FOR INTERNAL SCRUTINY (date of this version: 22/2/2024)

UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR10086

Monday 23rd December 1963

20:00 to 23:29

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 a NOTES PERMITTED examination: candidates may consult up to THREE A4 pages (6 sides) of notes. CALCULATORS MAY NOT BE USED IN THIS EXAMINATION.**

Year 3 Courses

Convener: ITO-Will-Determine
External Examiners: ITO-Will-Determine

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

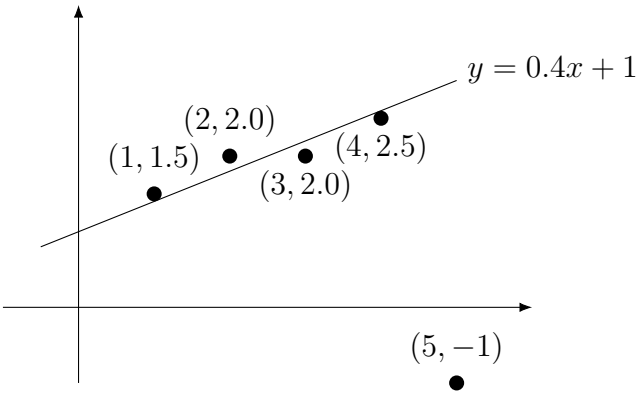
FOR INTERNAL SCRUTINY (date of this version: 22/2/2024)

1. Below are a few questions regarding linear regression.

   (a) We are approached by a customer who wants to optimize the loss function

   $$L(w, b) = \sum_{i=1}^{n} \min((wx_i + b - y_i)^2, 2), \tag{1}$$

   where (x_i, y_i) is one of the n samples. We need to perform gradient descent to find the best w and b. The customer provides their data points and a line $y = wx + b$, where $w = 0.4$ and $b = 1$, as their estimate.



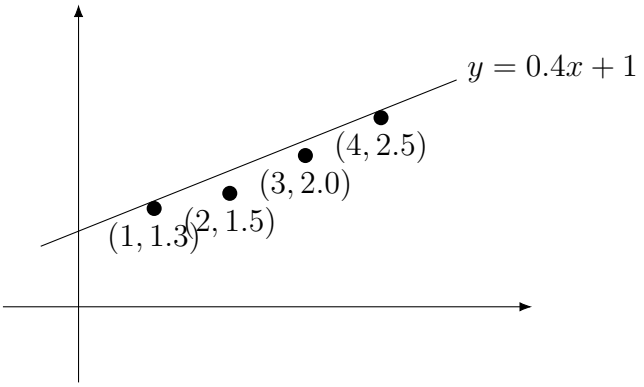
   The customer thinks that the point $(5, -1)$ is probably an outlier, but is unsure whether to remove it or not. How much would the gradient $\frac{\partial L}{\partial w}$ change (when $w = 0.4$) if we remove the data point $(5, -1)$? *[4 marks]*

   Is the loss function L convex in w? Why or why not? *[6 marks]*

   (b) We are approached by another customer who wants to optimize the loss function

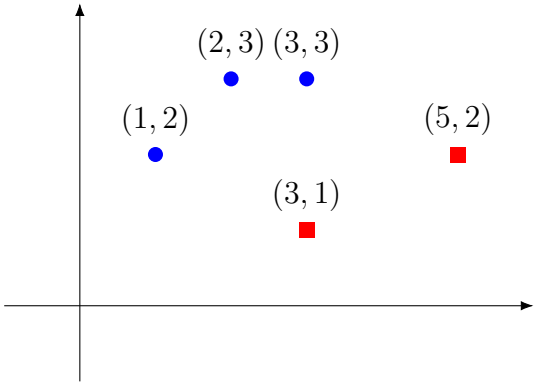
   $$L(w, b) = \sum_{i=1}^{n} \max((wx_i + b - y_i)^2 - 2, 0), \tag{2}$$

   where (x_i, y_i) is one of the n samples. We need to perform gradient descent to find the best w and b. The customer provides their data points and a line $y = wx + b$, where $w = 0.4$ and $b = 1$, as their estimate.



The customer is worried that the line does not seem to be a good fit as all the points are on one side of the line. What is the gradient $\frac{\partial L}{\partial w}$ when $w = 0.4$? Based on the gradient, does the customer need to worry? *[4 marks]*

2. Below is a 2D data set. Points with a label $+1$ are labeled ●, and points with a label -1 are labeled ■. Answer the following questions.



   (a) Give a line that has a zero-one loss of 0. In particular, the line should be in the form of $y = ax + b$, where you need to specify the values of a and b. *[3 marks]*

   In addition, show which side the 5 points lie with respect to the line. The working needs to have the halfspaces that the points belong to. Conclude that this line perfectly separates the two classes. *[3 marks]*

   (b) What would be the decision boundary if we run SVM on this data set? In particular, the line should be in the form of $y = ax + b$, where you need to specify the values of a and b. *[3 marks]*

   In addition, list the support vectors of this classifier, and use the support vectors to measure the margin of this classifier. What is the margin achieved by this classifier? *[3 marks]*

3. Suppose we want to design a few normalization layers for a neural network library, and we need to derive their backward functions for backpropagation. We will use L to denote our loss function. Answer the following questions.

   (a) We first start with a centering layer. Specifically, we want to design a function g such that

$$g(\mathbf{x}) = g\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}\right) = \begin{bmatrix} x_1 - \bar{x} \\ x_2 - \bar{x} \\ \vdots \\ x_d - \bar{x} \end{bmatrix}, \tag{3}$$

where $\bar{x} = \frac{1}{d}\sum_{i=1}^{d} x_i$. Find the matrix A such that $g(\mathbf{x}) = A\mathbf{x}$ and conclude that g is linear in $\mathbf{x}$. *[4 marks]*

If we denote $g_i(\mathbf{x})$ to be the i-th element of the output of g, derive $\frac{\partial g_i}{\partial x_j}$ and conclude that the backward function is

$$\frac{\partial L}{\partial x_j} = \sum_{i=1}^{d} \frac{\partial L}{\partial g_i}\left(\mathbb{1}_{i=j} - \frac{1}{d}\right), \tag{4}$$

where $\mathbb{1}_c$ is 1 if c is true and 0 otherwise. *[4 marks]*

(b) Instead of just centering, we can also divide the result by the standard deviation. Specifically, we want to design a function g such that

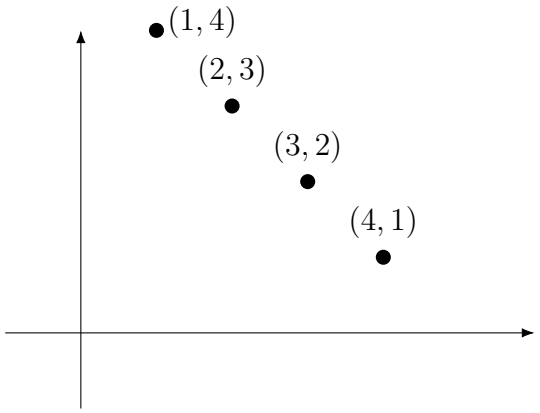
$$g(\mathbf{x}) = g\left(\begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_d \end{bmatrix}\right) = \begin{bmatrix} \frac{x_1-\bar{x}}{\sigma} \\ \frac{x_2-\bar{x}}{\sigma} \\ \vdots \\ \frac{x_d-\bar{x}}{\sigma} \end{bmatrix}, \tag{5}$$

where $\bar{x} = \frac{1}{d}\sum_{i=1}^{d} x_i$ and $\sigma = \sqrt{\frac{1}{d}\sum_{i=1}^{d}(x_i - \bar{x})^2}$. Show that

$$\frac{\partial L}{\partial x_j} = \sum_{i=1}^{d} \frac{\partial L}{\partial g_i}\left(\frac{\mathbb{1}_{i=j} - 1/d}{\sigma} - \frac{(x_i-\bar{x})(x_j-\bar{x})}{\sigma^2}\frac{1}{d\sigma}\right). \tag{6}$$

[8 marks]

4. One of your friends is trying to perform PCA on the following data set, but they **forget** to center the data.



(a) Since this is a 2D data set, after PCA, your friend would get two directions, with the first one having a larger eigenvalue. What is the first direction they would get if they forget to center the data? In particular, the direction should be in the form (a, b), where you need to specify a and b and the the norm of the direction should be 1. *[2 marks]*

(b) What would the first PCA direction be if they center the data? [*2 marks*]

(c) In general, suppose the first PCA direction is v. Show that if we apply a linear transformation A to the data, the first PCA direction becomes Av, as long as the linear transformation satisfies $A^\top A = I$. [*4 marks*]