Please tick the box to confirm that your examination paper is complete ☐

Question 1

[9 marks]

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

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

Solution:

1. No. [3]
2. We have: $M = U\Sigma V^T$, $U^T U = V^T V = I$, $M^T = (V^T)^T \Sigma^T U^T = V\Sigma U^T$
   $M^T M = V\Sigma U^T U\Sigma V^T = V\Sigma^2 V^T$ [3]
3. Σ is a diagonal matrix, Σ^2 is also a diagonal matrix whose entry in the ith row and column is the square of the entry in the same position of Σ. So, V is the matrix of eigenvectors of $M^T M$ and Σ^2 is the diagonal matrix whose entries are the corresponding eigenvalues. [3]

Question 2 [9 marks]

In Linear Regression given the following cost function

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

where $h_{\boldsymbol{\theta}}\big(\mathbf{x}^{(i)}\big) = \boldsymbol{\theta}^T \mathbf{x}^{(i)}$, 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: $\theta := \theta - \alpha \nabla J(\theta)$.

- What is the name of the cost function above? [3]
- 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? [3]

Solution:

- Mean square error.
- $\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{n} \sum_{i=1}^{n} \big(h_{\boldsymbol{\theta}}(x^{(i)}) - y^{(i)} \big) x_j^{(i)}$
  $\theta_j = \theta_j - \alpha \frac{\partial J(\theta)}{\partial \theta_j} = \theta_j - \alpha \frac{1}{n} \sum_{i=1}^{n} \big(h_{\boldsymbol{\theta}}(x^{(i)}) - y^{(i)} \big) x_j^{(i)}$
- $\theta_j = \theta_j - \alpha \frac{1}{n} \sum_{i=1}^{n} \big(\big(h_{\boldsymbol{\theta}}(x^{(i)}) - y^{(i)} \big) x_j^{(i)} + \lambda \theta_i \big)$

Question 3 [8 marks]

Assume that we have to classify the dataset which has two dimensional points with label 0 and 1 in the Figure **??**.

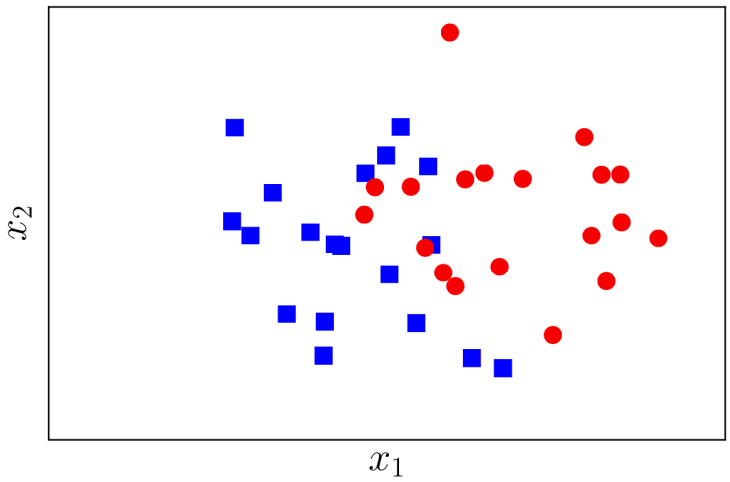


Figure 1: Binary dataset with 2 features

1. Is the problems linearly separable? If it is not, then how to make this problem become linearly separable? [4]
2. Which binary algorithm is optimal for this data set? And explain why you choose it.
   (You can consider Logistic Regression, SVM, or Perceptron . . .) [4]

Solution:

- Not linearly separable but close to linearly separable. Can use kernel svm to transfer data to linear space.
- SVM. Using kernel to transfer dataset to another space which can be linearly separable.
  Or Logistic Regression. Indeed, Logistic Regression does not require 2 classes are linearly separable however we still can find the linear boundary. Therefore LR is suitable for the dataset with 2 classes which are close to linearly separable. (Note that Perception can't work in this case)

Question 4 [12 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).

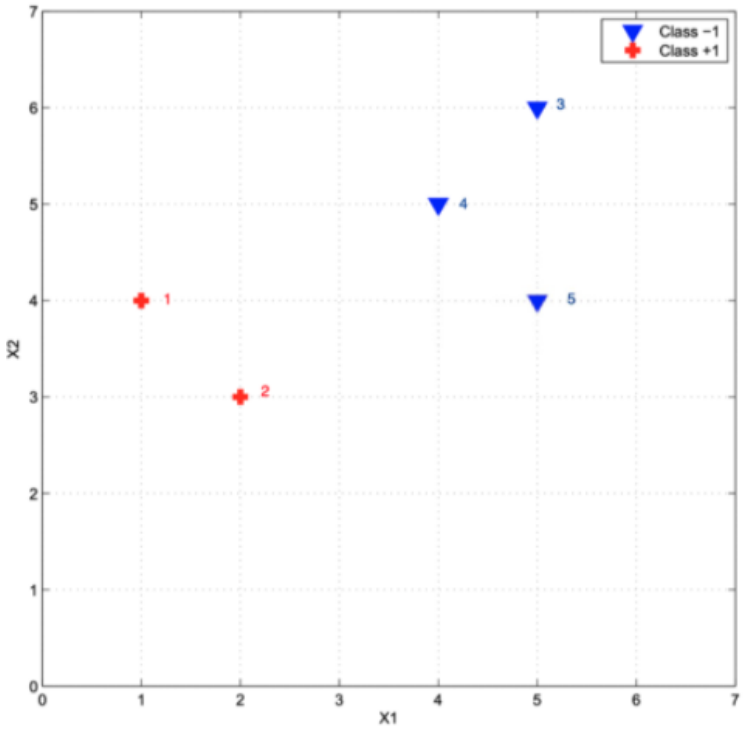


Figure 2: A tiny dataset for SVM

1. Identify the SVM hyperplane that maximizes the margin by providing 2 vector points on this hyperplane. [3]
2. Find the weight vector $\mathbf{w}$ and bias b. What is the equation corresponding to the decision boundary? [3]
3. Which data points will affect the hyperplane if they are removed from the dataset? [3]
4. Which techniques often applied to reduce overfitting in an SVM classifier? [3]

Solution:

1. SVM tries to maximize the margin between two classes. Therefore, the optimal decision boundary is diagonal and it crosses the point (3,4).
2. The line equation is $(x_2 - 4) = -1(x_1 - 3) = x_1 + x_2 = 7$ [2], 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 $b = 7/2$ [1]
3. Remove support vector. 2, 4
4. Using soft SVM reduce the miss-classification penalty, usually known as "C" in SVM.

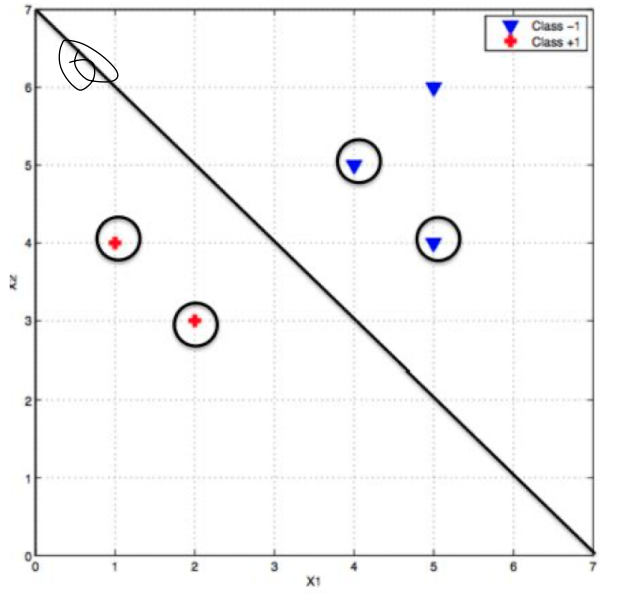


Figure 3: decision boundary

Question 5 [4 marks]

Which of the following are true about generative models?

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

They can be used for classification.

Naive Bayes is a generative model.

Question 6 [4 marks]

Which of below classifiers could have generated this decision boundary in Fig below

=> 1-NN

Question 7 [6 marks]

In ridge linear regression, we can minimize the residual sum of squares:

$$RSS(w) = \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

by using Gradient Descent or using direct solution:

$$\mathbf{w}^{LMS} = (\mathbf{X}^T\mathbf{X} + \lambda\mathbf{I})^{-1}\mathbf{X}^T\mathbf{y} \quad (2)$$

1. Explain in which case using Gradient Descent is more practical than using direct solution (2)? Why? [3]
2. Explain the meaning of λ in the solution (2)? [3]

Solution:

1. The solution needs to calculate the inverse of the matrix. It requires high complexity, and it is impractical for very large D or N.
2. λ is regularization term which forces w to be small and prevent over-fitting

Question 8 [4 marks]

In neural networks, what are the purposes of nonlinear activation functions such as sigmoid, tanh, and ReLU?

Solution: help to learn nonlinear decision boundaries

Question 9 [4 marks]

Alice uses multi-layer neural networks and notices that the training error is going down and converges to a local minimum. Then when she tests on the new data, the test error is abnormally high. What is probably going wrong and what do you recommend her to do?

Solution:

1. The training data size is not large enough. Collect a larger training data and retrain it
2. Play with learning rate and add regularization term to the objective function
3. Use a different initialization and train the net-work several times. Use the average of predictions from all nets to predict test data

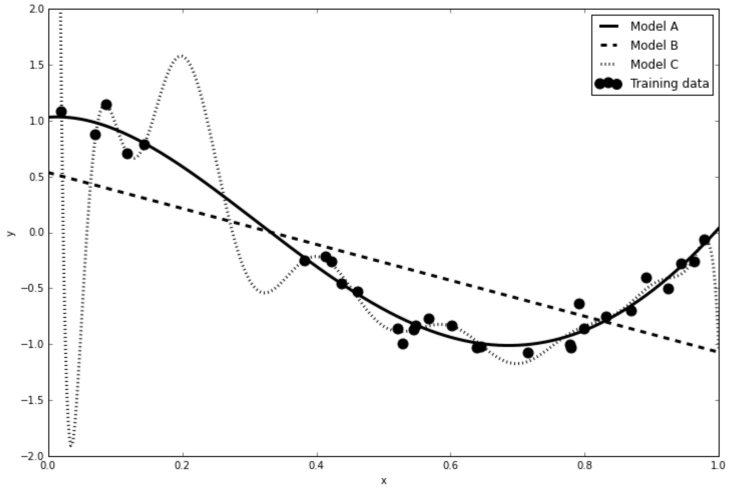
Question 10 [6 marks]

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]
2. The models A, B, and C independently reported RSS 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. [2]

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

3. Describe a procedure for this particular dataset to determine a suitable model complexity, i.e. the polynomial order. [2]



Solution:

1. B-1(linear), A-4, C-15 (highest degrees as it fits almost data points)
2. C has a lower SSE => using high degrees can fit almost data points => low loss.
3. A seems suitable as it can capture the tendency of the dataset, C seems to be overfitting while B is underfitting

Question 11 [4 marks]

What are the differences between Kmeans and GMM (Gaussian Mixture model)? Solution: mainly about determintistic vs probabilistic clustering, k-means is hard assignment, GMM is soft assignment. Both are clustering and have similar iterative algorithm style.

Question 12 [4 marks]

In convolutional neural networks (CNN), what are the main purposes of

1. Convolutional layer, Relu layer, and Maxpooling layer?
2. Dropout and Batch normalisation

Convolutional layer: is used as set of filters (become our parameters which will be learned by the network) to search the smilar pattern or information in the given dataset. Relu layer: prevent the exponential growth in the computation required to operate the neural network. Maxpooling layer: reduce the dimension. dropout: prevent the overfitting. reduce the complexity of the model. batch normalisation: standardize the inputs to a network, accelerates the training

Question 13 [4 marks]

Which methods you can use to do multi-class classification by using only binary classification, and explain how these methods work?.

Using one and rest

Question 14 [10 marks]

There are 2 version of the figure for K Means on Canvas

Version 1

Using the k-means algorithm for clustering the data represented in the Figure 5. There are 3 clusters with initial central points: 1, 3, and 4. Illustrate the partition changes of dataset after two updates. [5]

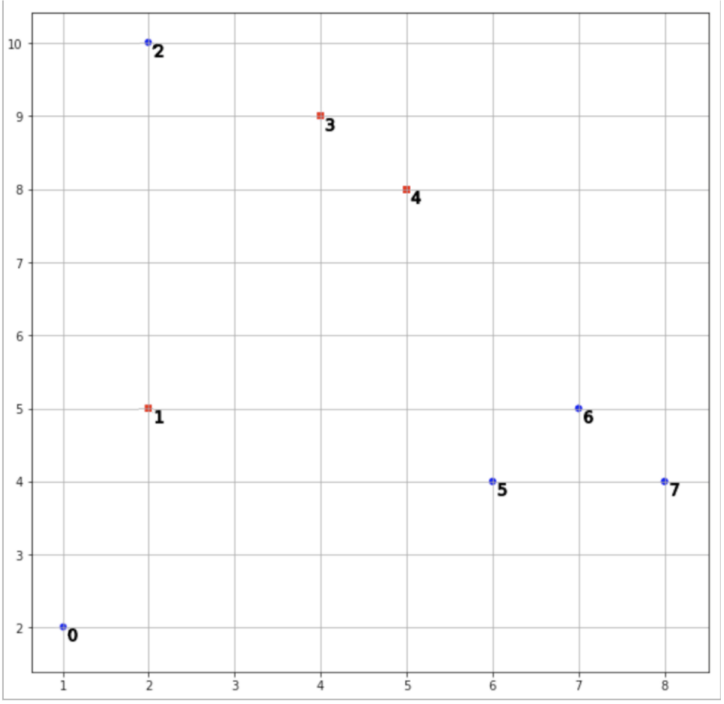


Figure 4: Initial data points

How is k-means clustering algorithm different from the k-nearest neighbor?

Solution:

1. At the initial stage: Central point 1,3,4
   First Update:
   Cluster 1: 0,1
   Cluster 2: 2,3
   Cluster 3: 4,5,6,7
   Second Update:
   Cluster 1: 0,1
   Cluster 2: 2,3,4
   Cluster 3: 5,6,7
2. K-mean and KNN : 3 over 5 main points will got full mark

K-means	KNN
It is an Unsupervised learning technique	It is a Supervised learning technique
It is used for Clustering	It is used mostly for Classification, and sometimes even for Regression
K in K-Means is the number of clusters the algorithm is trying to identify/learn from the data. The clusters are often unknown since this is used with Unsupervised learning.	K in KNN is the number of nearest neighbours used to classify or (predict in case of continuous variable/regression) a test sample
It is typically used for scenarios like understanding the population demomgraphics, market segmentation, social media trends, anomaly detection, etc. where the clusters are unknown to begin with.	It is used for classification and regression of known data where usually the target attribute/variable is known before hand.
In training phase of K-Means, K observations are arbitrarily selected (known as centroids). Each point in the vector space is assigned to a cluster represented by nearest (euclidean distance) centroid. Once the clusters are formed, for each cluster the centroid is updated to the mean of all cluster members. And the cluster formation restarts with new centroids. This repeats until the centroids themselves become mean of clusters, i.e., when updating centroids to mean doesnot change them. The prediction of a test observation is done based on nearest centroid.	K-NN doesnot have a training phase as such. But the prediction of a test observation is done based on the K-Nearest (often euclidean distance) Neighbours (observations) based on weighted averages/votes.

Version 2

Using the k-means algorithm for clustering the data represented in the Figure 5. There are 3 clusters with initial central points: 1, 3, and 4. Illustrate the partition changes of dataset after two updates.

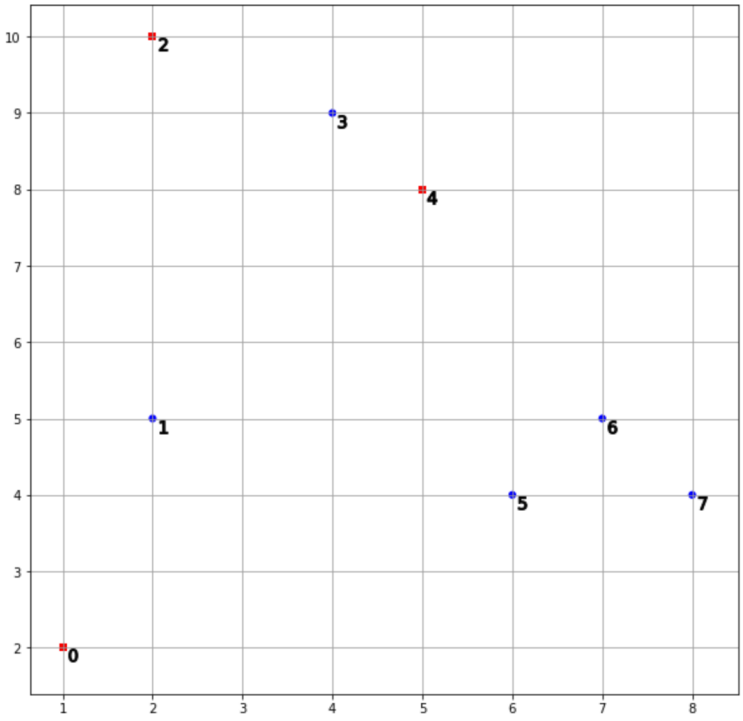


Figure 5: Initial data points

How is k-means clustering algorithm different from the k-nearest neighbor?
Solution:

1. At the initial stage: Central point 0,2,4
   First Update:
   Cluster 1: 0,1
   Cluster 2: 2
   Cluster 3: 3, 4,5,6,7
   Second Update:
   Cluster 1: 0,1
   Cluster 2: 2,3
   Cluster 3: 4, 5,6,7
2. K-mean and KNN : 3 over 5 main points will got full mark

Question 15

[9 marks]

The table below includes the dataset recording all attributes of both mammals and non-mammals Version 1

The table below includes the dataset recording all attributes of both mammals and non-mammals

Name	Give Birth	Can Fly	Live in Water	Have Legs	Class
human	yes	no	no	yes	mammals
python	no	no	no	no	non-mammals
salmon	no	no	yes	no	non-mammals
whale	yes	no	yes	no	mammals
frog	no	no	sometimes	yes	non-mammals
komodo	no	no	no	yes	non-mammals
bat	yes	yes	no	yes	mammals
pigeon	no	yes	no	yes	non-mammals
cat	yes	no	no	yes	mammals
leopard shark	yes	no	yes	no	non-mammals
turtle	no	no	sometimes	yes	non-mammals
penguin	no	no	sometimes	yes	non-mammals
porcupine	yes	no	no	yes	mammals
eel	no	no	yes	no	non-mammals
salamander	no	no	sometimes	yes	non-mammals
gila monster	no	no	no	yes	non-mammals
platypus	no	no	no	yes	mammals
owl	no	yes	no	yes	non-mammals
dolphin	yes	no	yes	no	mammals
eagle	no	yes	no	yes	non-mammals

Given a new data sample has following attributes:

Give Birth	Can Fly	Live in water	Have Legs	Class
no	no	yes	yes	?

1. Calculate the following probabilities: $P(A|M)[2], P(A|N)[2], P(A|M)P(M)[2], P(A|N)P(N)[2]$ where A, M, and N stand for Attribute, Mammal, and Non-mammal, respectively.
2. Which class the given new data sample is classified to based on Naive Bayes? [1]

Solution: $P(A|M) = \frac{1}{7} * \frac{6}{7} * \frac{2}{7} * \frac{5}{7} = 0.025$
$P(A|N) = \frac{12}{13} * \frac{10}{13} * \frac{3}{13} * \frac{9}{13} = 0.113$
$P(A|M)P(M) = 0.025 * \frac{7}{20} = 0.0085$
$P(A|N)P(N) = 0.113 * \frac{13}{20} = 0.074$
$P(A|M)P(M) <= P(A|N)P(N)$ non mammals

Version 2

Give Birth	Can Fly	Live in water	Have Legs	Class
no	no	yes	no	?

Solution: $P(A|M) = \frac{1}{7} * \frac{6}{7} * \frac{2}{7} * \frac{2}{7} = 0.01$

$P(A|N) = \frac{12}{13} * \frac{10}{13} * \frac{3}{13} * \frac{4}{13} = 0.05$
$P(A|M)P(M) = 0.01 * \frac{7}{20} = 0.0035$
$P(A|N)P(N) = 0.05 * \frac{13}{20} = 0.0325$
$P(A|M)P(M) <= P(A|N)P(N)$ non mammals

Version 3

Given a new data sample has following attributes:

Give Birth	Can Fly	Live in water	Have Legs	Class
no	no	no	yes	?

1. Calculate the following probabilities: $P(A|M), P(A|N), P(A|M)P(M)$, and $P(A|N)P(N)$.

2. What is the class of a new data sample?

Your Answer:

1. P(A|M)=1/7*6/7*5/7*5/7=0.0625

P(A|N)=12/13*10/13*6/13*9/13=0.2269

P(A|M)P(M)=1/7*6/7*5/7*5/7*7/20 = 0.0219

P(A|N)P(N)=12/13*10/13*6/13*9/13*13/20 =0.1475

2. as P(A|N)P(N)>P(A|M)P(M), the class of a new data sample:

is non-mammals.