The Australian National University
Research School of Computer Science

Australian National University

COMP3670/6670 Introduction to Machine Learning
Semester 2, 2020

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. (6 points) Let $\{\mathbf{v}_1, \mathbf{v}_2\}$ be linearly independent vectors in $\mathbb{R}^n$. Let $\mathbf{v}_3$ be a vector in $\mathbb{R}^n$ that does not lie in the span of $\mathbf{v}_1, \mathbf{v}_2$. Prove that $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is linearly independent.
2. (7 points) Consider the matrix

$$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}$$

Find its eigenvalues. What does this matrix geometrically do when applied to a vector? Explain how this relates to the set of eigenvalues for this matrix.

- **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 $\boldsymbol{v} \in \mathbb{R}^2$,

  $$T(\boldsymbol{v}) \cdot \boldsymbol{v} = 0$$

  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

  $$f(\boldsymbol{x}) = \boldsymbol{x}^T \boldsymbol{a}\boldsymbol{a}^T \boldsymbol{x}$$

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

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

  Consider the following scenario. I flip a fair coin.
  If the coin comes up heads, I roll a fair 4 sided die (with sides $\{1, 2, 3, 4\}$), and then I tell you the result of rolling the die.
  If the coin comes up tails, I roll a fair 6 sided die (with sides $\{1, 2, 3, 4, 5, 6\}$), and then I tell you the result of rolling the die.
  Let X denote the number I tell you.

  1. (3 points) What is the set of all possible outcomes $\mathcal{X}$ for X?
  2. (4 points) Compute $P(X = x)$ for all $x \in \mathcal{X}$.
  3. (4 points) I tell you that $X = 1$. How likely is it that the coin flipped heads?
  4. (4 points) We repeat the above experiment, but this time whatever die is selected, is rolled twice. I inform you that the outcome for both rolls was a 1. How likely is it that the coin flipped was heads?

- **Section 4. Clustering and Gaussian Mixture Model (GMM)** (15 points)

  Both Kmeans and GMM can be viewed as aiming to find $\boldsymbol{\theta}$ to optimise $p(\boldsymbol{\mathcal{X}}|\boldsymbol{\theta})$. Here, $\boldsymbol{\mathcal{X}}$ is the dataset, and $\boldsymbol{\theta}$ is related to the model. Answer the following questions.

  1. (2 points) In kmeans, use no more than 2 sentences to describe what $\boldsymbol{\theta}$ contains.

  2. (3 points) In kmeans, use no more than 2 sentences to describe the probabilistic meaning of $p(\boldsymbol{\mathcal{X}}|\boldsymbol{\theta})$.

  3. (3 points) Assume that samples in $\boldsymbol{\mathcal{X}}$ are from 3 classes. After training a GMM with 3 components on $\boldsymbol{\mathcal{X}}$, we use this GMM as a classifier to predict which class a new sample $\boldsymbol{x}$ belongs to. In no more than 3 sentences, describe the prediction process. (Use math symbols where relevant; you do not have to explain the symbols if they are same with lecture slides, *e.g.*, $\boldsymbol{\mu}$).

  4. (3 points) Is it correct to say that the kmeans method enables us to find $\boldsymbol{\theta}$ that minimises $p(\boldsymbol{\mathcal{X}}|\boldsymbol{\theta})$? Explain your answer in 2 sentences.

  5. (4 points) Suppose we have the following 10 data points. They are partitioned into two classes, red and blue. Which model could generate this partition, kmeans only, GMM only, both, or neither? Explain your answer in three sentences. (Note: where GMM is relevant, the classification principal is similar to Question 3 above, except that this question has two classes.)

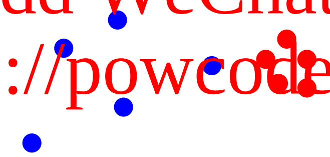
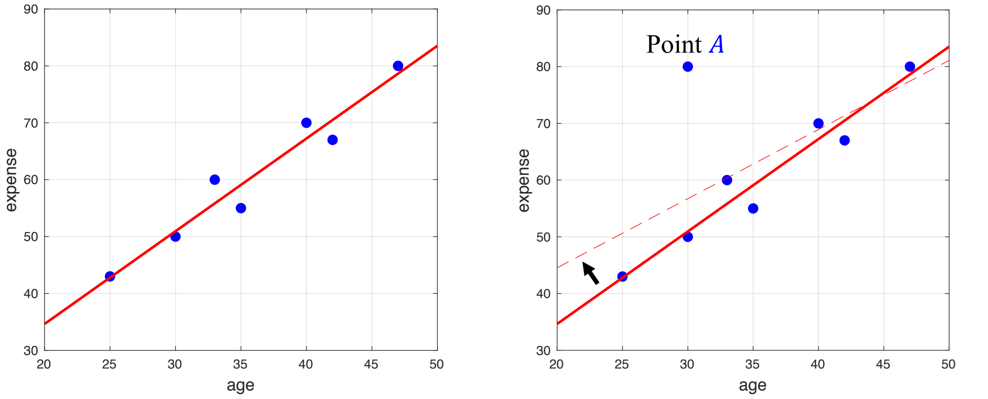


Figure 1: 10 data points divided into two classes, blue and red.

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

  You are doing a machine learning internship at a bank, analysing user age and their daily expense. You collected seven samples and plotted them in Fig. 2(a).



(a) Regression with normal data (b) Regression with normal and abnormal data

Figure 2: Linear regression with normal and abnormal data.

1. (3 points) From your observation of Fig. 2(a), use 1 sentence to describe the relationship between user age and expense.

After obtaining Fig. 2(a), you collected a new data point A. Together with the previous seven points, you do linear regression for a second time and obtain the dashed line in Fig. 2(b).

2. (4 points) Generally when adding new samples to the dataset, it is expected that the fitted line will be different. In our example, there is quite a **large** difference between the new model (dashed line) and the old model (solid line). In two sentences, explain why the change is large. (Hint: you don't have to explain why there is a "change". Focus on "large".)

3. (6 points) You originally used the squared error, *i.e.*, for the nth training sample $(\boldsymbol{x}_n, y_n)$,

$$l(\boldsymbol{x}_n, y_n, \boldsymbol{\theta}) = (y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n)^2,$$

   where $\boldsymbol{x}_n$ is the feature of the sample, y_n is the label, and $\boldsymbol{\theta}$ contains the model parameters. Your supervisor tells you that Point A is an outlier and that it is best to exclude its impact on your model. Write down an amended loss function that can achieve this goal. Explain how it excludes the impact of outliers on your linear model. Note: you will get partial marks if your loss function can merely alleviate the impact of A.

- **Section 6. Principal Component Analysis (PCA) and Linear Regression** (20 points)

  We are given a centered dataset, $(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)$, where $x_n \in \mathcal{R}$, $y_n \in \mathcal{R}$, and N is the number of samples. "Centered" means $\sum_{n=1}^{N} x_n = 0$, and $\sum_{n=1}^{N} y_n = 0$. Now for this dataset, we apply linear regression and PCA. For linear regression, our model is $y = \theta x + \theta_0 = \boldsymbol{\theta x}$, where we treat y_n as labels and x_n as the feature. For PCA, we obtain the first and second principal components: $\boldsymbol{pc}_1$ and $\boldsymbol{pc}_2$. An example is shown in Fig. 3.

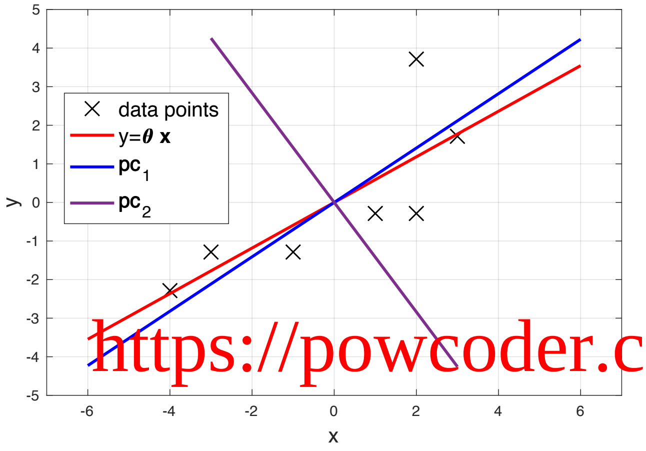


Figure 3: Linear regression with normal and abnormal data.

1. (3 points) Are $\boldsymbol{pc}_1$ and $\boldsymbol{pc}_2$ orthogonal? Explain your answer in two sentences.

2. (4 points) In usual cases, the regression output $\boldsymbol{\theta}$ is not in the same direction with $\boldsymbol{pc}_1$ (and $\boldsymbol{pc}_2$). Explain why $\boldsymbol{\theta}$ and $\boldsymbol{pc}_1$ are usually different in direction. You can use whatever is relevant to help you illustrate, such as figures or maths. (Hint: differences of PCA and linear regression in their optimisation objective.)

3. (4 points) On your paper, draw an example dataset for which $\boldsymbol{\theta}$ and $\boldsymbol{pc}_1$ are of the same direction. Your figure should contain the x-axis, the y-axis, at least 3 data points, as well as $\boldsymbol{\theta}$ and $\boldsymbol{pc}_1$ (the latter two should be overlapping). If necessary, write the coordinates of the data points.

4. (5 points) Let $\boldsymbol{a} = [x_1, x_2, ..., x_N]^T \in \mathcal{R}^N$, and $\boldsymbol{b} = [y_1, y_2, ..., y_N]^T \in \mathcal{R}^N$. On this *centered* dataset, show that when $\boldsymbol{a}^T\boldsymbol{b} = 0$, the regression output is the x-axis. (We assume the MSE as loss function)

5. (4 points) Continuing from Question 4, calculate the covariance matrix of this dataset (you do not have to do standardization). When $\sum_{i=1}^{N} x_i^2 > \sum_{i=1}^{N} y_i^2$, show that the first principal component $\boldsymbol{pc}_1$ is horizontal.

- **Section 7. Classification** (12 points)

  You have developed a linear classifier to classify the sentiment of a sentence into positive and negative. Assume that positive sentences and negative sentences have equal numbers in both training and testing sets. Your classifier obtains an accuracy of 40% on the test data.

  1. (2 points) Is this classifier meaningful? Explain your answer in two sentences.

  2. (3 points) Without re-training the classifier, use three sentences to describe how you improve the previous classifier and why it becomes better.

  PCA is a useful technique to project data samples onto a lower-dimensional subspace that preserves data variance. Oftentimes, it is used to preprocess features before training a classifier.

  3. (4 points) You have four data points of two classes. Their class labels (A or B) and coordinates are listed below.

     - Class A: (10, 1) and (-10, 1)
     - Class B: (10, -1) and (-10, -1)

     For this case, is it a useful step to project the data onto the first principle component before training a classifier? If yes, draw the decision boundary after the projection. If no, briefly explain your answer.

  4. (3 points) Explain why PCA is helpful for classifier training in many real-world cases.

——— End of the paper ———