Sample Exam Questions: 2018-2019

Question 1

[10 marks]

Let A be a $m * n$ matrix where m is 80 million and n is 4096. The eigen decomposition of $A^T A$ shows that 4086 eigen values are zero. The remaining ten eigen values (in descending order) are:

$$[10^2, 9^2, 8^2, 7^2, 6^2, 5^2, 4^2, 3^2, 2^2, 1^2]$$

1. What are the non-zero singular values of A? Show the calculation steps?
2. We would like to create a representation that can capture all the information of A. How many dimensions at least this representation should have and why?

Question 2

[10 marks]

A student has decided to use the k-NN algorithm for a binary classification problem.

1. Consider the training set with ten data points shown in Fig Q2a. Use the k-NN algorithm with three nearest neighbours, i.e. 3-NN, to determine the class of an unknown data point $(x = 4, y = 4)$. Clearly indicate how you determined the nearest neighbours.

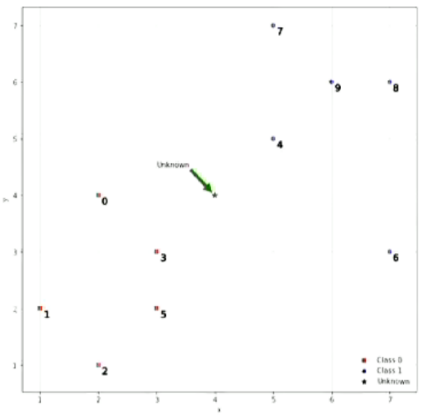


Figure 1: Fig. Q2a

Question 3

[10 marks]

A factory has three machines A, B, and C. From the history, we know that the defect rates of the three machines are respectively 2%, 1%, and 3%. In last year, the three machines respectively produced 2000, 3000, and 5000 products. If two products are chosen at random from these total products and is found to be defective, what is the probability that it was produced by machine A?

Question 4 [10 marks]

Given the dataset below, use the Naive Bayes classifier to predict whether the last borrower will default or not. Show your calculations as well as the final result.

Id	Home Owner	Marital Status	Annual income	Default borrower
1	Yes	Single	High	No
2	No	Married	Medium	No
3	No	Single	High	No
4	Yes	Married	Medium	No
5	No	Divorced	Low	Yes
6	No	Married	High	No
7	Yes	Divorced	Low	Yes
8	No	Single	Medium	No
9	No	Single	Medium	Yes
10	Yes	Divorced	Low	?

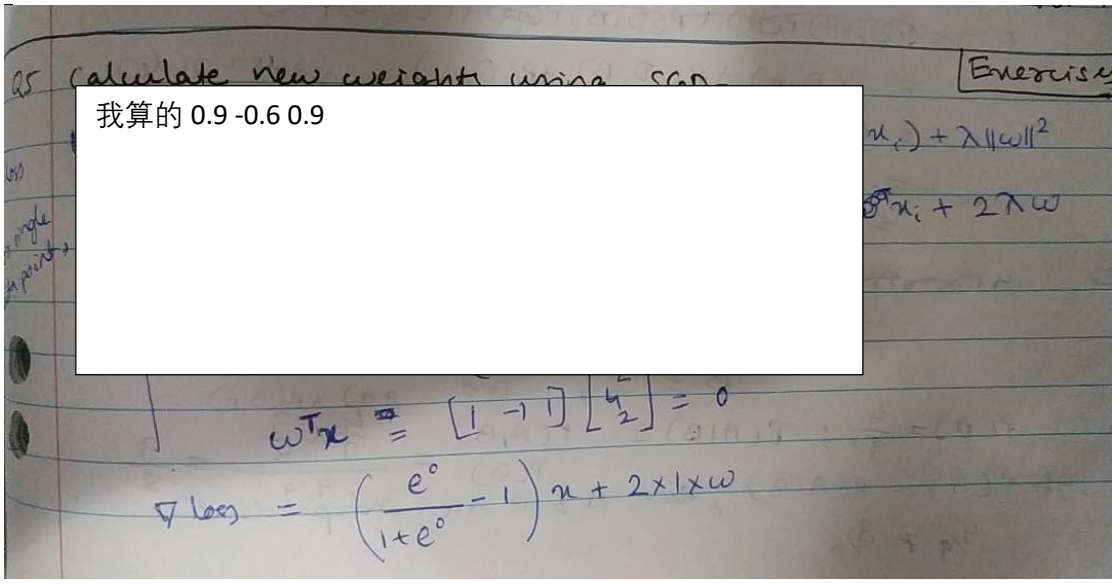
Question 5 [10 marks]

We are going to optimise the regularised logistic regression by employing the stochastic gradient descent method (SGD). The objective function is as follows:

$$\frac{1}{N}\sum_{i=1}^{N}(\log(1+e^{w^T x_i}) - y_i w^T x_i) + \lambda\|w\|^2$$

The initial w: $w_0^T = [0,0,0], \lambda = 1$, and the learning rate $\mu = 0.1$. Given that the first randomly selected example is $(x_0^T = [0,0,0], y_0 = 0)$ and the second is $(x_1^T = [1,0,0], y_1 = 1)$, what are the new weight values after two updates?

Let $f(x) = \log x, g(x) = e^x$. Then, their first derivatives are as:$f'(x) = \frac{1}{x}$ and $g'(x) = e^x$.



Question 6 [20 marks]

You have collected three datasets with two dimensional points for binary classification. Your ultimate goal is to classify the point $x^* = (x_1^*, x_2^*) = (0, 2.5)$. You are provided with the following datasets: A, B, and C.

- For dataset A: The positive observations are $(2, 2), (2, 3)$, and $(3, 2)$. The negative observations are $(-1, 1)$, $(0, 0)$, $(1, -1)$, and $(-1, 0)$.
- Dataset B contains all the data from dataset A, but has one extra positive observation at $(2, -2)$.

For each of the above datasets, answer the following questions.

1. Are the problems linearly separable in the original input space? You can use scatter plots to show these.
2. If the problem is not linearly separable, provide an example of simple feature transformation ϕ such that the resulting problem is linearly separable with the lowest number of features possible. Otherwise, use the two dimensional feature $\phi(x_1, x_2) = (x_1, x_2)$.
3. Write down the equation for the decision boundaries in the original input space.
4. Write down the coordinates of all support vector(s) for both classes based on your choice for the feature.
5. Does x^* belong to the positive or negative class?

SVM tries to maximize the margin between two classes. Therefore, the optimal decision boundary is diagonal and it crosses the point (3,4). It is perpendicular to the line between support vectors (4,5) and (2,3), hence it is slope is m = -1. Thus the line equation is $(x_2 - 4) = -1(x_1 - 3) = x_1 + x_2 = 7$. From this equation, we can deduce that the weight vector has to be of the form (w_1, w_2), where $w_1 = w_2$. It also has to satisfy the following equations:
$2w_1 + 3w_2 + b = 1$ and
$4w_1 + 5w_2 + b = -1$
Hence $w_1 = w_2 = -1/2$ and $b = 7/2$

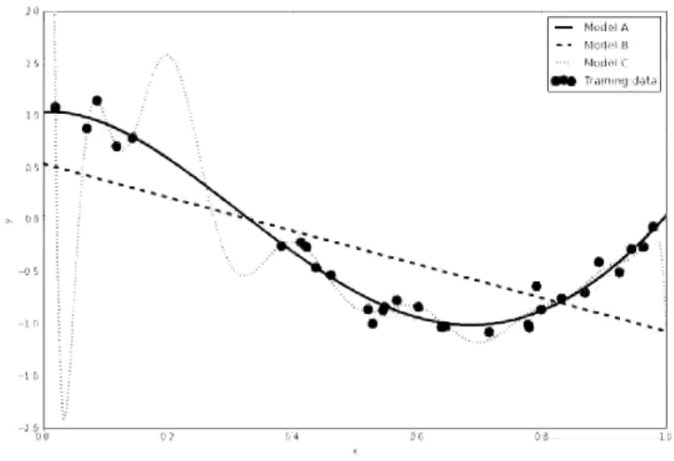
Question 7 [10 marks]

Consider a linear regression problem of estimating a non-linear function f with 30 training data points $\{(x_i, y_i)\}_{i=1}^{30}$. As shown in Fig Q8, three linear regressions were independently performed with polynomial features of polynomial orders 1, 4 and 15.

1. Identify which polynomial degrees (out of 1, 4 and 15) could correspond to models A, B and C. Why?
2. The models A, B, and C independently reported sum-of-squared-error (SSE) values between all training data points and corresponding estimates as 0.35, 6.78 and 0.15, respectively. Explain why model C has a lower SSE, although it seems to have a spurious fit.

$$SSE = \sum_{i=1}^{30} (y_i - \hat{f}(x_i))^2$$

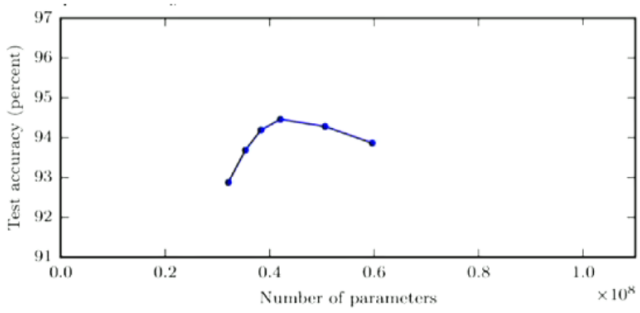
3. Describe a procedure for this particular dataset to determine a suitable model complexity, i.e. the polynomial order.



Question 8

[10 marks]

You are building a deep convolutional neural network for a classification problem on a large dataset of images. This network has three layers but as you add more elements in each layer, the performance start to degrade as indicated in the graph below. Describe two strategies to avoid this problem and explain why.



这道题 1.dropout 2.pooling 两个策略对么

Question 9

[9 marks]

Given that the SVD of a matrix $M = U\Sigma V^T$.

1. Is it correct to say: "The matrix M^TM can be decomposed as $M^TM = V\Sigma V^T$". If it is not, how to make it become correct.
2. Choose a correct matrix to fill in the question mark: $M^TMV = ?\,\Sigma^2$
3. Based on the above, what are eigenvectors and eigenvalues of M^TM ?

Question 10 [4 marks]

Give a vector $a = (1, 3, 4)$ and other vectors $x_1 = (4, 3, 5), x_2 = (0.4, 10, 50)$, and $x_3 = (1, 4, 10)$.

1. Report the minimum distance of a to x_1, x_2, and x_3 and the nearest neighbor of a.
2. Find which of x_1, x_2, x_3 makes the smallest angle with a and report that angle.

Question 11 [12 marks]

Alice and Bob both need to buy a bicycle. The bike store has a stock of 4 green, 3 yellow and 2 red bikes. Alice randomly picks one of the bikes and buy it. Immediately after, Bob does the same. The sale price of the green, yellow and red bikes are \$300, \$200 and \$100, respectively.

Let A be the event that Alice bought a green bike, and B be the event that Bob bought a green bike.

1. What is $\mathbf{P}(A)$? What is $\mathbf{P}(A|B)$? Are A and B independent events? Justify your answer.
2. What is the probability that Alice and Bob bought bicycles of different colors?
3. What is the probability that at least one of them bought a green bike?
4. Given that Bob bought a green bike, what is the expected value of the amount of money spent by Alice?

Question 12 [9 marks]

In Linear Regression given the following cost function

$$J(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left(y^{(i)} - h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) \right)^2 \tag{1}$$

where $h_{\boldsymbol{\theta}}(\mathbf{x}^{(i)}) = \boldsymbol{\theta}^T \mathbf{x}^{(i)}$, feature vector $\mathbf{x}^{(i)} \in R^d$ of the i-th sample, and there are n data samples. We usually use the gradient descent to learn the minimum value of the cost function: $\theta := \theta - \alpha \nabla J(\theta)$.

1. What is the name of the cost function above?
2. Show step-by-step the gradient descent update for this cost function.
3. How will the gradient update change if we add the regularization term?

Question 13 [10 marks]

Support vector machines learn a decision boundary leading to the largest margin from both classes. You are training SVM on a tiny dataset with 4 points shown in Figure 2. This dataset consists of two examples with class label -1 (denoted with triangles), and two examples with class label $+1$ (denoted with plus).

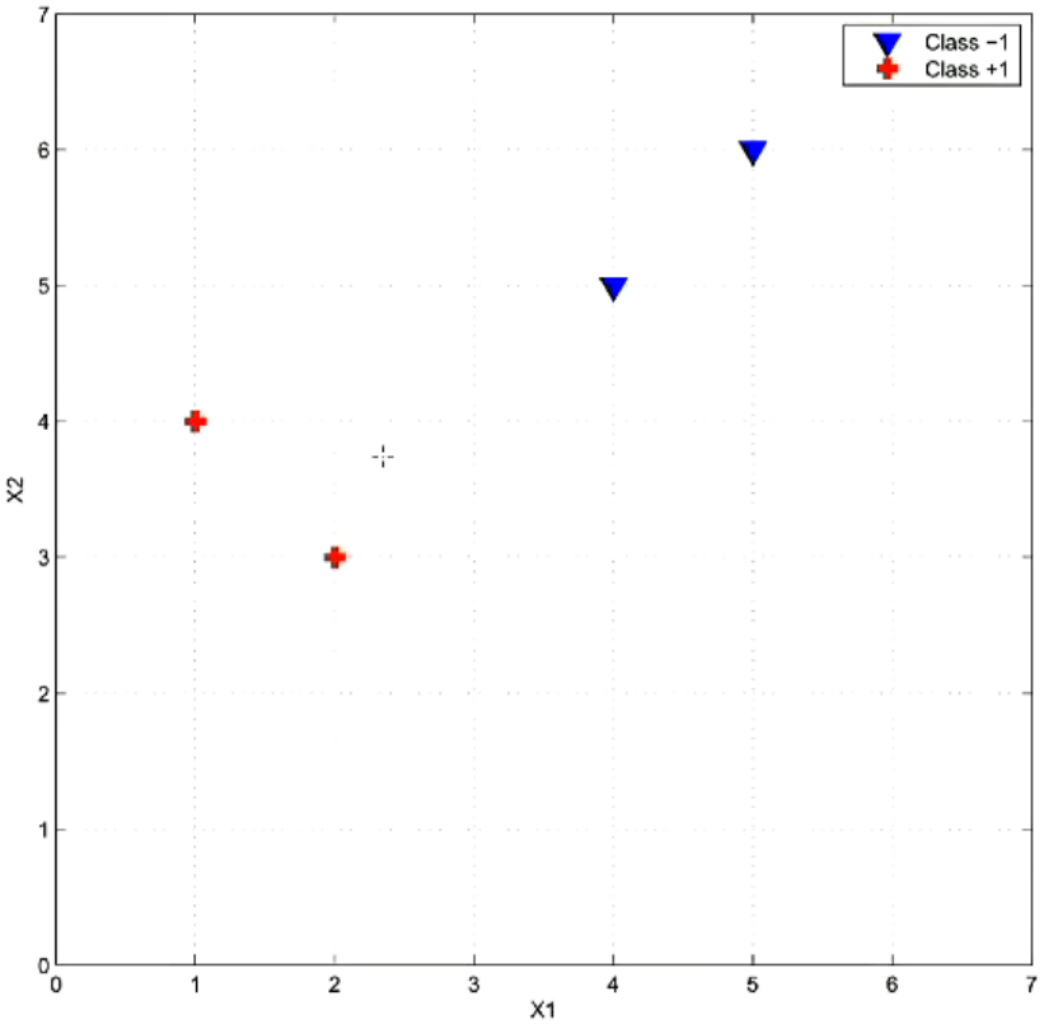


Figure 2: A tiny dataset for SVM

Find the weight vector $\mathbf{w}$ and bias b. What is the equation corresponding to the decision boundary?

SVM tries to maximize the margin between two classes. Therefore, the optimal decision boundary is diagonal and it crosses the point (3,4). It is perpendicular to the line between support vectors (4,5) and (2,3), hence it is slope is m = -1. Thus the line equation is $(x_2 - 4) = -1(x_1 - 3) = x_1 + x_2 = 7$. From this equation, we can deduce that the weight vector has to be of the form (w_1, w_2), where $w_1 = w_2$. It also has to satisfy the following equations:
$2w_1 + 3w_2 + b = 1$ and
$4w_1 + 5w_2 + b = -1$
Hence $w_1 = w_2 = -1/2$ and $b = 7/2$

Question 14

[10 marks]

Consider the problem of binary classification using the Naive Bayes classifier. You are given two dimensional features (X_1, X_2) and the categorical class conditional distributions in the tables below. The entries in the tables correspond to $P(X_1 = x_1|C_i)$ and $P(X_2 = x_2|C_i)$ respectively. The two classes are equally likely.

$X_1 =$ \ Class	C_1	C_2
-1	0.3	0.1
0	0.2	0.6
1	0.5	0.3

$X_2 =$ \ Class	C_1	C_2
-1	0.5	0.3
0	0.2	0.6
1	0.3	0.1

Given a data point $(1, 1)$, calculate the following posterior probabilities: $P(C_1|X_1 = 1, X_2 = 1)$ and $P(C_2|X_1 = 1, X_2 = 1)$

(e) [5 pts] Consider the problem of binary classification using the Naive Bayes classifier. You are given two dimensional features (X_1, X_2) and the categorical class conditional distributions in the tables below. The entries in the tables correspond to $P(X_1 = x_1|C_i)$ and $P(X_2 = x_2|C_i)$ respectively. The two classes are *equally likely*.

X_1 = \ Class	C_1	C_2
-1	0.2	0.3
0	0.4	0.6
1	0.4	0.1

X_2 = \ Class	C_1	C_2
-1	0.4	0.1
0	0.5	0.3
1	0.1	0.6

Given a data point $(-1, 1)$, calculate the following posterior probabilities:

$P(C_1|X_1 = -1, X_2 = 1) =$ Using Bayes' Rule and conditional independence assumption of Naive Bayes

$$\frac{P(X_1=-1,X_2=1|C_1)P(C_1)}{P(X_1=-1,X_2=1)} = \frac{P(X_1=-1|C_1)P(X_2=1|C_1)P(C_1)}{P(X_1=-1|C_1)P(X_2=1|C_1)P(C_1)+P(X_1=-1|C_2)P(X_2=1|C_2)P(C_2)} = 0.1$$

$P(C_2|X_1 = -1, X_2 = 1) = 1 - P(C_1|X_2 = -1, X_1 = 1) = 0.9$

The following questions are multiple choices questions. Please check **ALL CORRECT CHOICES** and circle your answers. Note that every question should have at least one right answer.

Question 15

[5 marks]

Which of the following are true about generative models?

A. They model the joint distribution P(class = C AND sample = x).

B. They can be used for classification.

C. The Perceptron is a generative model.

D. Linear discriminant analysis is a generative model.

(4) [3 pts] Which of the following are true about generative models?

■ They model the joint distribution P(class = C AND sample = x)

□ The perceptron is a generative model

■ They can be used for classification

■ Linear discriminant analysis is a generative model

Question 16

[5 marks]

Suppose we train a hard-margin linear SVM on $n > 100$ data points in R^2, yielding a hyperplane with exactly 2 support vectors. If we add one more data point and retrain the classifier, what is the maximum possible number of support vectors for the new hyperplane (assuming the $n + 1$ points are linearly separable)?

A. 2

B. 3

C. n

D. $n + 1$

Question 17 [5 marks]

Suppose we are given data comprising points of several different classes. Each class has a different probability distribution from which the sample points are drawn. We do not have the class labels. We use k-means clustering to try to guess the classes. Which of the following circumstances would undermine its effectiveness?

A. Each class has the same mean.

B. Choose $k = n$,the number of sample points.

C. Some of the classes aren't normally distributed

D. The variance of each distribution is small in all directions.

(12) [3 pts] Suppose we are given data comprising points of several different classes. Each class has a different probability distribution from which the sample points are drawn. We do not have the class labels. We use k-means clustering to try to guess the classes. Which of the following circumstances would undermine its effectiveness?

☐ Some of the classes are not normally distributed

☐ The variance of each distribution is small in all directions

■ Each class has the same mean

■ You choose $k = n$, the number of sample points