FOR INTERNAL SCRUTINY (date of this version: 11/7/2024)

UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

INFR10086

Monday 23 $^{\underline{\text{rd}}}$ 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

1. For this question, we will look at properties of a two-layer neural network with rectified linear units (ReLUs).

   (a) A multilayer perceptron typically uses the sigmoid function
$$\sigma(x) = \frac{1}{1 + \exp(-x)} \tag{1}$$
   as the activation function. Show that the sigmoid function is *not* convex. [*4 marks*]

   (b) A rectified linear unit (ReLU) is an activation function of the form
$$\mathrm{ReLU}(x) = \begin{bmatrix} \max(0, x_1) \\ \max(0, x_2) \\ \vdots \\ \max(0, x_d) \end{bmatrix}. \tag{2}$$
   Show that ReLU is convex. [*4 marks*]

   (c) For n functions $f_1, \dots, f_n$, in which $f_i \in \mathbb{R}^d \to \mathbb{R}$, a non-negative weighted sum of them is a function g, such that
$$g(x) = \lambda_1 f_1(x) + \lambda_2 f_2(x) + \cdots + \lambda_n f_n(x), \tag{3}$$
   for all $x \in \mathbb{R}^d$, where $\lambda_1, \dots, \lambda_n \geq 0$. Show that for n convex functions $f_1, \dots, f_n$, in which $f_i \in \mathbb{R}^d \to \mathbb{R}$ for $i = 1, \dots, n$, their non-negative weighted sum is convex. [*4 marks*]

   (d) Consider a two-layer neural network of the form
$$f(x) = w^\top \mathrm{ReLU}(Vx). \tag{4}$$
   This neural network is parameterized by w and V.

   i. Show that regardless of what w is, this network is convex in w. [*2 marks*]

   ii. Show that when w is element-wise non-negative, i.e., $w_1, \dots, w_d \geq 0$, this network is convex in V. [*6 marks*]

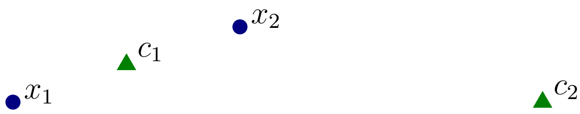
2. Consider the following 2D data set that containts two points x_1 and x_2 (labeled ●).

   (a) If the centroids do not change after further k-means updates, we say that the centroids have reached a local optimum.

   Suppose we initialize k-means with the two centroids c_1 and c_2 (labeled ▲ in the figure below), one of which is exactly at the center of the two points while the other is significantly further away from both points.

[QUESTION CONTINUES ON NEXT PAGE]

[QUESTION CONTINUES FROM PREVIOUS PAGE]



Show that this initialization is a local optimum of k-means. [*4 marks*]

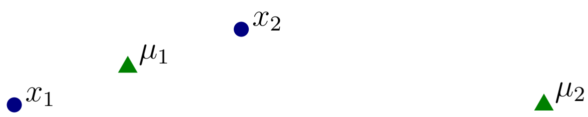
(b) Suppose we initialize k-means with the two centroids c_1 and c_2 (labeled ▲ in the figure below).

Where would the centroids be if we run k-means until it reaches a local optimum? [*4 marks*]

(c) Based on the above results, which local optimum has a better k-means objective? Can we conclude that all local optima of the k-means objective are the global optimum? [*3 marks*]

(d) When training a Gaussian mixture model (GMM) with expectation maximization (EM), if the mean vectors do not change after further updates, we say that EM have reached a local optimum.

Suppose we initialize a two-component GMM with two mean vectors μ_1 and μ_2 (labeled ▲ in the following figure), one of which is exactly at the center of the two points while the other is significantly further away from both points.



Show that this initialization is *not* a local optimum of EM. [*4 marks*]

3. In this question, we will look at the connection between linear regression and the Gaussian distribution.

Recall that a 1D Gaussian distribution $\mathcal{N}(\mu, \sigma^2)$ has a density function

$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right). \tag{5}$$

In linear regression, we assume that $y \sim \mathcal{N}(w^\top x, 1)$, where w is the weight vector. For simplicity, there is no bias term.

[QUESTION CONTINUES ON NEXT PAGE]

[QUESTION CONTINUES FROM PREVIOUS PAGE]

(a) Given an i.i.d. training set $(x_1, y_1), \ldots, (x_n, y_n)$, each of which follows $y_i \sim \mathcal{N}(w^\top x_i, 1)$, show that the log-likelihood is

$$\log \prod_{i=1}^{n} p(y_i|x_i) = -\frac{n}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^{n}(y_i - w^\top x_i)^2. \tag{6}$$

[4 marks]

(b) Given a training set $(x_1, y_1), \ldots, (x_n, y_n)$, discuss how maximizing the log-likelihood is equivalent to solving the mean-square error. *[2 marks]*

(c) Consider a data set $(x_1, y_1), \ldots, (x_n, y_n)$, where $x_i = x_0$. In other words, all samples in the data set share the same input while having potentially different output.

i. Show that

$$\nabla_w \log \prod_{i=1}^{n} p(y_i|x_i) = \left(\sum_{i=1}^{n} y_i - n w^\top x_0\right) x_0. \tag{7}$$

[5 marks]

ii. Show that the optimal solution in this case is any w that satiesfies

$$w^\top x_0 = \frac{1}{n}\sum_{i=1}^{n} y_i. \tag{8}$$

[4 marks]