FOR INTERNAL SCRUTINY (date of this version: 18/11/2022)

UNIVERSITY OF EDINBURGH

COLLEGE OF SCIENCE AND ENGINEERING

SCHOOL OF INFORMATICS

MACHINE LEARNING

November 2020

23:50 to 23:59

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

THIS EXAMINATION WILL BE MARKED ANONYMOUSLY

FOR INTERNAL SCRUTINY (date of this version: 18/11/2022)

1. Suppose we have data points $\mathbf{x}_1, \ldots, \mathbf{x}_n$, each of which is a $\mathbb{R}^d$ vector. We assume that the data points follow the generative process

$$\mathbf{x}_i \sim \mathcal{N}(W\mathbf{h}_i, I), \tag{1}$$

where $\mathbf{h}_1, \ldots, \mathbf{h}_n$ is another set of known vectors.

As a reminder, when we write $\mathbf{x} \sim \mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$, it means

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\boldsymbol{\Sigma}|^{1/2}} \exp\left(-\frac{1}{2}(\mathbf{x}_i - \boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}_i - \boldsymbol{\mu})\right). \tag{2}$$

You might also need

$$\frac{\partial \mathbf{x}^\top \mathbf{a}}{\partial \mathbf{x}} = \mathbf{a} \qquad\qquad \frac{\partial \mathbf{a}^\top A \mathbf{a}}{\partial \mathbf{a}} = (A + A^\top)\mathbf{a}. \tag{3}$$

a) Show that the log likelihood is

$$L = \sum_{i=1}^{n} \left[-\frac{d}{2}\log(2\pi) - \frac{1}{2}\|\mathbf{x}_i - W\mathbf{h}_i\|^2\right]. \tag{4}$$

[3 marks]

b) Suppose $\mathbf{h}_1, \ldots, \mathbf{h}_n$ are vectors produced by a neural network, and we need to train the network with back-propagation. What is gradient with respect to $\mathbf{h}_i$ for any $i = 1, \ldots, n$? *[4 marks]*

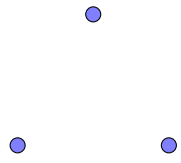
c) Derive the Hessian of the likelihood with respect to $\mathbf{h}_i$. Show that the negative of the Hessian is positive semidefinite, meaning that the log likelihood is concave in $\mathbf{h}_i$. *[4 marks]*

d) If $W^\top W$ is invertible, show that the optimal solution for $\mathbf{h}_i$ is $(W^\top W)^{-1} W^\top \mathbf{x}_i$. *[4 marks]*

2. Answer the following questions about clustering and dimensionality reduction.

   a) Suppose there is a data set of two classes. The two classes are well separated, meaning that there is a hyperplane that can separate the two classes. Suppose we use PCA to project the data points to a 2-dimensional space. We do not observe distinct clusters in this 2-dimensional space. Discuss how this can happen. Use simple plots to justify your answer. *[4 marks]*

   b) Consider the following data set with 3 points in 2-dimensional space. Each point is of the same distance to the other two points.



   Recall that when the k-means algorithm converges, the centroids and the assignments of each point to the centroids would not change after subsequent updates. If we run k-means with $k = 2$ on this data set, how many solutions would the k-means algorithm converges to? Draw the solutions that the k-means algorithm converges to, including the three points and using crosses $\times$ to indicate where the two centroids are. *[4 marks]*

   c) A Gaussian mixture model (GMM) with K components assumes a one-hot, latent vector $\mathbf{z} = [z_1, \ldots, z_K]$, meaning that $\sum_{k=1}^{K} z_k = 1$ and $z_k \in \{0, 1\}$ for $k = 1, \ldots, K$. The element $z_i = 1$ when the i-th component is chosen. The values $\pi_1, \ldots, \pi_K$ are the prior probability for choosing one of the K components. In other words, the probability of $\mathbf{z}$ can be written as

$$p(\mathbf{z}) = \prod_{k=1}^{K} \pi_k^{z_k}, \tag{5}$$

   Component i in GMM is a Gaussian with mean $\boldsymbol{\mu}_i$ and covariance $\boldsymbol{\Sigma}_i$. The conditional probability of observing a sample $\mathbf{x}$ from a component $\mathbf{z}$ is

$$p(\mathbf{x}|\mathbf{z}) = \prod_{k=1}^{K} \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k)^{z_k}. \tag{6}$$

   Show that the marginal distribution $p(\mathbf{x})$ is of form

$$p(\mathbf{x}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k). \tag{7}$$

   [4 marks]

   d) When training a GMM with 2 components on the three-point data set above, you realize the log likelihood becomes `nan`. What could be the reason for this? *[3 marks]*

3. Recall that in binary classification, the input type is $\mathbb{R}^d$ and the output type is $\{+1, -1\}$. A linear classifier can be written as

$$f(\mathbf{x}) = \begin{cases} +1 & \text{if } \mathbf{w}^\top \phi(\mathbf{x}) \geq 0 \\ -1 & \text{if } \mathbf{w}^\top \phi(\mathbf{x}) < 0 \end{cases} \tag{8}$$

where ϕ is a feature function.

You are working on a term project with two other friends. One is excited about using hinge loss to train linear classifiers, and wants to understand the loss better. The hinge loss is defined as

$$\ell_{\text{hinge}}(\mathbf{w}; \mathbf{x}, y) = \max(0, 1 - y\mathbf{w}^\top \phi(\mathbf{x})). \tag{9}$$

You all know that if $\ell(\mathbf{w})$ is a convex function, then

$$\ell(\alpha \mathbf{w}_1 + (1-\alpha)\mathbf{w}_2) \leq \alpha \ell(\mathbf{w}_1) + (1-\alpha)\ell(\mathbf{w}_2) \tag{10}$$

for any $0 \leq \alpha \leq 1$.

a) Show that the hinge loss is an upper bound on the zero-one loss

$$\ell_{01}(\mathbf{w}; \mathbf{x}, y) = \mathbb{I}[y\mathbf{w}^\top \phi(\mathbf{x}) < 0], \tag{11}$$

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

b) Is the hinge loss differentiable? If you think it is, derive the derivative of hinge loss with respect to w. If you think it is not, provide a point where the derivative does not exist, and provide a subgradient at that point. [*3 marks*]

c) A linear classifier is sometimes written as

$$g(\mathbf{x}) = \operatorname*{argmax}_{y \in \{+1,-1\}} \mathbf{w}^\top \psi(\mathbf{x}, y). \tag{12}$$

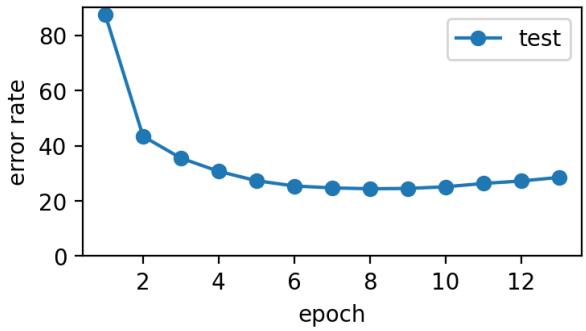
Show that if we choose $\psi(\mathbf{x}, y) = \frac{1}{2} y \phi(\mathbf{x})$, then $g(\mathbf{x}) = f(\mathbf{x})$ for all $\mathbf{x}$. [*5 marks*]

d) The other friend discovers a new loss function

$$\ell_{\text{new}}(\mathbf{w}; \mathbf{x}, y) = \max_{\hat{y} \in \{+1,-1\}} \left(\mathbb{I}[y\hat{y} < 0] - \mathbf{w}^\top \psi(\mathbf{x}, y) + \mathbf{w}^\top \psi(\mathbf{x}, \hat{y}) \right). \tag{13}$$

Show that this loss function is the same as the hinge loss. [*4 marks*]

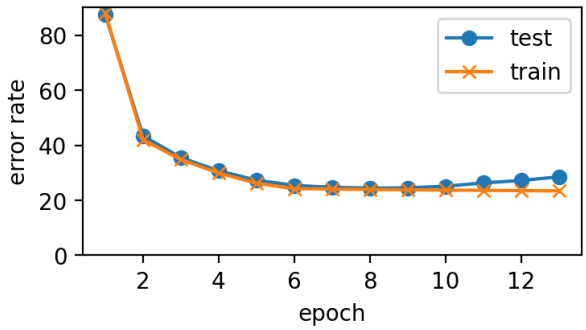
We now know the properties of the loss, and a model can be trained. Once the training is done, you obtain the following plot, the error rates on the test set.



After looking at the plot, one of your friends concludes that the model must be overfitting, while the other friend is not convinced.

e) What, if any, would you conclude from this plot? Why? [*2 marks*]

To further study this problem, you plot the error rates on the training set.



After looking at the plot, one friend thinks this is definitely overfitting, and suggests you explore regularizers. The other friend is not convinced, and suggests you tune the step size of the training algorithm to see if the models are underfitting.

f) What would you do in this case? Why? [*2 marks*]