FOR INTERNAL SCRUTINY (date of this version: 17/7/2023)

UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR10086 MACHINE LEARNING

Sunday 13 th August 2023

00:00 to 00:00

INSTRUCTIONS TO CANDIDATES

1. **Note that ALL QUESTIONS ARE COMPULSORY.**
2. **DIFFERENT QUESTIONS MAY HAVE DIFFERENT NUMBERS OF TOTAL MARKS. Take note of this in allocating time to questions.**
3. **This is an OPEN BOOK examination: books, notes and other written or printed material MAY BE CONSULTED during the examination. The use of electronic devices or electronic media is NOT PERMITTED.**

Year 3 Courses

Convener: D.Armstrong

External Examiners: J.Bowles, A.Pocklington, R.Mullins

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

1. The exponential distribution is commonly used to model how long a customer needs to wait to be served. Formally, the probability density to get served at time $x \geq 0$ is

$$p(x) = \lambda e^{-\lambda x}, \tag{1}$$

where $\lambda > 0$ is a parameter. Typically, the larger λ is, the shorter a customer needs to wait.

Suppose you want to estimate the λ parameter for a coffee shop. You follow n independent customers and record $x_1, x_2, \ldots, x_n$ where x_i denotes how long customer i waits.

(a) Show that the log likelihood of λ given $x_1, x_2, \ldots, x_n$ is

$$L = n \log \lambda - \lambda \sum_{i=1}^{n} x_i. \tag{2}$$

[4 marks]

(b) To estimate λ, we follow the maximum likelihood principle and aim to find the λ that maximizes the log likelihood. Suppose we want to use gradient descent to find the maximum. We see that the gradient of L with respect to λ is

$$\frac{\partial}{\partial \lambda} L = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i. \tag{3}$$

Is L concave with respect to λ? [Hint: If L is concave in λ that means $-L$ is convex in λ.] *[5 marks]*

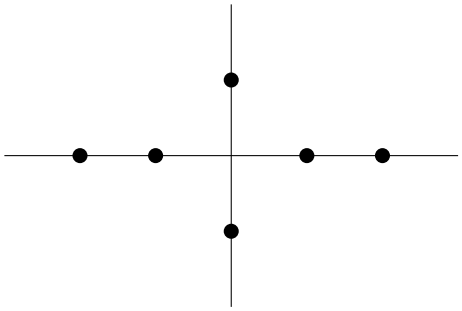
(c) In this particular example, we don't actually need to do gradient descent to know the optimal λ. Solve

$$\frac{\partial}{\partial \lambda} L = 0 \tag{4}$$

to obtain the optimal λ. *[3 marks]*

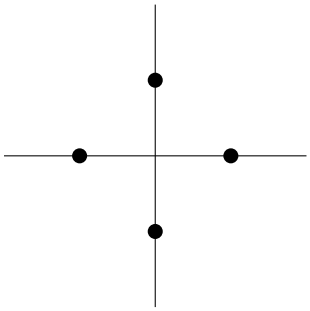
(d) Is the optimal λ from $\frac{\partial}{\partial \lambda} L = 0$ a unique solution? Why or why not? *[3 marks]*

2. (a) Consider the following points in two-dimensional space.



What happens if you do principal component analysis on this data set? Copy the axes and points to your answer sheet and draw two vectors indicting the first and the second principal components. [*3 marks*]

(b) Now consider the following points in two-dimensional space, where the four points are equally spaced.



How many directions would be qualified to be the first principal component? Copy the axes and points to your answer sheet and draw the vectors to indicate the potential directions. [*5 marks*]

(c) Recall that the Lloyd's algorithm (the iterative algorithm for solving k-means taught in class) terminates when the centroids are not updated. If the same set of data points are assigned to a centroid, the centroid is not updated. If no data points are assigned to a centroid, then the centroid is also not updated. We refer to the centroids, when the Lloyd's algorithm terminates, as the solutions of the Lloyd's algorithm or the solutions to which the Lloyd's algorithm converges.

Given the same four-point data set as the one in 2. (b), how many solutions would the Lloyd's algorithm converges to when $k = 2$? Note that the centroids can be initialized any where.

Copy the axes and points to your answer sheet and draw all the solutions. [*7 marks*]

3. Recall the setting of binary classification. A data point is represented as a vector $\mathbf{x} \in \mathbb{R}^d$, and its label is $y \in \{-1, +1\}$. In this question, we look at the exponential loss

$$\ell_{\exp}(\mathbf{x}, y; \mathbf{w}) = e^{-y\mathbf{w}^\top \mathbf{x}}, \tag{5}$$

for a linear classifier with the weight $\mathbf{w} \in \mathbb{R}^d$.

(a) Suppose we have a training set $\{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$. Show that the exponential loss on the training set

$$L = \frac{1}{n} \sum_{i=1}^{n} e^{-y_i \mathbf{w}^\top \mathbf{x}_i} \tag{6}$$

is convex in $\mathbf{w}$. [*7 marks*]

(b) Show that the exponential loss $\ell_{\exp}$ is an upper bound of the zero-one loss

$$\ell_{01}(\mathbf{x}, y; \mathbf{w}) = \begin{cases} 1 & \text{if } y\mathbf{w}^\top \mathbf{x} < 0 \\ 0 & \text{otherwise} \end{cases} \tag{7}$$

for any $\mathbf{x}$, y, and $\mathbf{w}$. [*3 marks*]

(c) Suppose we achieve an average exponential loss of 0.3 on the training set. What can we say about the prediction accuracy on the training set? [*3 marks*]

(d) During training (when minimizing the training loss with respect to $\mathbf{w}$), is it possible to achieve a zero-one loss of 0 but a positive exponential loss on a training set? If you think this could happen, should you stop training once the zero-one loss reaches 0? Why or why not? [*7 marks*]