The Australian National University
School of Computing

COMP3670/6670 Introduction to Machine Learning
Semester 2, 2021

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 9 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. (6 points) Show that it is impossible to have a set of three **non-zero** vectors $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ in $\mathbb{R}^2$ such that both of the following properties hold:
     (a) The set $\{\mathbf{v}_1, \mathbf{v}_2\}$ is linearly independent.
     (b) $\mathbf{v}_3 \cdot \mathbf{v}_1 = 0$ and $\mathbf{v}_3 \cdot \mathbf{v}_2 = 0$, where $\cdot$ is the standard Euclidean dot product.
  2. (7 points) Consider an arbitrary matrix $\mathbf{A} \in \mathbb{R}^{2\times 2}$

     $$\mathbf{A} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}$$

     We would like $\mathbf{A}$ to satisfy the following properties:
     (a) $\mathbf{A}$ has a **unique** eigenvalue λ.
     (b) All the entries of $\mathbf{A}$ are positive.

     Does there exist a matrix $\mathbf{A}$ satisfying both constraints? If so, state with proof the set of all matrices that satisfy the condition. If not, prove none exist." (There is no mark for the "yes/no" answer.)

- **Section 2. Analytic Geometry and Vector Calculus** (12 points)

  1. (6 points) Find all matrices $\boldsymbol{T} \in \mathbb{R}^{2\times 2}$ such that for any $\boldsymbol{v} \in \mathbb{R}^2$,

  $$\boldsymbol{v}^T\boldsymbol{v} = (\boldsymbol{T}\boldsymbol{v})^T\boldsymbol{v} = (\boldsymbol{T}\boldsymbol{v})^T\boldsymbol{T}\boldsymbol{v}$$

  2. (6 points) Let $\boldsymbol{x}, \boldsymbol{a} \in \mathbb{R}^{n\times 1}$, and define $f : \mathbb{R}^{n\times 1} \to \mathbb{R}$ as

  $$f(\boldsymbol{x}) = 2\sin(3\boldsymbol{x}^T\boldsymbol{A}\boldsymbol{x})$$

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

- **Section 3. Probability** (15 points)

  Consider the follow scenario. A bag contains three coins, coin A, coin B, coin C. The coins have numbers on each side. $A = \{2, 4\}, B = \{1, 3\}, C = \{2, 3\}$. All the coins are fair.

  Two coins are selected from the bag uniformly at random, and flipped. Let S be a random variable that represents the sum of the two numbers shown, with outcomes in $\mathbb{N}$. Let X be a random variable that represents the two coins chosen, with outcomes $\{AB, BC, AC\}$.

  1. (3 points) For each possible outcome for X, what is the conditional probability mass function $p(S = s|X = x)$ for the sum S?
  2. (2 points) How likely is it that the sum of the two coins is 5?
  3. (4 points) The coins are flipped, and the sum $S = 5$. How likely is it that one of the coins flipped was coin C? How does this compare to the likelihood that one of the coins flipped was C, before observing the sum? Explain your observations.
  4. (6 points) The same coins as the previous question are flipped **again**, and this time the sum $S = 7$. How likely is it now that one of the coins flipped was C? Compare your result to the likelihood that one of the coins flipped was C, prior to any observations. (Hint: You now have two observations, a flip which gave a sum of $S = 5$, and a flip which gave a sum of 7).

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

  Suppose you have N data points in $\mathbb{R}$: $x_1, x_2, ..., x_n, x_{n+1}, ..., x_N$. Now you want to partition them into two clusters C_1 and C_2. First, you assign $x_1, x_2, ..., x_n$ into C_1 and assign $x_{n+1}, ..., x_N$ into C_2. After that, you want to calculate the new cluster centers μ_1 and μ_2 of C_1 and C_2. For each cluster, your objective is to minimise the averaged squared distances between each data point and its assigned center.

  1. (3 points) Calculate μ_1 and μ_2 that achieve your objective (M-step). (Only showing result without the optimisation process will lead to 0 mark.) Hint: after the assignment, you can only use the first n points to calculate μ_1, and the rest points to calculate μ_2.
  2. (2 points) Are μ_1 and μ_2 a global minimum solution to this M-step (calculating the means of clusters)? Use no more than 3 sentences to explain your answer.

  Given μ_1 and μ_2 calculated in Question 1 above, you now want to update the assignment (E-step). This time again you aim to minimise the sum of squared distances between each data point and its center.

  3. (3 points) What is your optimal assignment strategy (you can only assign a data point to one cluster, *i.e., hard assignment*)? Why is it optimal for your aim? (You can use whatever is helpful to illustrate, such as figures and maths. Only showing the result without a clear demonstration/proof process with lead to 0 mark)
  4. (2 points) Is this assignment a global minimum solution to the E-step under hard assignment? Use no more than 3 sentences to explain your answer.

  Following the above questions, we do further explorations. The GMM performs soft assignment, *i.e.,* assigning a data point into multiple clusters, and this assignment is accompanied by responsibilities. Now, we want to explore a similar scheme in k-means. Specifically, we define

$$r_{m2} = \frac{(x_m - \mu_1)^2}{(x_m - \mu_1)^2 + (x_m - \mu_2)^2}, r_{m1} = \frac{(x_m - \mu_2)^2}{(x_m - \mu_1)^2 + (x_m - \mu_2)^2},$$

  where r_{m2} denotes how likely x_m belongs to C_2, and r_{m1} denotes how likely x_m belongs to C_1.

  1. (2 point) Suppose x_m is assigned to C_1 under *hard assignment*, then under *soft assignment*, which cluster will bear a higher responsibility for x_m? Use two sentences to explain.
  2. (4 points) Given μ_1 and μ_2, when looking at a single data point x_m, does soft assignment have a lower loss value than the hard assignment? Prove it. Hint: your loss function will be the sum of the squared distance from x_m to each center multiplied by the "responsibility" $r_{mk}, k = 1, 2$.

- **Section 5. Linear Regression** (13 points)

1. (4 points) In least squares linear regression, our empirical risk is

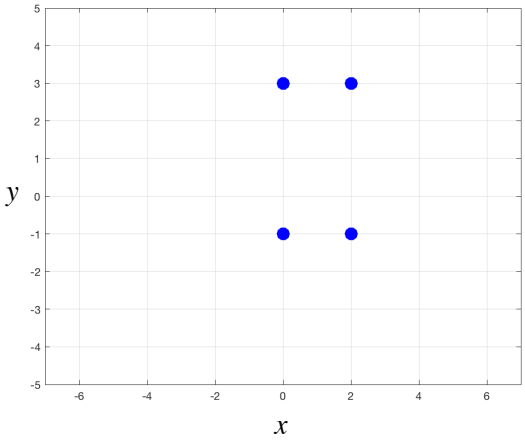
$$\boldsymbol{R} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \boldsymbol{\theta}^T\boldsymbol{x}_n)^2,$$

where $\boldsymbol{x}_n \in \mathbb{R}^d$ is a training sample, $y_n \in \mathbb{R}$ is its label. $\boldsymbol{\theta}$ contains the model parameters. Now we use a sigmoid function in this empirical risk, *i.e.,*

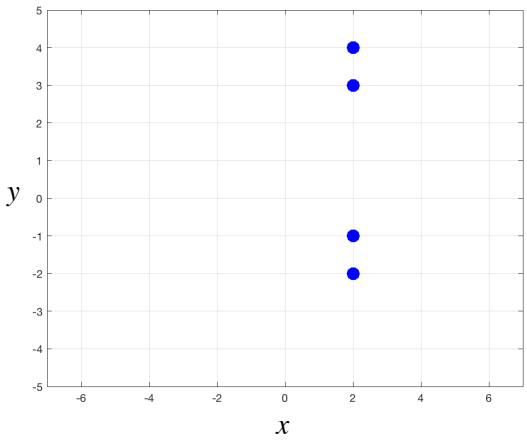
$$\boldsymbol{R} = \frac{1}{N}\sum_{n=1}^{N}(y_n - \text{sigmoid}(\boldsymbol{\theta}^T\boldsymbol{x}_n))^2.$$

In no more than 4 sentences, state the reason why using sigmoid function in this way is not desirable.

2. Now we have four data points shown in the figure below. We use the least square error in regression.



(a)



(b)

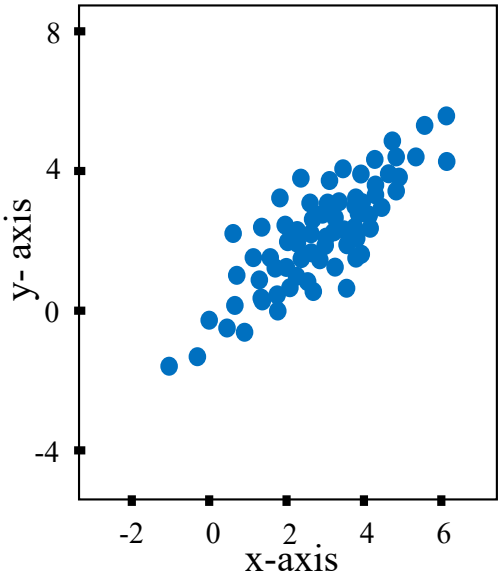
   1) (3 points) In (a), draw the linear regression output and the first principal component. Are they of the same direction/orientation? Use no more than 3 sentences to explain why.
   2) (3 points) In (b), draw the regression output, using language to describe when necessary.

   Note: you can use any drawing software (PowerPoint, pdf editor, etc) to directly do it on figure; otherwise, approximately draw the four points on your paper and then draw the regression/PCA lines.

3. (3 points) In the lecture slides, to derive compact formula, we include a dummy feature $x_0 = 1$ into the sample vector $\boldsymbol{x}$, and correspondingly $\boldsymbol{\theta}$ includes the coefficient θ_0 of this dummy feature. Now I want to remove this dummy feature and its coefficient from the regression model. Will the resulting model give better test accuracy after training (sufficient training samples, no over-fitting)? In 3 sentences justify your answer.

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

1. (4 points) Give the main steps for principal component analysis. For example, given $\boldsymbol{X} \in \mathbb{R}^{D\times N}$, list the steps to reduce dimensions to $D-1$, and calculate a new $\boldsymbol{X} \in \mathbb{R}^{(D-1)\times N}$.
2. (2 points) As shown in the figure below, we generate a 2D dataset with size 100×2. The eigenvalues and corresponding eigenvectors are given under the figure. Compute θ, which is defined as the angle ($\leq 90°$) between the first principal component and $x-$axis.



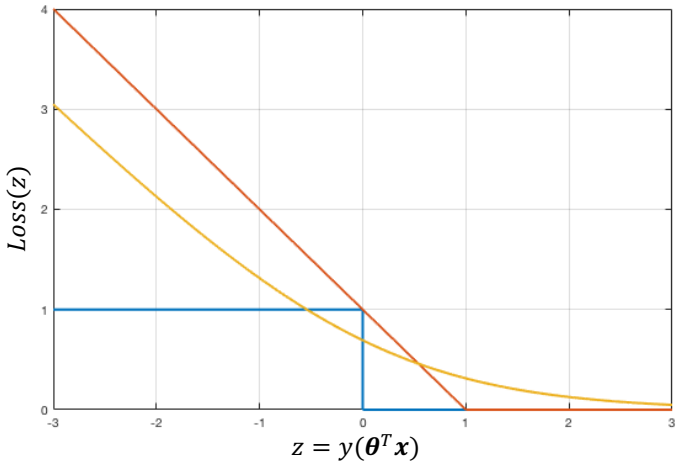
$$\lambda_1 = 316, \lambda_2 = 1683$$

$$v_1 = [-0.71, 0.71]^T, v_2 = [-0.71, -0.71]^T$$

3. (2 points) Following the previous question, if we reduce the original data to one-dimensional data using PCA, calculate the percentage of information loss.
4. (3 points) Suppose we have sufficient training data. If we use PCA to first project the training data onto a few principal components, *i.e.,* performing dimension reduction, will this lower-dimensional training set give a better linear regression model than the original higher-dimensional training data? In three sentences, justify your answer.
5. (3 points) Suppose you perform the following PCA operations on your d-dimensional data points of which the covariance matrix has d distinct eigenvalues. Operation 1: project the data onto a j-dimensional space, and then project the j-dimensional data onto a g-dimensional space. Operation 2: project the d-dimensional data directly onto the g-dimensional space. Here $d > j > g$. Are the PCA results of the two operations the same? In no more than 5 sentences explain your answer.

- **Section 7. Classification** (17 points)

  In your lecture slides, the zero-one loss, hinge loss and logistic loss (with basis e) can be plotted as the following figure. The loss functions are written as the function of $z = y(\boldsymbol{\theta}^T \boldsymbol{x})$. Here, $y \in \{-1, 1\}$ is the label of data sample $x \in \mathbb{R}^d$, where d is the feature dimension. $\boldsymbol{\theta} \in \mathbb{R}^d$ contains the model parameters.



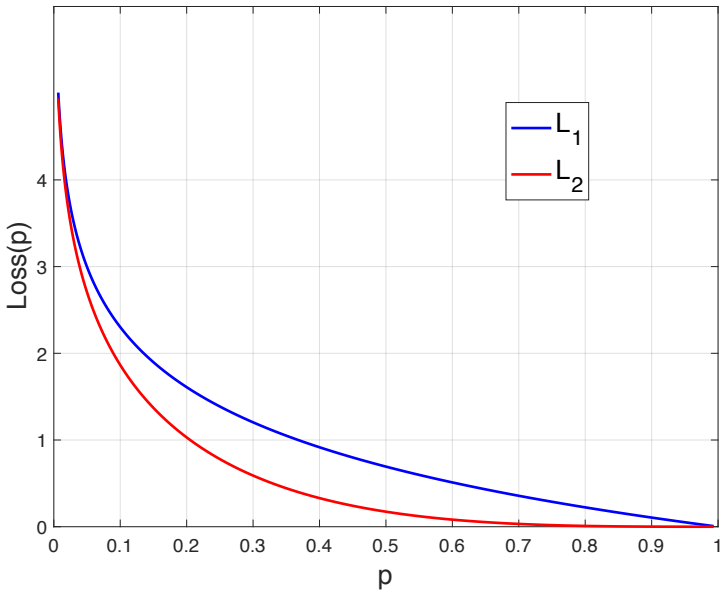
  1. (1 point) What does a small z mean? What does a large z mean? Use one sentence to answer each.
  2. (3 points) If we use the least squares loss function to train a classifier, please write down the loss function 1) using $\boldsymbol{\theta}^T \boldsymbol{x}$ as argument, and 2) using z as argument.
  3. (1 point) In the figure above, draw the least squares loss function in the same figure with the other three loss functions.

     Note: you can use any drawing software (PowerPoint, pdf editor, etc) to directly do it on figure; otherwise, roughly replicate this figure on paper and then add the curve for the least squares loss.
  4. (2 points) From what you have drawn, explain in 2 sentences why least squares is not a good choice for classifier training.
  5. (2 points) Following the slides, we now transform z into probability:

$$p = p(y|\boldsymbol{x}) = \text{sigmoid}(z) = \frac{1}{1+e^{-z}} = \frac{1}{1+e^{(-y\boldsymbol{\theta}^T\boldsymbol{x})}}.$$

     Now rewrite the logistic loss using p as argument. For your convenience, the logistic loss is $\text{Loss}_L(z) = \log(1+e^{-z})$ when using z as argument.
  6. (2 points) In the loss function $\text{Loss}_L(p)$ you just derived, what does a small p mean? what does a large p mean?
  7. (2 points) Suppose you are working on a two-way classification task. During classifier training, while some samples can be easily and successfully classified into the correct class, others are rather difficult and oftentimes misclassified. You want to solve this issue by designing a loss function that allows the classifier to focuse more on such difficult and easily misclassified samples. We have two loss functions for you to choose, all expressed as a function of p: $L_1(p)$ and $L_2(p)$. We draw these loss functions in the figure below against p. Which loss function would you like to choose? Use no more than 3 sentences to justify your answer.
  8. You want a K-class classifier, where the number of classes $K > 2$. According to the lecture slides, this classifier contains K linear functions

$$g_k(\boldsymbol{x}) = \boldsymbol{\theta}_k^T \boldsymbol{x} + \theta_{k0}, k = 1, ..., K.$$



Here, $g_k(\boldsymbol{x}) \in \mathbb{R}$ characterises how likely sample $\boldsymbol{x}$ belongs to the kth class. Besides, vector $[\boldsymbol{\theta}_k^T, \theta_{k0}]^T$ is also called the prototype of the kth class. If a test sample is closest to the prototype of the ith class, it will be classified into the ith class. Suppose for some reason, that you find it very undesirable to train the classifier using methods like gradient descent. In other words, you find using what we've learned in the classification lecture results in a poor classifier.

1) (2 points) In two sentences, give a potential reason why the classifier trained by gradient descent gives very poor performance (suppose your make no mistakes in programming and math). Trivial answers (*e.g.,* my computer is down) will receive 0.

2) (2 points) Propose a way to calculate the prototype vectors that could give decent classification performance without training the classifier.

——— End of the paper ———