**The Australian National University**
**School of Computing**


COMP3670/6670 Introduction to Machine Learning
Semester 2, 2022

Final Exam

- Write your name and UID on the first page (you will be fine if you forget to write them).
- This is an open book exam. You may bring in any materials including electronic and paper-based ones. Any calculators (programmable included) are allowed. No communication devices are permitted during the exam.
- Reading time: 30 minutes
- Writing time: 180 minutes
- For all the questions, write your answer CLEARLY on papers prepared by yourself.
- There are totally 8 pages (including the cover page)
- Points possible: 100
- This is not a hurdle.
- When you are asked to provide a justification to your answer, if your justification is incorrect, you will get 0.

## Section 1. Linear Algebra and Matrix Decomposition (13 points)

1. Let $\mathbf{A} = \begin{bmatrix} a & b & c \\ d & e & f \\ g & h & i \end{bmatrix}$ be a square $3 \times 3$ matrix, such that the determinant of $\mathbf{A}$ is 2.

   What is the determinant of the following matricies?

   (1). (1 pt)
   $$\begin{bmatrix} 3a & b & c \\ 3d & e & f \\ 3g & h & i \end{bmatrix}$$

   (2). (1 pt) $\mathbf{A}^3$

   (3). (1 pt) $\mathbf{A} + \mathbf{A}$.

   (4). (1 pt) $\mathbf{A} = \begin{bmatrix} d & e & f \\ g & h & i \\ a & b & c \end{bmatrix}$

2. (7 pt) We say that $\mathbf{A}$ is an *isometry* if it preserves Eucledian distance, e.g. for all
   $$||\mathbf{Ax}||_2 = ||\mathbf{x}||_2$$
   Let $\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$ be an arbitrary $2 \times 2$ matrix. Find the set of all $\mathbf{A} \in \mathbb{R}^{2\times 2}$ such that

   - $\mathbf{A}$ is an isometry.
   - All entries along the main diagonal are strictly positive.

3. (2 pt) What is the geometric intuition behind multiplying by an isometric matrix? Other than preserving the length of a vector, what sort of transformation is performed?

   Hint: Substitute $\sin\theta$ for the free coefficient in your isometric matricies.

**Section 2. Analytic Geometry and Vector Calculus** (11 points)

1. (5 pt) Let $\mathbf{A} \in \mathbb{R}^{n \times n}$, and $\mathbf{x}, \mathbf{b} \in \mathbb{R}^n$ and define $f : \mathbb{R}^n \to \mathbb{R}$ as

$$f(\mathbf{x}) = (\mathbf{A}\mathbf{x})^T$$

   Compute $\nabla_{\mathbf{x}} f(\mathbf{x})$.

2. (6 pt) Let $\mathbf{A} \in \mathbb{R}^{n \times n}$ be a matrix such that for all $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, we have $\mathbf{x}^T \mathbf{A} \mathbf{y} = \mathbf{x}^T \mathbf{y}$. Prove that $\mathbf{A} = \mathbf{I}$.

**Section 3. Probability** (16 points)

1. (5 pt) Let $X$ and $Y$ be random variables, with outcomes $\{x_1, x_2\}$ and $\{y_1, y_2\}$ respectively. Prove that if $x_1$ is evidence in favour of $y_1$, $(P(Y = y_1 \mid X = x_1) > P(Y = y_1))$ then $x_2$ must be evidence against $y_1$. $(P(Y = y_1 \mid X = x_2) < P(Y = y_1))$. Hint: Try assuming that $x_2$ is not evidence against $y_1$, and derive a contradiction from the laws of probability.

2. There are two types of coins: *real* coins and *fake* (counterfeit) coins. Counterfeit coins flip heads with probability $p$, and tails with probability $1 - p$. Real coins flips heads with probability $q$, and tails with probability $1 - q$. The coins otherwise look totally identical, with no other distinguishing features. Both $p$ and $q$ are fixed, known constants, with $0 < p < q < 1$ (the real coin is more biased towards heads than the fake coin.)

   (1). (4 pt) Show that the probability of the real coin generating a sequence containing $h$ heads when flipped $N$ times is $P(H = h) = \binom{N}{h} q^h (1 - q)^{N-h}$.

   (2). (7 pt) A fake coin and a real coin are placed into a bag, and one is drawn uniformly at random. We flip the coin $N$ times, and observe a sequence with $h$ heads, and $N - h$ tails. Verify the that new probability that the coin is real after observing this sequence is

$$P(coin = real \mid h \text{ heads in } N \text{ flips }) = \frac{q^h (1 - q)^{N-h}}{p^h (1 - p)^{N-h} + q^h (1 - q)^{N-h}}$$

**Section 4. Clustering and Gaussian Mixture Model (GMM)** (16 points)

We describe a new clustering algorithm named token clustering.

Initialization. Same as k-means, it randomly initializes $K$ cluster centers (called tokens in this new method).

E step. It differs from k-means in this step. Specifically, for the first sample, this method finds its nearest token. During this process, if the distance between this sample to its nearest token is below a threshold $h$, this sample will be assigned to it; otherwise, this sample will immediately become a new token, resulting in $K+1$ tokens. For the next sample, find its nearest token among all the tokens (old+new), and repeat this procedure. Iterate until all the samples have been computed.

M Step. It updates each cluster center (token) by computing the mean of its assigned training samples.

Iterate the E and M steps until the change in tokens is below a threshold.

1. (2 pt) In the E-step, the training samples are processed one by one. If the data samples are processed in different orders, will the clustering results be the same? In 3 sentences explain your answer.

2. (3 pt) Compared with kmeans, is token clustering better at dealing with outlier points? An outlier point is dissimilar to all other points. In 3 sentences explain your answer.

3. (4 pt) When threshold $h$ increases from 0 to positive infinite, the total number of tokens will change accordingly. Discuss when token clustering will give more, the same, and fewer clusters than kmeans. Here, we asssume both methods have the same initialization.

Given an image below that contains an object (horse), you want to separate this object from the background. That is, you want to decide for every pixel whether it belongs to the foreground (horse) or background. To this end, you propose to use the Gaussian Mixture Model (GMM).

4. (3 pt) In a few sentences, describe how to use GMM to find the foreground of this image.

5. (2 pt) Would using 2 components be the best option for any kind of images? Explain your answer in 2 sentences.

6. (2 pt) For our image, if you are going to further improve the foreground (horse) segmentation performance, how would you do that? Use 2 sentences to describe your strategy.

**Section 5. Linear Regression** (15 points)

Given a face image, we are interested in predicting the *identity* and *age* of the person of this face. To this end, we collect a training set with $N$ samples: $(\boldsymbol{I}_1, \boldsymbol{y}_1)$, $(\boldsymbol{I}_2, \boldsymbol{y}_2)$,..., $(\boldsymbol{I}_N, \boldsymbol{y}_N)$, where $I_n, n = 1, ..., N$ is the $n$th image, and $\boldsymbol{y}_n$ denotes its label. In the training set, we have many different identities and ages. Some images contain the same person but of different ages; some contain different people of the same age.

1. (2 pt) Describe two methods that can transform this face image into a vector.

2. (1 pt) In 1 sentence describe what label $\boldsymbol{y}_n, n = 1, ..., N$ includes.

3. (1 pt) In 3 sentences, explain how you can train an age estimation model from the training set, and how you perform inference given a test image.

4. Traditionally, we design two different models, one using a loss $L_{\text{ID}}$, the other one using loss function $L_{\text{age}}$, for face identification and age estimation, respectively. But this would increase carbon consumption due to training. To make your system greener, you want to design a *single* model that can perform both face identification and age estimation tasks, given the vector $\boldsymbol{x} \in \mathbb{R}^d$ of a face image.

   1). (1 pt) For this multi-task system, please write down the loss function, which contains a hyperparameter to control the relative importance of the two tasks.

   2). (2 pt) Will the multi-task system have a higher age estimation accuracy than the traditional single-task age estimation model? Please explain your answer in a few sentences. Suppose the hyperparameter is optimally chosen.

   3). (2 pt) Will your answer to the above question change if we replace face identification with mustache prediction (predicting whether a face has mustache)? We assume mustache labels (1 for existence, 1 for non-existence) are provided.

5. Now, forget about face identification; we only focus on age estimation. Based on some property, each training sample can be assigned to one of $K$ groups (for example, each group contains faces of a certain ethnicity). Assume that we use the traditional average squared loss, which is computed on individual training samples. Under this loss, if a certain group has only a few training samples, the age estimation model oftentimes has low accuracy for this group during inference.

   4). (2 pt) In 3 sentences explain why this problem might happen.

   5). (4 pt) Without changing the training set, please design a loss function that can improve model performance on the worst-performing group. Define your notations as needed. Under such a loss function, what might be sacrificed compared with the traditional loss?

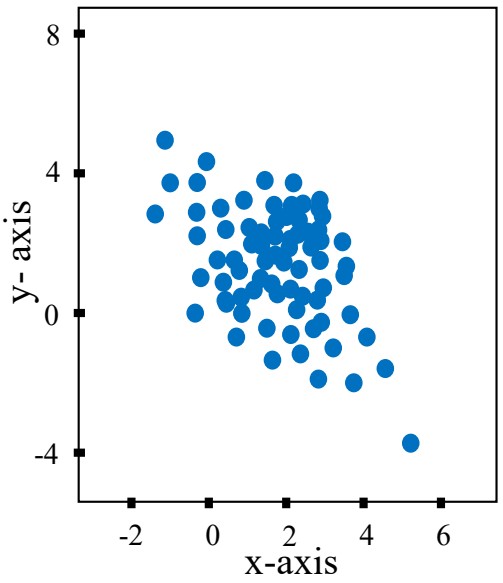
**Section 6. Principal Component Analysis (PCA) and Linear Regression** (14 points)

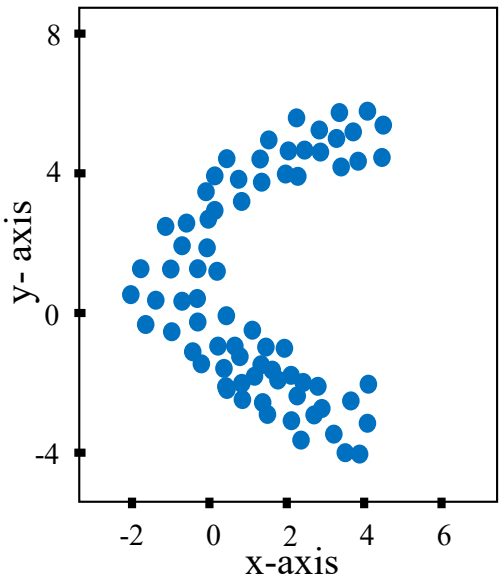
As shown in the figure below, we generate two 2D dataset with size $100 \times 2$.

1. (2 pt) For the dataset *on the left*, are the linear regression result and the first principal component of the same direction? Using two sentences to explain your answer.

2. (2 pt) For the dataset *on the left*, If we project the dataset onto the subspace spanned by the first principal component, will the first principal component and regression result be of the same direction?

3. (2 pt) For the dataset *on the left*, Suppose you add the same offset $[5, -2]^T$ to every point in this dataset, will and how the first principal component change? Explain your answer in 2 sentences.




4. (2 pt) For the dataset *on the right*, will it be helpful to use PCA for dimensionality reduction? In 3 sentences explain your reason.

5. (2 pt) You left multiple all the data points in a dataset by an orthogonal matrix. For the transformed dataset, will the principal components remain the same? Will the eigenvalues of the sample covariance matrix remain the same? For each question, explain your answer in 2 sentences.

6. (4 pt) We perform PCA to the data points before training a classifier. Show a scenario where PCA harms classifier training? Show a scenario where PCA benefits classifier training. You can draw figures, write equations, text descriptions, *etc.*

**Section 7. Classification** (15 points)

You are training a 2-way classifier using the logisitc loss. The training set has three training samples: $(\boldsymbol{x}_1, 1), (\boldsymbol{x}_2, 1), (\boldsymbol{x}_3, -1)$, where $\boldsymbol{x_n} \in \mathbb{R}^d$, $n = 1, 2, 3$, and their class labels are 1, 1, and -1, respectively. During training, we store and view the prediction results (output of the Sigmoid function) after a certain iteration.

$$\hat{\mathbf{Y}}_1 = \begin{bmatrix} 0.6 \\ 0.7 \\ 0.4 \end{bmatrix}$$

1. (1 point) On this training set (with 3 samples), compute the classification accuracy. Briefly write down your steps.

2. (2 points) Does the classification model converge after this iteration? Explain your answer in 3 sentences.

3. (2 points) In a few sentences explain why logistic regression is superior to linear regression for the classification problem.

In $K$-way classification, we make use of $K$ linear functions of the form $y_k(\boldsymbol{x}) = \boldsymbol{\theta}_k^T \boldsymbol{x} + \theta_{k0}$. Vector $[\boldsymbol{\theta}_k^T, \theta_{k0}]$ is also called the prototype of the $k$th class. If its dot product with $\boldsymbol{x}$ (plus a dummy feature) is the largest among all the prototypes, then $\boldsymbol{x}$ will be assigned into this class.

4. Among the $K$ classes, two classes ($i$ and $j$) in the training set are particularly hard to distinguish.

   1). (2 points) Use 2 sentences to interpret this using the prototype concept.

   2). (2 points) Write down an improved loss function that can better distinguish the two classes. Note: you may denote the previous loss function as $L_0$. Other reasonable loss functions should be fine as well.

5. For a training sample $\boldsymbol{x}$ whose ground truth label is class $k$. During training, we obtain its $K$ prediction scores from the $K$ linear functions in a vector $\boldsymbol{s} \in \mathbb{R}^K$. We then **normalize** $\boldsymbol{s}$ by its $\ell_2$ norm, yielding $\tilde{\boldsymbol{s}}$. To compute the loss value of $\boldsymbol{x}$, we would like to compute the squared difference between $\tilde{\boldsymbol{s}}$ and the ground truth label (class $k$).

   1). (2 points) To allow such computation, write down the mathematical format of the ground truth.

   2). (2 points) Use no more than 3 sentences to explain why an normalization step ($\ell_2$ normalization here) is needed.

6. (2 points) The 0-1 loss was designed for classification. Does it work for linear regression? In 3 sentences explain your answer.

——— End of the paper ———