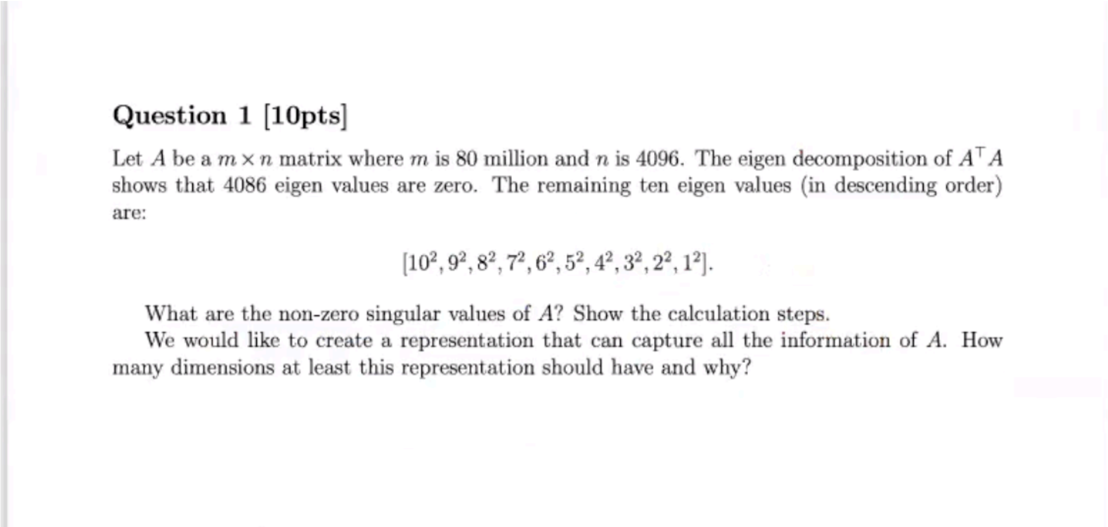


Question 1 [10pts]

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

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

What are the non-zero singular values of A? Show the calculation steps.

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?

1. We can learn that the rank of A is 4096-4086 = 10

Because of

$$A = U \sum V^T$$

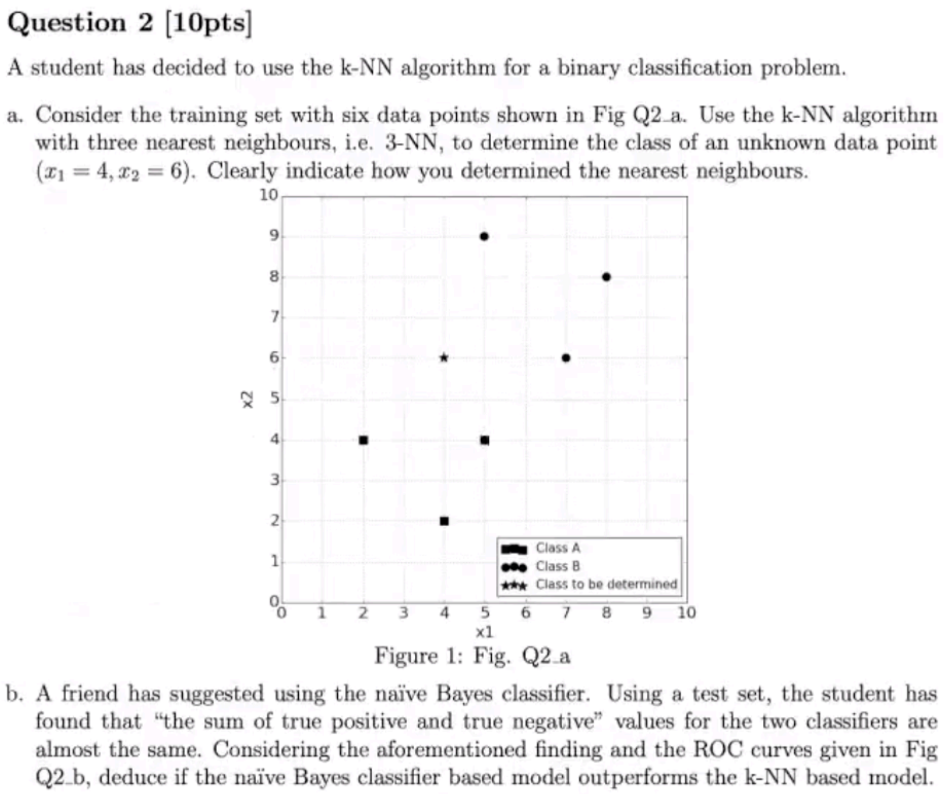
We can get

$$A^T A = (U\Sigma V^T)^T (U\Sigma V^T) = V\Sigma U^T * U\Sigma V^T = V\Sigma^2 V^T$$

So , we now know that the non-zero singular values of A have a square relationship with Σ^2

Non-zero singular values of A is [10, 9, 8, 7, 6, 5, 4, 3, 2, 1]

2. 4096 x 10



Question 2 [10pts]

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

a. Consider the training set with six data points shown in Fig Q2_a. Use the k-NN algorithm with three nearest neighbours, i.e. 3-NN, to determine the class of an unknown data point ($x_1 = 4, x_2 = 6$). Clearly indicate how you determined the nearest neighbours.

10
9
8
7
6
5
4
3
2
1
0
x2
0 1 2 3 4 5 6 7 8 9 10
x1
■ Class A
● Class B
★ Class to be determined

Figure 1: Fig. Q2_a

b. A friend has suggested using the naïve Bayes classifier. Using a test set, the student has found that "the sum of true positive and true negative" values for the two classifiers are almost the same. Considering the aforementioned finding and the ROC curves given in Fig Q2_b, deduce if the naïve Bayes classifier based model outperforms the k-NN based model.

1. a. I will calculate the Euclidean distance
2. between the test data(x1=4, x2=6) and the sample data. For example, the first point (2,4)from class A, the distance of (2, 4) and (4, 6) is sqrt((x1-x2)^2 + (y1-y2)^2) = sqrt(8), and the point from

算法流程

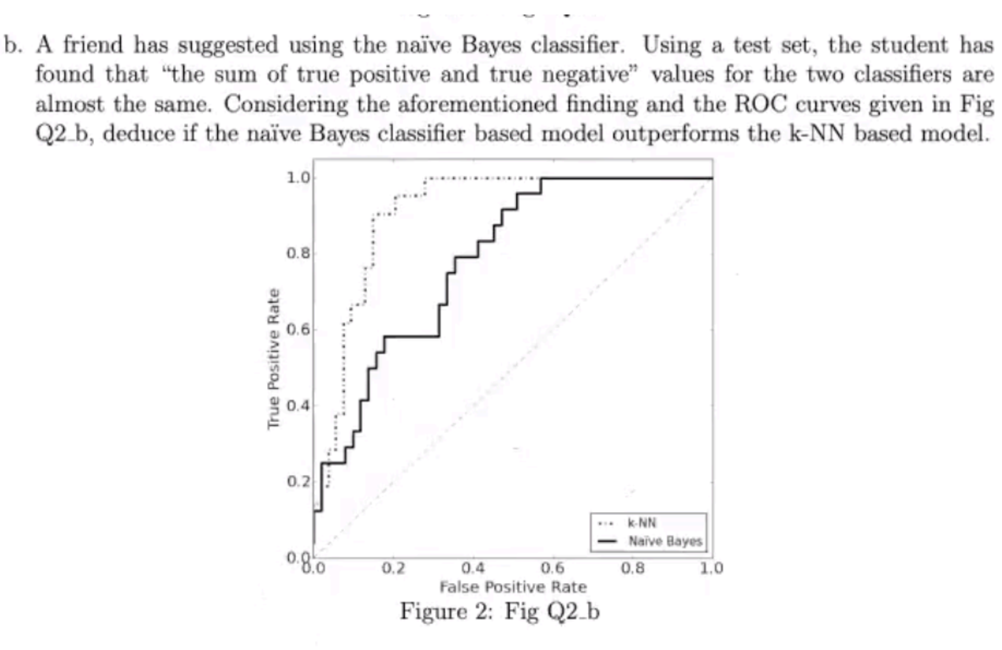
总体来说，KNN 分类算法包括以下 4 个步骤：[4]

①准备数据，对数据进行预处理[4] 。

②计算测试样本点（也就是待分类点）到其他每个样本点的距离[4] 。

③对每个距离进行排序，然后选择出距离最小的 K 个点[4] 。

④对 K 个点所属的类别进行比较，根据少数服从多数的原则，将测试样本点归入在 K 个点中占比最高的那一类[4] 。

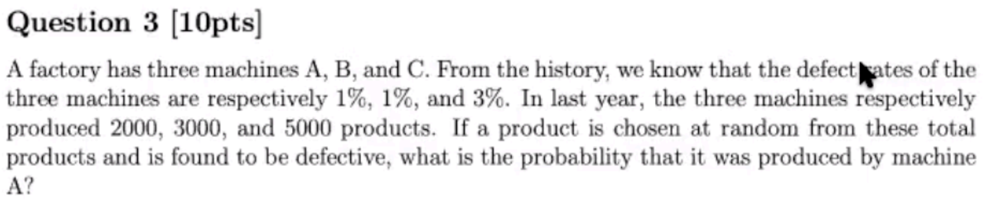


b. A friend has suggested using the naïve Bayes classifier. Using a test set, the student has found that "the sum of true positive and true negative" values for the two classifiers are almost the same. Considering the aforementioned finding and the ROC curves given in Fig Q2_b, deduce if the naïve Bayes classifier based model outperforms the k-NN based model.

Figure 2: Fig Q2_b

从上面给出的四个点可以发现，ROC 曲线图中，越靠近(0,1)的点对应的模型分类性能越好，所以，可以确定的是 **ROC 曲线图中的点对应的模型，它们的不同之处仅仅是在分类时选用的阈值(Threshold)不同，每个点所选用的阈值都对应某个样本被预测为正类的概率值**

From the above figure, it can be found that in the ROC curve graph, the closer the point (0, 1), the better the classification performed of the model. Therefore, the models corresponding to the points in the ROC curve are different in the threshold value replaced during classification. The threshold value selected by each point corresponds to the probability value of a certain sample being predicted as a positive class. So the KNN is better than the Naïve Bayes.



Question 3 [10pts]

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

Let D = defective items

Total production = 2000 + 3000 + 5000 = 10000

P(A) = 2000/10000 = 0.2

P(B) = 3000/10000 = 0.3

P(C) = 5000/10000 = 0.5

P(D|A) = 0.01

P(D|B) = 0.01

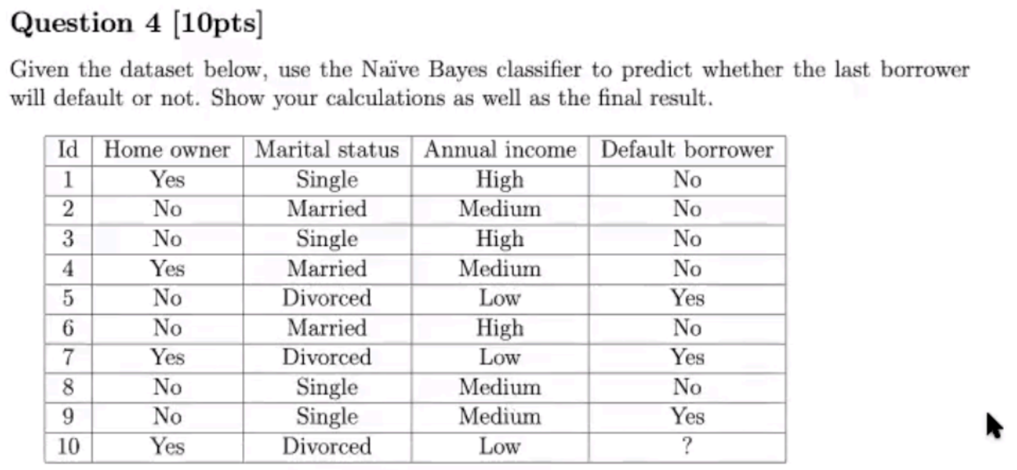
P(D|C) = 0.03

Now

P(A|D) = P(D|A) P(A) / P(D)

P(D) = (0.01 x 2000 + 0.01 x 3000 + 0.03 x 5000) / 10000 = 0.02

P(A|D) = (0.01 x 0.2) / 0.02 = 0.1



Question 4 [10pts]

Given the dataset below, use the Naïve 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	?

Let the final result is attribute, the probability of attribute is P(A)

P(A|Y) = 1/3 x 2/3 x 2/3 = 4/27

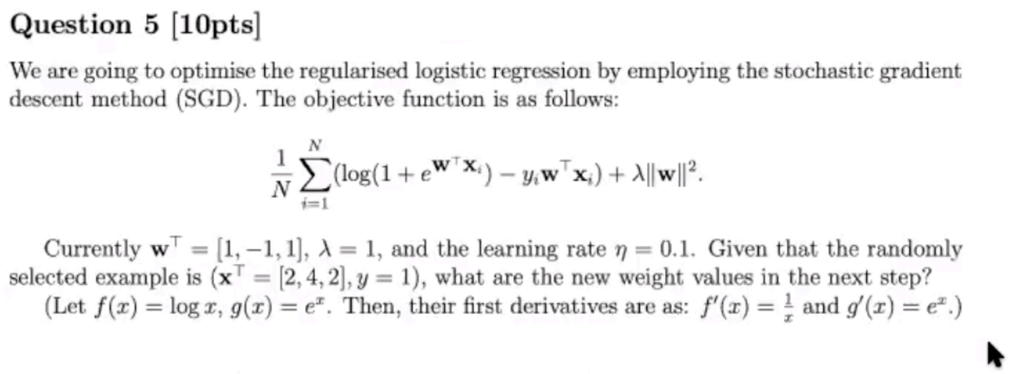
P(A|N) = 2/6 x 1/6 x 1/6 = 1/108 We should assume that there is one ‘Marital status = Divorced’ , because when the ‘Default borrower = NO’ , the marital status has no dicorved.

P(A|Y)P(Y) = 4/27 x 1/3 = 4/81

P(A|N))(N) = 0 x 6/9 x 1/108= 0

P(A|Y)P(Y) > P(A|N))(N)

So the last borrower will default.



Question 5 [10pts]

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

$$\frac{1}{N}\sum_{i=1}^{N}(\log(1+e^{\mathbf{w}^\top\mathbf{x}_i})-y_i\mathbf{w}^\top\mathbf{x}_i)+\lambda\|\mathbf{w}\|^2.$$

Currently $\mathbf{w}^\top=[1,-1,1]$, $\lambda=1$, and the learning rate $\eta=0.1$. Given that the randomly selected example is $(\mathbf{x}^\top=[2,4,2],y=1)$, what are the new weight values in the next step?

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

Let the function is f(x)

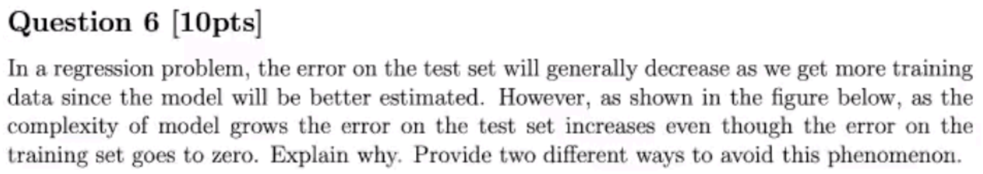
So

$$\frac{\partial loss}{\partial w}=\frac{1}{1+\ e^{w^Tx_i}}e^{w^Tx_i}*x-yx+2\lambda||W||$$

= 0.5 x [2, 4, 2] – [2, 4, 2] + 2[1, -1, 1] = [1, -4, 1]

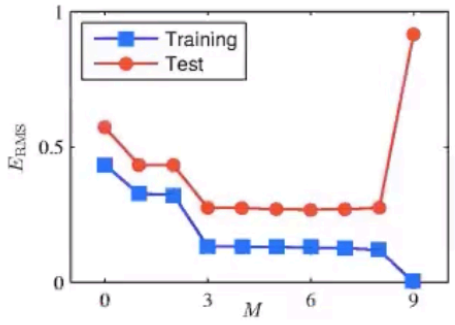
W_1 = W_0 – $\eta\ f'(x)$ = [1, -1, 1]- 0.1[1, -4, 1] = [0.9, -0.6, 0.9]

$$\frac{\partial loss}{\partial w}=\frac{1}{1+\ e^{w^Tx_i}}\ e^{w^Tx_i}*x-yx+2\lambda||W||$$



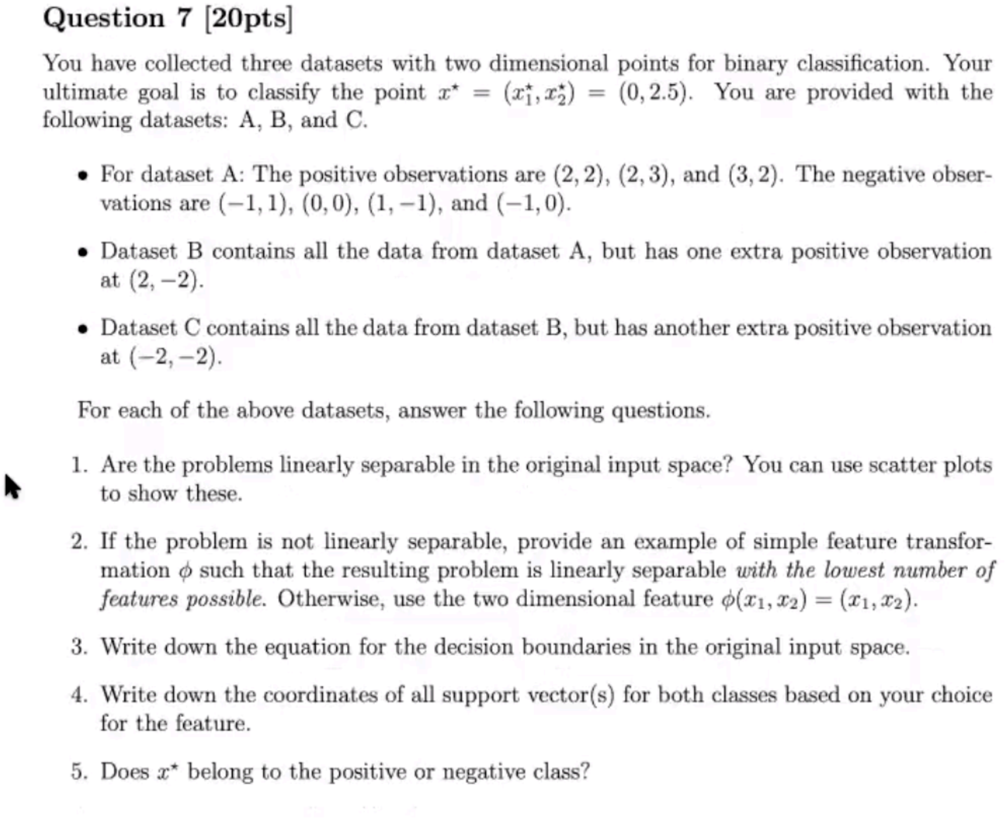
Question 6 [10pts]

In a regression problem, the error on the test set will generally decrease as we get more training data since the model will be better estimated. However, as shown in the figure below, as the complexity of model grows the error on the test set increases even though the error on the training set goes to zero. Explain why. Provide two different ways to avoid this phenomenon.



After the M=8, it is overfitting. The error on the training dataset is small while the error on the test dataset is large.

1. Reducing the complexity of the model to **prevent the model from overfitting the training set.**
2. Increase the size of training data.



Question 7 [20pts]

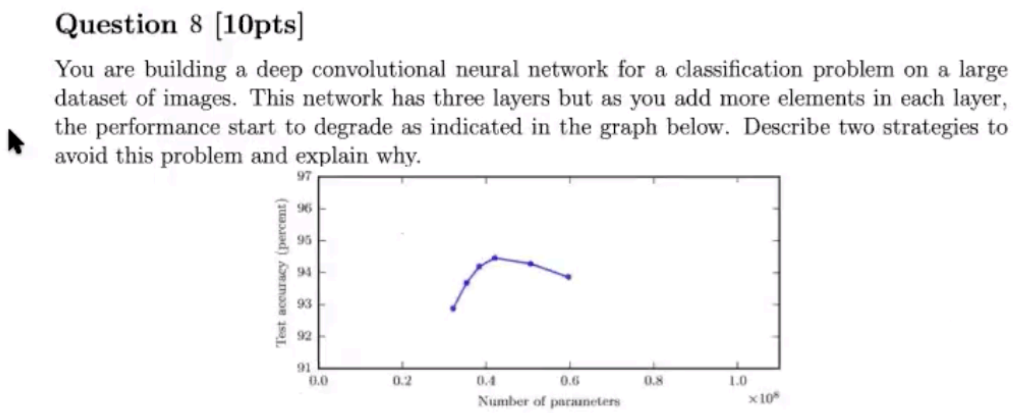
You have collected three datasets with two dimensional points for binary classification. Your ultimate goal is to classify the point $x^\star = (x_1^\star, x_2^\star) = (0, 2.5)$. You are provided with the following datasets: A, B, and C.

- 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 $(2, -2)$.
- Dataset C contains all the data from dataset B, but has another 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 $x^\star$ belong to the positive or negative class?

1. Hk
2. 2-dimensional cannot divided dataset C, so we use three-dimensional, so we use the 3D and make the kernel as small as possible. $\emptyset(x, y, x^2 + y^2)$
3. A: y = -x + 2
   B: x-1.5 = 0
   C: $x^2 + y^2 = 5$
4. A: (0, 0), (1, -1), (-1, 1), (1, 1)
   B: (2, 3), (2, 2), (2, -2), (1, -1)
   C: (-1, 1), (2, 2), (1, -1), (2, -2), (-2, -2)
5. A: 2 < 2.5, so the x* belongs to the positive class.
   B: 2 > 1.5, so the x* belongs to the negative class.
   C: 2.5 x 2.5 = 6.26 > 5, so the x* belongs to the positive class.



Question 8 [10pts]

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.

1. **Dropout. Randomly delete some hidden neurons to prevent the model from being overly dependent on some neurons.**
2. **Decrease the number of layers. Reduce the complexity of the model and prevent the model from overfitting the training set**

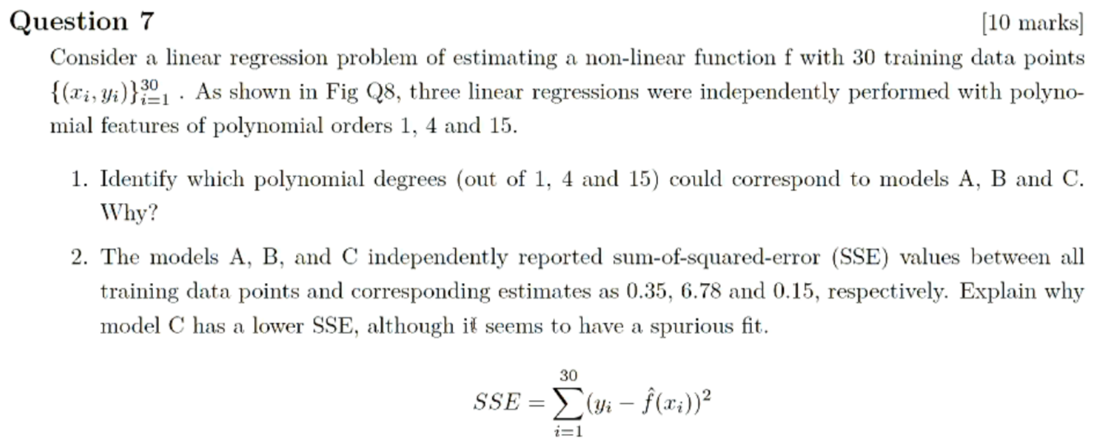
Question6

Consider the random variables X, Y, Z which have the following joint distribution:

p(X, Y, Z) = p(X)p(Y |X)p(Z|Y).

Show that X and Z are conditionally independent given Y

$$P(X,Z|Y) = \frac{P(X,Y,Z)}{P(Y)} = \frac{P(X)P(Y|X)P(Z|Y)}{P(Y)} = P(Z|Y)\frac{P(X)P(Y|X)}{P(Y)} = P(Z|Y)P(X|Y)$$



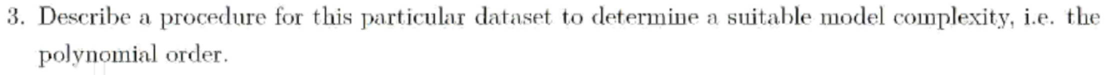
Question 7

[10 marks]

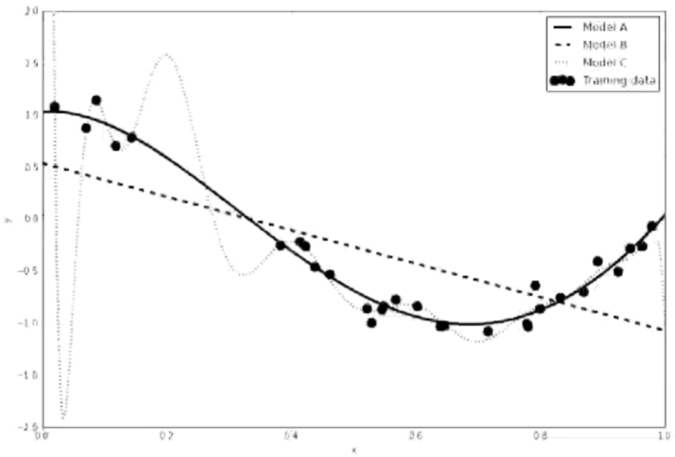
Consider a linear regression problem of estimating a non-linear function f with 30 training data points $\{(x_i, y_i)\}_{i=1}^{30}$. As shown in Fig Q8, three linear regressions were independently performed with polynomial features of polynomial orders 1, 4 and 15.

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.

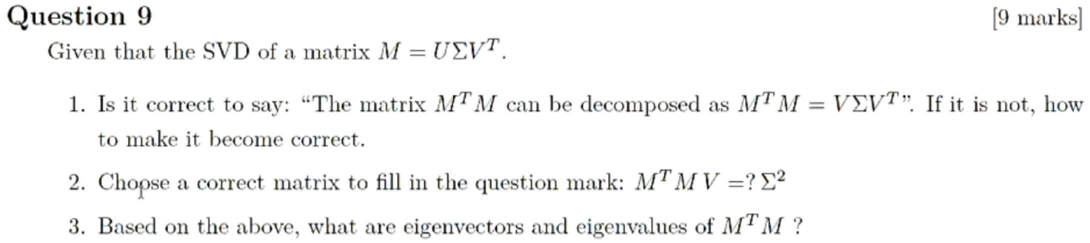
$$SSE = \sum_{i=1}^{30}(y_i - \hat{f}(x_i))^2$$



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



1. The ‘1’ is model B, the ‘4’ is model A, the ‘15’ is model C. It depends on the highest power (n) of the polynomial. The higher the highest power, the more times the line will bend, up to (n-1) times.
2. The model C has overfitting. About the whole training data points, the line of model C passes through almost all points, so its predicted value is close to the true value. Therefore, the model C has the smaller error.
3. Draw a simulated line to see the trend of the model. It depends on the highest power (n) of the polynomial. The higher the highest power, the more times the line will bend, up to (n-1) times.



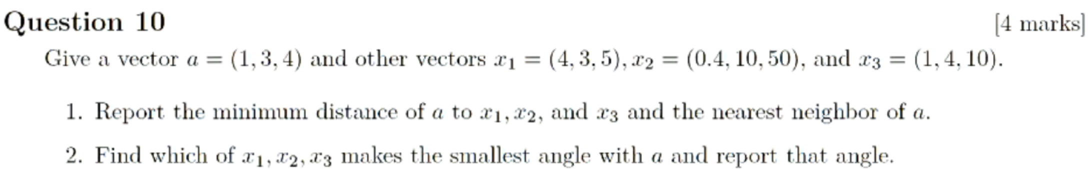
Question 9

[9 marks]

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

1. Is it correct to say: “The matrix $M^T M$ can be decomposed as $M^T M = V\Sigma V^T$”. If it is not, how to make it become correct.
2. Chopse 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 $M^T M$?

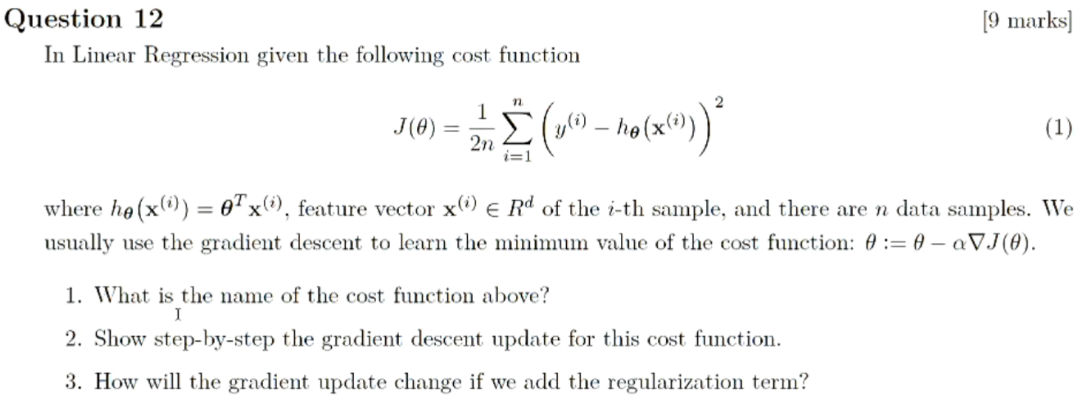
$$M^T M = (U\Sigma V^T)^T (U\Sigma V^T) = V\Sigma U^T * U\Sigma V^T = V\Sigma^2 V^T$$



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 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: $\theta := \theta - \alpha\nabla J(\theta)$.

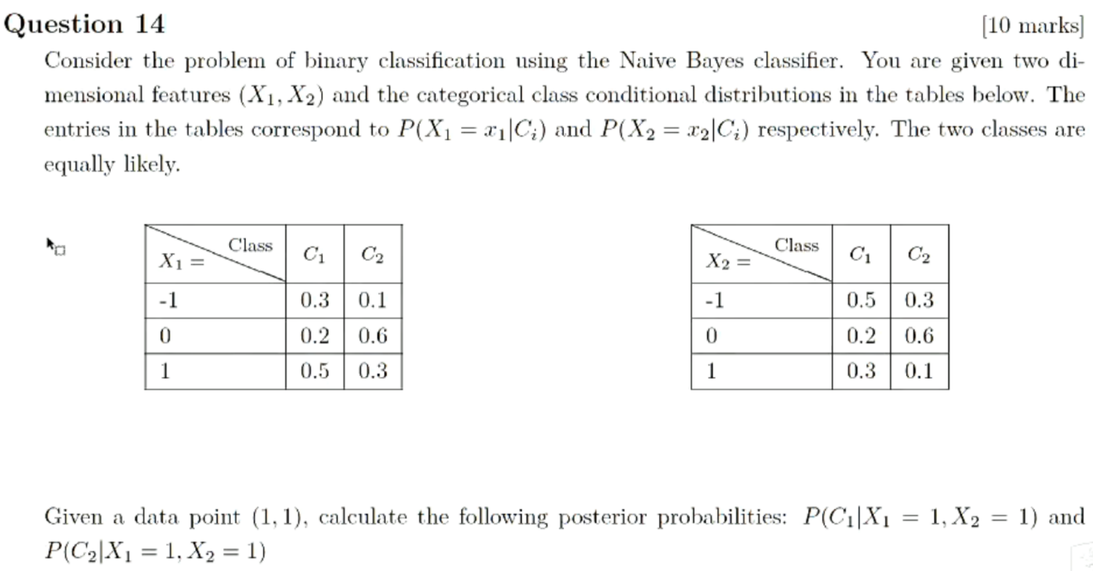
1. What is the name of the cost function above?
2. Show step-by-step the gradient descent update for this cost function.
3. How will the gradient update change if we add the regularization term?

1. Mean square error
2. $\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{n}\sum_{i=1}^{n}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) * x_j^{(i)}$

$$\theta_j = \theta_j - \alpha\frac{\partial J(\theta)}{\partial \theta_j} = \theta_j - \alpha\frac{1}{n}\sum_{i=1}^{n}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) * x_j^{(i)}$$

3. If will change to

$$\theta_j = \theta_j - \alpha\frac{\partial J(\theta)}{\partial \theta_j} = \theta_j - \alpha[\frac{1}{n}\sum_{i=1}^{n}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right) * x_j^{(i)} + \lambda\theta_j]$$



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

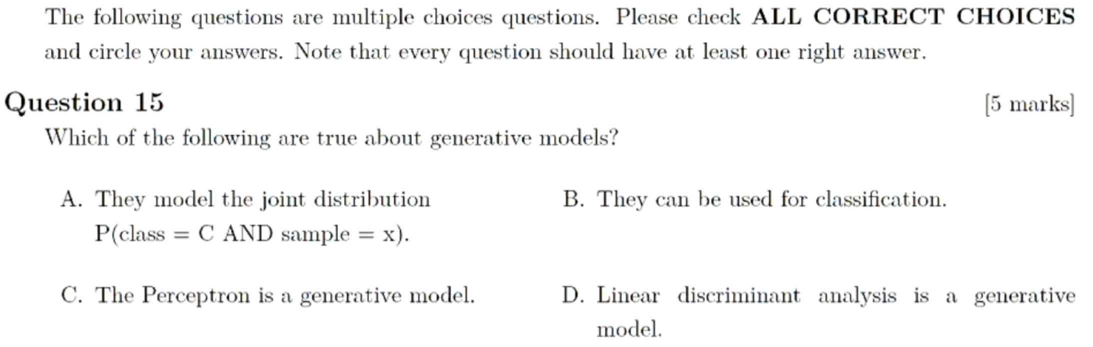
$X_2 =$ \ Class	C_1	C_2
-1	0.5	0.3
0	0.2	0.6
1	0.3	0.1

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

P(C1|X1 = 1, X2 = 1) =[P(X1=1, X2=1|C1)P(C1)] / [P(X1 = 1, X2 = 1)]

=[P(X1=1|C1)P(X2=1|C1)P(C1)] / [P(X1=1|C1)P(X2=1|C1)P(C) + P(X1=1|C2)P(X2=1|C2)P(C2)]

=5/6



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?

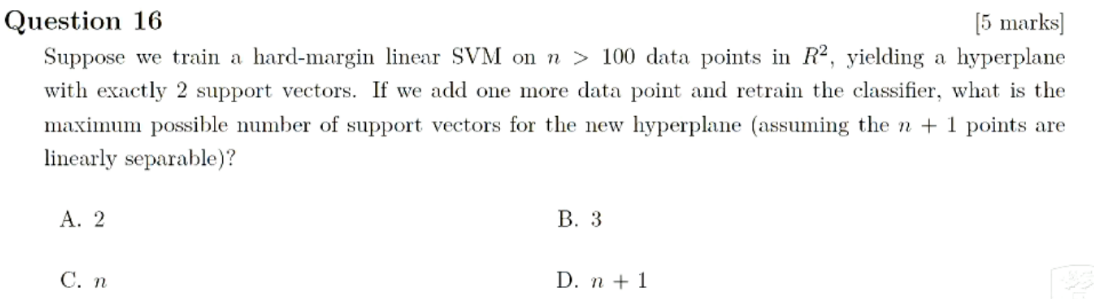
A. They model the joint distribution P(class = C AND sample = x).

B. They can be used for classification.

C. The Perceptron is a generative model.

D. Linear discriminant analysis is a generative model.

ABD



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)?

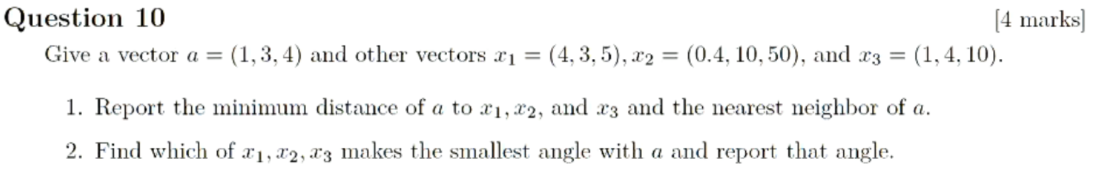
A. 2

B. 3

C. n

D. $n + 1$

D

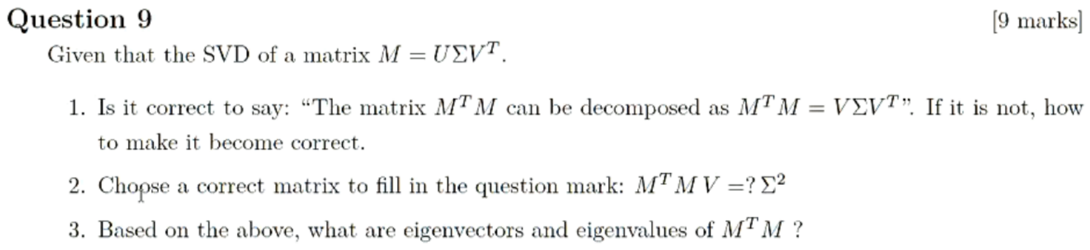


Question 10

[4 marks]

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

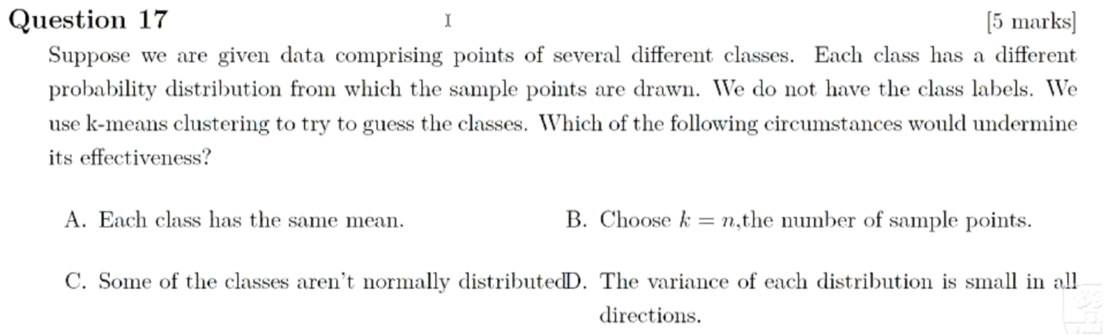
1. Report the minimum distance of a to x_1, x_2, and x_3 and the nearest neighbor of a.
2. Find which of x_1, x_2, x_3 makes the smallest angle with a and report that angle.



Question 9 [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 $M^T M$?



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.

ABC