to be completely accurate but should capture the key features. **Explain your reasoning**

(ii) If you were to use PCA to reduce the dimensionality of this data to just 1 dimension, show how the points will be mapped onto the new dimension (your drawing should give the general idea of the mapping).

(iii) Describe one way to determine how many dimensions should be kept in PCA.

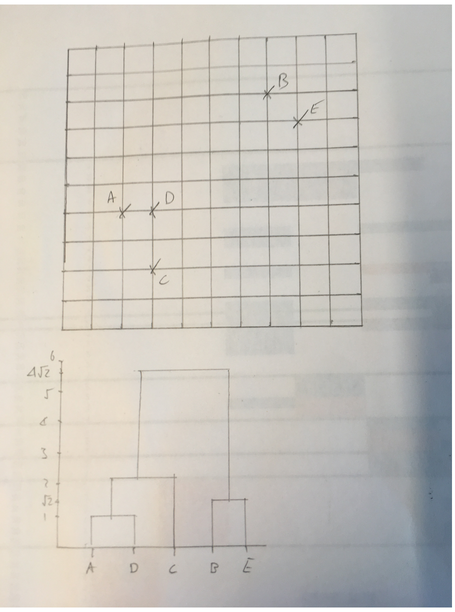
[6 marks]

(c) Document vectorisation and Page Rank are both methods for ranking documents. Document vectorisation allows documents to be ranked by their similarity to a query document. Page Rank provides a way to rank a collection of *linked documents* by considering their *authority* which is derived from the connectivity of each document.

Explain briefly how these two methods might be combined to implement a basic search engine for collection of linked documents that takes both document content and document authority into account. Briefly discuss any limitations of your approach. **[8 marks]**

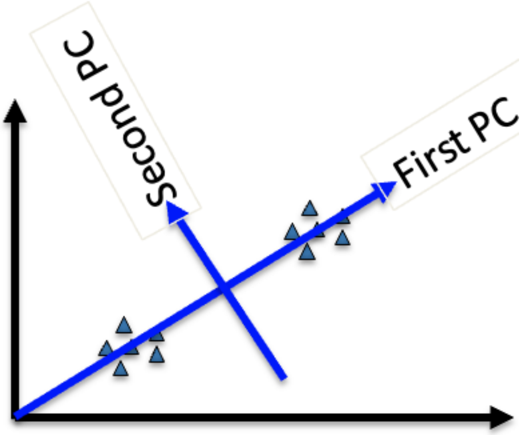
Model answer / LOs / Creativity:

(a) The correct clustering is $(((A, D, 1.0), C, 2.0), (B, E, \sqrt{(2)}), 4\sqrt{2})$. A plot of the data and the resulting dendrogram is shown below:



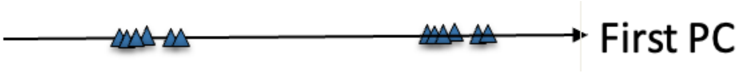
[3 marks for the correct clustering, 3 marks for correct distances]

(b) (i) The first PC should align with the direction of greatest variation in the data which in this case runs roughly through the centres of the two visible "clusters". The second PC is perpendicular to this. Note that the directions of the PCs are arbitary.



. [2 marks]

(ii) A correct sketch should show the data projected on to the first PC. The key feature is that two distinct "clusters" should remain obvious.



[2 marks]

(iii) The most common way is by computing the amount of variance explained by each principal component, and retaining those the explain about 90% of the total variance in the data, but this is highly problem dependent [2 marks].

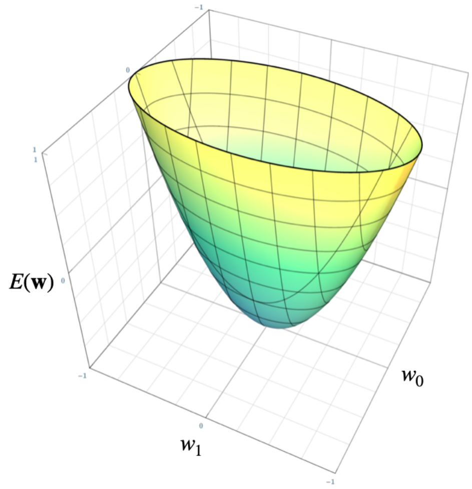
(c) Several answers are possible here. The key is to realise what is required of a search engine. It should return only results that are relevant, and should return those results in a sensible order. Credit will be given for sensible proposals that are carefully thought-through in terms of their implications for both the quality of results and the cost. The optimal solution involves filtering by content similarity first, then computing authority over that subset of documents to return the most authoritative documents first. [8 marks]

Learning outcomes: Demonstrate knowledge and understanding of core ideas and foundations of unsupervised and supervised learning on vectorial data; Explain principles and techniques for mining textual data; Demonstrate understanding of the principles of efficient web-mining algorithms. All parts of the question involve some creativity.

Non-alpha only

Question 3 Classification

(a) Which algorithm (Gradient Descent or Iterative Reweighted Least Squares) would be better to learn the weights w_0 and w_1 of a logistic regression model for a problem with the loss function $E(\mathbf{w})$ below, which is an elliptic quadratic function? **Justify** your answer by explaining how these two algorithms would work **in the context of this problem**. **[6 marks]**



(b) Logistic regression models for binary classification can be trained by maximising the log-likelihood:

$$ln(\mathcal{L}(\mathbf{w})) = \sum_{i=1}^{N} y^{(i)}\ ln\ p_1(\mathbf{x}^{(i)}, \mathbf{w}) + (1 - y^{(i)})\ ln\ (1 - p_1(\mathbf{x}^{(i)}, \mathbf{w})).$$

where $\mathbf{w}$ are the weights of the logistic regression model, $y^{(i)} \in \{0, 1\}$ is the output variable of training example i, $\mathbf{x}^{(i)} \in \mathcal{X}$ are the input variables of training example i, $\mathcal{X}$ is the input space, N is the number of training examples, and $p_1(\mathbf{x}^{(i)}, \mathbf{w})$ is the probability of example i to belong to class 1 given $\mathbf{x}^{(i)}$ and $\mathbf{w}$.

How would you modify the log-likelihood function above so that it also works for problems with $M > 2$ classes? **Explain** your function.

PS: Please create a **single** log-likelihood function and make sure to define any variable or symbol that is different from the ones defined above.

[7 marks]

Question 3 continued over the page

(c) Prove that the kernel below is a valid kernel based on the kernel composition rules below and the fact that $\mathbf{x}^T\mathbf{z}$ is a valid kernel.

$$k(\mathbf{x},\mathbf{z}) = 10(e^{(x^Tz)})^2 + 2 + \mathbf{x}^T\mathbf{z}$$

Kernel composition rules, given two valid kernels $k_1(\mathbf{x},\mathbf{z})$ and $k_2(\mathbf{x},\mathbf{z})$:

1	$k(\mathbf{x},\mathbf{z}) = ck_1(\mathbf{x},\mathbf{z})$, where $c > 0$ is a constant
2	$k(\mathbf{x},\mathbf{z}) = f(\mathbf{x})k_1(\mathbf{x},\mathbf{z})f(\mathbf{z})$, where $f()$ is any function
3	$k(\mathbf{x},\mathbf{z}) = q(k_1(\mathbf{x},\mathbf{z}))$, where $q()$ is a polynomial with non-negative coefficients
4	$k(\mathbf{x},\mathbf{z}) = e^{k_1(x,z)}$
5	$k(\mathbf{x},\mathbf{z}) = k_1(\mathbf{x},\mathbf{z}) + k_2(\mathbf{x},\mathbf{z})$
6	$k(\mathbf{x},\mathbf{z}) = k_1(\mathbf{x},\mathbf{z})k_2(\mathbf{x},\mathbf{z})$

[7 marks]

Model answer / LOs / Creativity:

(a) Iterative Reweighted Least Squares would be better. This is because this algorithm uses a Taylor polynomial of degree two to approximate the loss function. The minimum of the Taylor polynomial is then obtained by setting the gradient to zero. As this loss function is a quadratic function, this approximation is perfect. Therefore, the minimum of the Taylor polynomial is also the minimum of this function and the algorithm would be able to find the optimum in a single step [3 marks].

In contrast, Gradient Descent updates the weights in the direction of the steepest descent. However, as this function is elliptical, such updates are not steps that go directly to the optimum. They are likely to overshoot the optimum in the direction of the w_0 axis. Even though Gradient Descent will eventually reach the optimum if a suitable learning rate is used, several steps (weight updates) would be typically necessary for that [3 marks].

(b)

$$ln(\mathcal{L}(\mathbf{w})) = \sum_{i=1}^{N}\sum_{k=1}^{M} y_k^{(i)}\ ln\ p_k(\mathbf{x}^{(i)},\mathbf{w}),$$

where $y_k^{(i)}$ is 1 if example i belongs to class k and 0 otherwise; and $p_k(\mathbf{x}^{(i)},\mathbf{w})$ is the probability of example i to belong to class k computed using $\mathbf{w}$.

This function sums the log of the probability of each example to belong to its true class k. This is because the $y_k^{(i)}$ multiplying $ln\ p_k(\mathbf{x}^{(i)},\mathbf{w})$ will only have value 1 for the true class to which this example belongs, resulting in the log of the probability of the example to belong to its true class being added to the summation. For all other classes, it will have value 0, resulting in zero being added to the summation.

(c) This can be proved, for example, by using the following kernel composition rules:

- (5) to get $10(e^{(\mathrm{x}^T\mathrm{z})})^2 + 2$ and $(\mathbf{x}^T\mathbf{z})$
- (3) to get $e^{(\mathrm{x}^T\mathrm{z})}$ from $10(e^{(\mathrm{x}^T\mathrm{z})})^2 + 2$
- (4) to get $\mathbf{x}^T\mathbf{z}$ from $e^{(\mathrm{x}^T\mathrm{z})}$

Marking scheme: 7 marks for correct answer, 6 marks for almost correct answer, 3 marks for partially correct answer, 0 marks for wrong answer.

Learning outcomes: Demonstrate knowledge and understanding of core ideas and foundations of unsupervised and supervised learning on vectorial data; Demonstrate understanding of broader issues of learning and generalisation in machine learning and data analysis systems. The creative parts are parts (b) and (c).