Name and Student ID:

Machine Learning BLG527E, Nov 1, 2018, 120mins, Midterm Exam - ANSWERS
Signature:
Duration: 120 minutes.
Closed books and notes. Write your answers neatly in the space provided for them. Write your name on each sheet. Good Luck!

Question	Q1	Q2	Q3	Q4	Q5	Q6	TOTAL
MaxPoint	15	15	15	15	20	20	100
ExpectedPoints							
ActualPoints							

QUESTIONS

In the table below, x_1, x_2, x_3 and $x_i \in \{0,1\}$, i = 1,2,3 x_i represent the *i* feature vector and y $\in$ {+,-} represents the class label.

Id	x_1	x_2	x_3	y	$dist(x^t,x^6)$
1	1	0	0	+	**1**
2	0	1	0	+	3
3	0	0	1	-	**1**
4	0	0	0	-	2
5	1	1	1	-	**1**
6	1	0	1	?	

Q1) [15pts]
Classify the data point with Id=6 using a 1-NN classifier on the given training dataset (datapoints with Id=1..5). Use Euclidean distance as the distance measure.

We give the EuclideanDistance(x^t,x^6) on the dist column,
where EuclideanDist(x^t,x^6)=sum_{j=1..3}($x_j^t,x_j^6)^2$
x^6 is closest (distance=1) to x1,x3,x5 with labels +,-,-
therefore, we can label it as -.
Labeling it as + will be accepted as the correct answer well.

Q2) [15pts]
Cluster the dataset using 2-means clustering and the same distance measure as in 1a). Take datapoints with Id=5 and Id=4 as the initial cluster centers.
Which cluster does the datapoint with Id=6 belong to?
How would you classify the datapoint with Id=6?
Let m1: cluster centered at x4, m2: cluster centered at x5. Then the distances to m1 and m2 are as follows:

Id	Dist(x4)	Dist(x5)	y
1	1	2	+
2	1	2	+
3	1	2	-
4	0	3	-
5	3	0	-
6	1	0	

Data points with id 1..4 belong to m1 and data point with id 5 belong to m2.
Cluster centers become:

m1	0.25	0.25	0.25
m2	1	1	1

Data points with id 1..4 still belong to m1 and data point with id 5 belong to m2.

The k-means clustering algorithm converges.

The datapoint with id6 has
dist(m1,x6)=$2*0.75^2+0.25^2$ =1.1875
dist(m2,x6)=1
So x6 belongs to cluster2, which has only 1 datapoint with a negative label. Therefore, it could be labeled as -.

Question did not specify if x6 is to be included at k-means clustering or not. Accepted either way.

Name and Student ID:

Q3) [15pts]

Suppose you are given a financial regression dataset generated from a polynomial of degree of 2. Indicate whether you think the bias and variance of the following models would be relatively high (H) or low (L) considering the true model.

	Bias	Variance
Linear Regression	H	L
Polynomial regression with degree of 4	L	H
Polynomial regression with degree of 9	L	H

Q4) [15pts]

For the following classification dataset, you need to evaluate the relevance of each feature and select the most relevant feature (i.e. the most correlated with the label). Evaluate the relevance of a feature by means of its dot product with the label, for two binary vectors u, v, sim(u,v) = $u^T.v$.

Feature1 Weekend (F_1)	Feature2 Rain (F_2)	Feature3 Daytime (F_3)	Label (r): Take Taxi
Yes	No	Morning	+
Yes	Yes	Morning	+
Yes	Yes	Morning	+
No	Yes	Morning	+
No	Yes	Noon	+
No	No	Noon	-
Yes	No	Noon	-
Yes	Yes	Noon	-
No	No	Noon	-
No	No	Noon	-

Let -1,+1encoded Features and labels be:

Feature1 Weekend (F_1)	Feature2 Rain (F_2)	Feature3 Daytime (F_3)	Label (r): Take Taxi
1	1	1	1
1	1	1	1
1	1	1	1
-1	1	1	1
-1	1	-1	1
-1	-1	-1	-1
1	-1	-1	-1
1	1	-1	-1
-1	-1	-1	-1
-1	-1	-1	-1
Sim(Fi,label)			Sim(label,label)
2	8	8	10

We represent features and labels as -1,1 vectors and compute the dot-product. Feature2&F3 are closest to the label. Hence we can choose either. Changing 0-1 encoding could give a different result. It is also OK to compute the impurity when data are separated according to each single variable.

Q5)[20pts]

The probability of a single observation x is a normal with mean μ as follows:

$$\boldsymbol{P}(\boldsymbol{x}|\mu) = N(\mu, 1) \quad for\ x = 0,1,2 \dots n$$

You are given the data points X = {x_1 x_2,… x_n} that are drawn independently from $\boldsymbol{P}(\boldsymbol{x}|\mu)$.

Write down the log-likelihood of the data:

$$logL(X|\mu) = logP(X|\mu) = \log \prod_{t=1..n} p(x^t|\mu) = \sum_{t=1..n} \log(p(x^t|\mu))$$

$$= \sum_{t=1..n} \log(\frac{1}{\sqrt{2\pi 1}} \exp(-\frac{(x^t-\mu)^2}{2*1})) = C - \frac{1}{2}\sum_{t=1..n}(x^t-\mu)^2$$

where C is a constant.

Find the maximum likelihood estimate of the parameter μ:

Taking derivative of the logL wrto μ and setting it to zero:

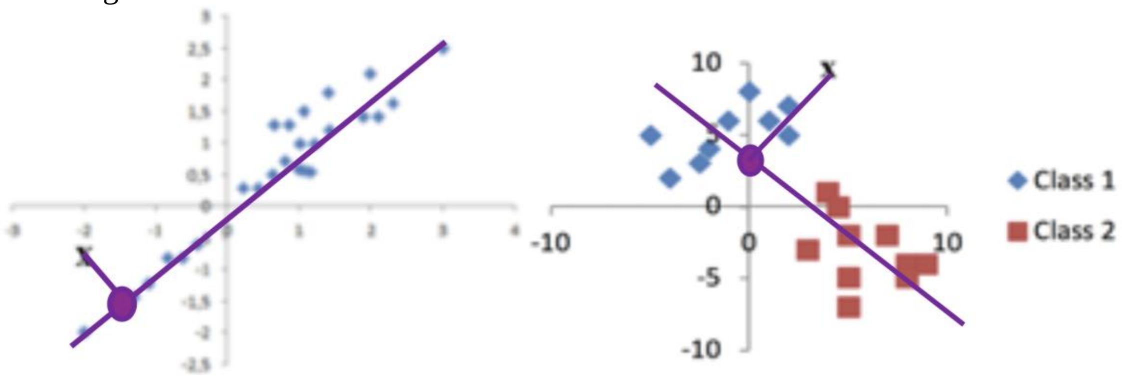
$$\sum_{t=1..n}(x^t-\mu) = 0 \quad \rightarrow \quad \mu = \frac{1}{n}\sum_{t=1..n} x^t$$

Name and Student ID:

Q6)[20pts]
Describe briefly the following dimensionality reduction methods:
Principal Component Analysis (PCA):
Is an unsupervised and linear dimensionality reduction method. Projects d dimensional inputs into k<d dimensional space by taking the dot product of input with the eigen vectors beloging to the highest k eigen values of the covariance of the inputs. PCA aims to project inputs to dimensions with the highest variance so that reconstruction error is minimized.

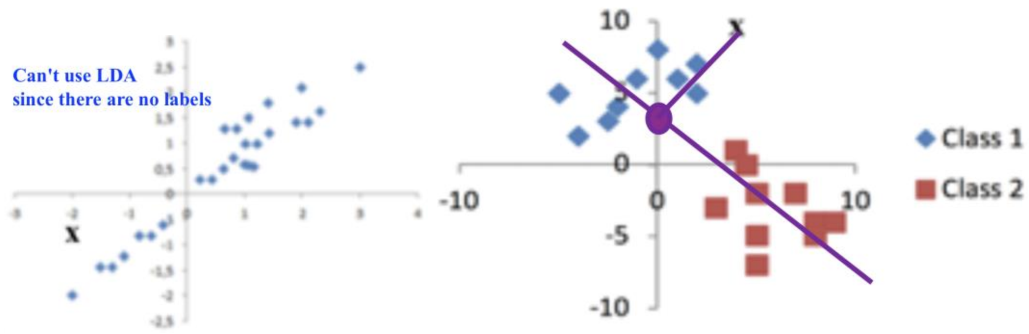
Linear Discriminant Analysis (LDA):
Is a supervised and linear dimensionality reduction method. Projects d dimensional inputs into k<d dimensional space by taking the dot product of input with the eigen vectors beloging to the highest k eigen values of the matrix $S_w^{-1}S_B$ where S_w and S_b are within and between class covariance matrices respectively. LDA aims to minimize the class variances and maximize the difference between class means at the projected space.

Plot where x would be projected if the datasets on the left and right were both projected according to PCA:



Plot where x would be projected if the datasets on the left and right were both projected according to LDA:

LDA projection for case on the right would not differ much from the PCA projection.



Name and Student ID:

If you need more space, use this page.