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.

   Solution. Assume for a contradiction that $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ are not linearly independent. Then there exists constants c_1, c_2, c_3, not all zero, such that

   $$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 = 0$$

   Now, if $c_3 = 0$, then the above equation becomes

   $$c_1\mathbf{v}_1 + c_2\mathbf{v}_2 = 0$$

   and therefore $\{\mathbf{v}_1, \mathbf{v}_2\}$ are linearly dependent, a contradiction. So, we have that $c_3 \neq 0$. Hence

   $$\begin{aligned} c_1\mathbf{v}_1 + c_2\mathbf{v}_2 + c_3\mathbf{v}_3 &= 0 \\ c_1\mathbf{v}_1 + c_2\mathbf{v}_2 &= -c_3\mathbf{v}_3 \\ \frac{-c_1}{c_3}\mathbf{v}_1 + \frac{-c_2}{c_3}\mathbf{v}_2 &= \mathbf{v}_3 \end{aligned}$$

   and hence $\mathbf{v}_3$ lies in the span of $\mathbf{v}_1, \mathbf{v}_2$, a contradiction.

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.

   Solution. We form the characteristic equation $\det(\mathbf{A} - \lambda\mathbf{I})$

   $$\det(\mathbf{A} - \lambda\mathbf{I}) = \det\begin{bmatrix} -\lambda & -1 \\ 1 & \lambda \end{bmatrix} = \lambda^2 - (-1)(1) = \lambda^2 + 1$$

   Setting it equal to zero and solving,

   $$\lambda^2 = 1 \Rightarrow \lambda = \pm\sqrt{-1}$$

   We find no eigenvalues, as we cannot square root a negative number (at least not in the real numbers). This matrix can be seen to perform a rotation of 90 degrees about the origin,

   $$\begin{bmatrix} 0 & -1 \\ 1 & 0 \end{bmatrix}\begin{bmatrix} x \\ y \end{bmatrix} = \begin{bmatrix} -y \\ x \end{bmatrix}$$

   which means that no vector in $\mathbb{R}^2$ will be transformed into a scalar multiple of itself. This is why we see no real eigenvalues.

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

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

Solution. Let

$$\mathbf{T} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}, \mathbf{v} = \begin{bmatrix} x \\ y \end{bmatrix}.$$

Then,

$$\begin{aligned} T(\mathbf{v}) \cdot \mathbf{v} = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} &= 0 \\ \begin{bmatrix} ax + by \\ cx + dy \end{bmatrix} \cdot \begin{bmatrix} x \\ y \end{bmatrix} &= 0 \\ (ax + by)x + (cx + dy)y &= 0 \end{aligned}$$

Choose $x = 0$, then

$$(a0 + by)0 + (c0 + dy)y = 0 \Rightarrow dy^2 = 0$$

Since $dy^2 = 0$ needs to hold for all y, this is only possible when $d = 0$. Now, choose $y = 0$, then

$$(ax + b0)x + (cx + d0)0 = 0 \Rightarrow ax^2 = 0$$

Since $ax^2 = 0$ needs to hold for all x, this is only possible when $a = 0$. Substituting this into the above,

$$(0x + by)x + (cx + 0y)y = 0$$

$$bxy + cxy = 0$$

Since this needs to hold for all x and for all y, it must be the case that $b = -c$. Hence,

$$\{\mathbf{T} \in \mathbb{R}^{2\times 2} : \forall \mathbf{v} \in \mathbb{R}^2, (\mathbf{T}\mathbf{v}) \cdot \mathbf{v} = 0\} = \left\{ \begin{bmatrix} 0 & \alpha \\ -\alpha & 0 \end{bmatrix} : \alpha \in \mathbb{R} \right\}$$

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

Solution. There are a few ways to do this. Students could take the derivative with respect to a component of $\mathbf{x}$, or could use product rule, or could recognize that

$$\boldsymbol{x}^T \boldsymbol{a}\boldsymbol{a}^T \boldsymbol{x} = \boldsymbol{x}^T \boldsymbol{a}\boldsymbol{x}^T \boldsymbol{a} = (\boldsymbol{x}^T \boldsymbol{a})^2$$

and use chain rule. I will use the latter approach. Let $h(\mathbf{x}) = \mathbf{x}^T \boldsymbol{a}$ and $g(x) = x^2$. Then $f(\mathbf{x}) = g(h(\mathbf{x}))$, and so

$$\frac{df}{d\mathbf{x}} = \frac{dg}{dh}\frac{dh}{d\mathbf{x}}$$

Then, $\frac{dg}{dh} = 2h$ and $\frac{dh}{d\mathbf{x}} = \boldsymbol{a}^T$, by identities provided in lecture slides. Hence,

$$\frac{df}{d\mathbf{x}} = 2h\boldsymbol{a}^T = 2\boldsymbol{x}^T \boldsymbol{a}\boldsymbol{a}^T.$$

as required.

- **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?
     Solution. We either roll a 4-sided die, and get a number from 1 to 4, or roll a 6-sided die, and get a number from 1 to 6. Hence $\mathcal{X} = \{1, 2, 3, 4, 5, 6\}$.
  2. (4 points) Compute $P(X = x)$ for all $x \in \mathcal{X}$.
     Solution. First consider the case where $x \in \{1, 2, 3, 4\}$. Let C denote the outcome of the coin, with $\mathcal{C} = \{\text{heads}, \text{tails}\}$ denoting the set of outcomes for the coin.

     $$\begin{aligned} P(X = x) &= \sum_{c \in \mathcal{C}} P(X = x \mid C = c) P(C = c) \\ &= P(X = x \mid C = \text{heads}) P(C = \text{heads}) + P(X = x \mid C = \text{tails}) P(C = \text{tails}) \\ &= 1/4 \cdot 1/2 + 1/6 \cdot 1/2 = 5/24 \end{aligned}$$

     Then consider the case where $x \in \{5, 6\}$, which will be the same as before, but now $P(X = x \mid C = \text{heads}) = 0$, as the four-sided die can't roll 5's or 6's.

     $$\begin{aligned} P(X = x) &= \sum_{c \in \mathcal{C}} P(X = x \mid C = c) P(C = c) \\ &= P(X = x \mid C = \text{heads}) P(C = \text{heads}) + P(X = x \mid C = \text{tails}) P(C = \text{tails}) \\ &= 0 \cdot 1/2 + 1/6 \cdot 1/2 = 1/12 \end{aligned}$$

     Hence,

     $$P(X = x) = \begin{cases} 5/24 & x \in \{1, 2, 3, 4\} \\ 1/12 & x \in \{5, 6\} \end{cases}$$

  3. (4 points) I tell you that $X = 1$. How likely is it that the coin flipped heads?
     Solution. We apply Bayes rule.

     $$\begin{aligned} &P(C = \text{heads} \mid X = 1) \\ &\qquad = \frac{P(X = 1 \mid C = \text{heads}) P(C = \text{heads})}{P(X = 1)} \\ &\qquad = \frac{1/4 \cdot 1/2}{5/24} = 3/5 \end{aligned}$$

  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?
     Solution. Each die roll is i.i.d.

     $$P(C = \text{heads} \mid \text{rolled one twice}) = \frac{P(\text{rolled one twice} \mid C = \text{heads}) P(C = \text{heads})}{P(\text{rolled one twice})}$$

     Now, $P(\text{rolled one twice} \mid C = \text{heads}) = P(X = 1 \mid C = \text{heads})^2 = 1/16$, and $P(\text{rolled one twice} \mid C = \text{tails}) = P(X = 1 \mid C = \text{tails})^2 = 1/36$

     $$\begin{aligned} &P(\text{rolled one twice}) \\ &= P(\text{rolled one twice} \mid C = \text{heads}) P(C = \text{heads}) + P(\text{rolled one twice} \mid C = \text{tails}) P(C = \text{tails} \\ &= P(X = 1 \mid C = \text{heads})^2 (1/2) + P(X = 1 \mid C = \text{tails})^2 (1/2) \\ &= 1/16 \cdot 1/2 + 1/36 \cdot 1/2 = 13/288 \end{aligned}$$

Hence,

$$P(C = \text{heads} \mid \text{rolled one twice}) = \frac{1/16 \cdot 1/2}{13/288} = 9/13$$

- **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.
     Solution. $\boldsymbol{\theta}$ contains the coordinates of centroids, as well as the cluster ID that each sample is assigned to. Deduct 0.5 mark if answering the number of clusters.

  2. (3 points) In kmeans, use no more than 2 sentences to describe the probabilistic meaning of $p(\boldsymbol{\mathcal{X}}|\boldsymbol{\theta})$.
     Solution. Given model parameters $\boldsymbol{\theta}$, the probability that dataset $\boldsymbol{\mathcal{X}}$ is generated by this model.

  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.*, $\boldsymbol{\mu}$).
     Solution. We obtain three Gaussian distributions from training. We compute $p(\boldsymbol{x}|\boldsymbol{\theta}_k)$, the probability that the test sample $\boldsymbol{x}$ is generated by the kth Gaussian distribution. We pick the Gaussian distribution that gives the largest probability as the label of this test sample.

  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.
     Solution. Incorrect. First, we aim to maximise instead of minimise $p(\boldsymbol{\mathcal{X}}|\boldsymbol{\theta})$. Second, kmeans enables us to find a local maximum instead of a global one.

  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.)

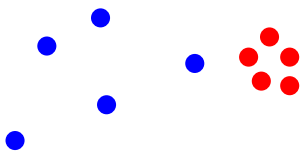
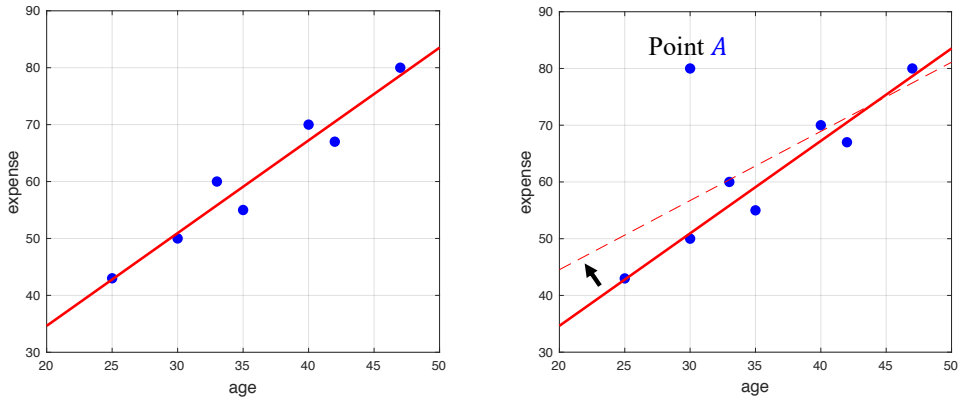
     

     Figure 1: 10 data points divided into two classes, blue and red.

     Solution. GMM only. It cannot be generated by kmeans because of the right most point is closer to the center of the red class, so it will be classified as red. GMM can generate this figure because the blue class has a larger variance under GMM.

- **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.
   Solution. Expense increases linearly as age increases.

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".)
   Solution. 1) Point A deviates a lot from the other points, thus rendering a large gradient/loss asking the model to be closer to it. 2) There are limited number of samples in this dataset, so the big gradient from A causes the model to move a lot.

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.
   Solution.
$$l(\boldsymbol{x}_n, y_n, \boldsymbol{\theta}) = \min((y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n)^2 - m, 0).$$
   m is a hyperparameter. From Fig. 1, its value should be approximately between 100 and 400 (it's not necessary for a student to make this estimation). There could be other equivalent ones.
   Explanation. For a proper value of m, the outlier point A will have a large $(y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n)^2$, thus causing term $(y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n)^2 - m$ to be greater than 0. In this case, there will be 0 loss value. For other points, $(y_n - \boldsymbol{\theta}^T \boldsymbol{x}_n)^2 - m$ will be less than 0, thus preserving the value after the min operator.

- **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 $y = \theta x + \theta_0 = \boldsymbol{\theta}\boldsymbol{x}$, 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. 5.

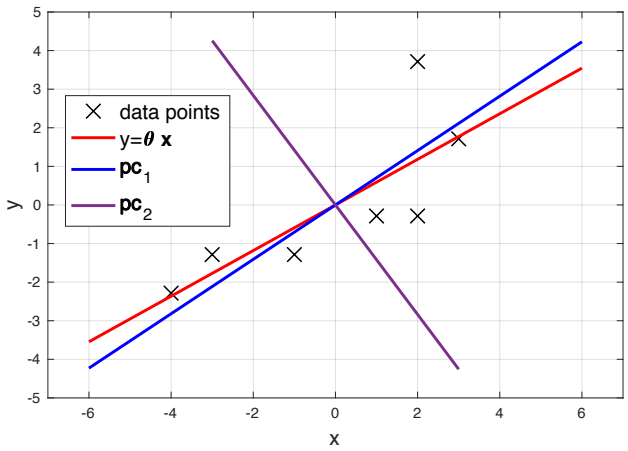


Figure 3: Linear regression with normal and abnormal data.

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

   Solution. Yes. The covariance matrx of this 2-D dataset has two different eigenvalues because the data distribution is oval. Eigenvectors corresponding to different eigenvalues are orthogonal.

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.)

   Solution. Both linear regression and PCA aim to find a model to minimise a loss function. Linear regression aims to find a subspace such that the predicted output and the model output are close. In comparison, PCA wants to find a subspace such that the orthogonal projection of samples onto this subspace is minimised. See figure below for an illustration.

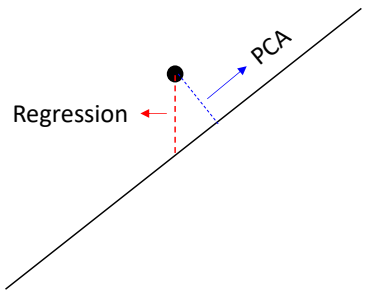


Figure 4: Difference between PCA and regression.

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.

Solution. Many scenarios. For example, if three different points are on the same line, $\boldsymbol{\theta}$ and $\boldsymbol{pc}_1$ overlap. Two examples are shown below.

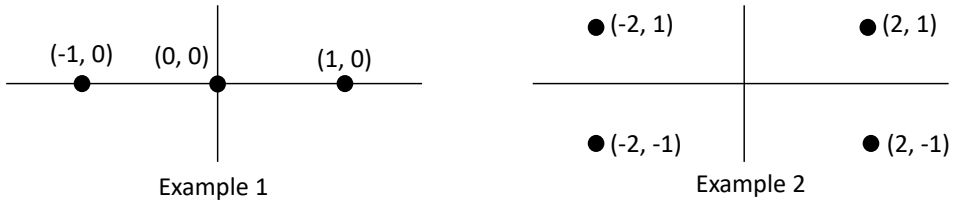


Figure 5: Two examples of overlapping PCA and regression outputs.

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)

Solution. The model of linear regression is

$$y = \theta_1 x + \theta_0 = \boldsymbol{\theta}\boldsymbol{x}.$$

The closed form solution of linear regression is

$$[\theta_1, \theta_0] = \boldsymbol{\theta} = (\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y},$$

where $\boldsymbol{X} = \begin{bmatrix} x_1 & x_2 & ... & x_N \\ 1 & 1 & ... & 1 \end{bmatrix}^T$, $\boldsymbol{y} = [y_1, y_2, ..., y_N]^T$.

Because $\boldsymbol{a}^T\boldsymbol{b} = 0$ and $y_1 + y_2 + ... + y_N = 0$, we obtain that

$$\boldsymbol{X}^T\boldsymbol{y} = [\sum_{i=1}^{N} x_i y_i, \sum_{i=1}^{N} y_i] = [0, 0].$$

Therefore, the output of linear regression is the x-axis, *i.e.*, $y = 0$.

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.

Solution. The covariance matrix is calculated as,

$$\boldsymbol{S} = \frac{1}{N}\begin{bmatrix} x_1 & x_2 & ... & x_N \\ y_1 & y_2 & ... & y_N \end{bmatrix}\begin{bmatrix} x_1 & x_2 & ... & x_N \\ y_1 & y_2 & ... & y_N \end{bmatrix}^T = \frac{1}{N}\begin{bmatrix} \sum_{i=1}^{N} x_i^2 & \sum_{i=1}^{N} x_i y_i \\ \sum_{i=1}^{N} x_i y_i & \sum_{i=1}^{N} y_i^2 \end{bmatrix} = \frac{1}{N}\begin{bmatrix} \sum_{i=1}^{N} x_i^2 & 0 \\ 0 & \sum_{i=1}^{N} y_i^2 \end{bmatrix}.$$

When $\sum_{i=1}^{N} x_i^2 \neq \sum_{i=1}^{N} y_i^2$, the two eigenvalues are

$$\lambda_1 = \sum_{i=1}^{N} x_i^2, \lambda_2 = \sum_{i=1}^{N} y_i^2.$$

When $\sum_{i=1}^{N} x_i^2 > \sum_{i=1}^{N} y_i^2$, the eigenvector corresponding to λ_1 is the principal component $\boldsymbol{pc}_1$. We calculate this eigenvector by solving the following equation system,

$$\frac{1}{N}\begin{bmatrix} \sum_{i=1}^{N} x_i^2 & 0 \\ 0 & \sum_{i=1}^{N} y_i^2 \end{bmatrix}\boldsymbol{b} = \lambda_1\boldsymbol{b}.$$

Substituting λ_1 with $\sum_{i=1}^{N} x_i^2$, we can obtain the eigenvector as

$$\boldsymbol{b} = [1, 0]^T,$$

which 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.
     Solution. No. Random guess would have 50% accuracy, so the classifier is not meaningful.

  2. (3 points) Without re-training the classifier, use three sentences to describe how you improve the previous classifier and why it becomes better.
     Solution. Negate every prediction, *i.e.*, if the prediction is negative then change it to positive and vice versa. Then we get a classifier that has 60% accuracy.

  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.
     Solution. No. The first principal component is the x-axis. If we project the points on the x-axix, (10, 1) and (10, -1) will overlap, and (-10, 1) and (-10, -1) will overlap. The two classes are no longer separable. In this question, the informative dimension is trimmed off.

  4. (3 points) Explain why PCA is helpful for classifier training in many real-world cases.
     Solution. In real-world cases, features have a higher dimension and there might not be sufficient training data. PCA allows us to remove the redundant features, alleviate overfitting and improve training efficiency. A student would get full marks if the answer contains "remove redundant features".

——— End of the paper ———