Name and Student ID:

Machine Learning BLG527E, Nov 6, 2014, 120mins, Midterm Exam (PART A)
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!

Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9	TOTAL
10	**6**	**14**	**15**	**10**	10	10	10	15	100

QUESTIONS

Q1) [10pts]
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**
1	1	0	0	-
2	0	1	0	-
3	0	0	1	-
4	0	0	0	+
5	1	1	1	+

1a) Construct the Naïve Bayes classifier for the given training dataset.

$p(y = -) = 3/5$, $p(y = +) = 2/5$
$p(x_1 = 1|y = -) = 1/3$, $p(x_1 = 0|y = -) = 2/3$, $p(x_1 = 1|y = +) = ½$ $p(x_1 = 0|y = +) = ½$
$p(x_2 = 1|y = -) = 1/3$, $p(x_2 = 0|y = -) = 2/3$, $p(x_2 = 1|y = +) = ½$ $p(x_2 = 0|y = +) = ½$
$p(x_3 = 1|y = -) = 1/3$, $p(x_3 = 0|y = -) = 2/3$, $p(x_3 = 1|y = +) = ½$ $p(x_3 = 0|y = +) = ½$

1b) Classify the $(x_1 = 1, x_2 = 1, x_3 = 0)$ data sample.

$p(y = -)\ p(x_1 = 1|y = -)\ p(x_2 = 1|y = -)\ p(x_3 = 0|y = -) = 3/5 \times 1/3 \times 1/3 \times 2/3 = 2/45$
$p(y = +)\ p(x_1 = 1|y = +)\ p(x_2 = 1|y = +)\ p(x_3 = 0|y = +) = 2/5 \times 1/2 \times 1/2 \times 2/2 = 1/20$

$p(y = +| x_1\ x_2\ x_3) > p(y = -| x_1\ x_2\ x_3)$ data sample belongs to +

Q2) [6pts]
Suppose you are given a financial regression dataset generated from a polynomial of degree of 4. 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	L
Polynomial regression with degree of 9	L	H

Name and Student ID:

Q3) [14pts]

3a) Generate a decision tree using Gini index (2p(1-p)) as impurity measure.

Weekend (x_1)	Rain (x_2)	Daytime (x_3)	Take Taxi C
Yes	No	Morning	+
Yes	Yes	Morning	+
Yes	Yes	Morning	+
No	Yes	Evening	+
No	Yes	Evening	+
No	No	Noon	-
Yes	No	Noon	-
Yes	Yes	Noon	-
No	No	Evening	-
No	No	Evening	-

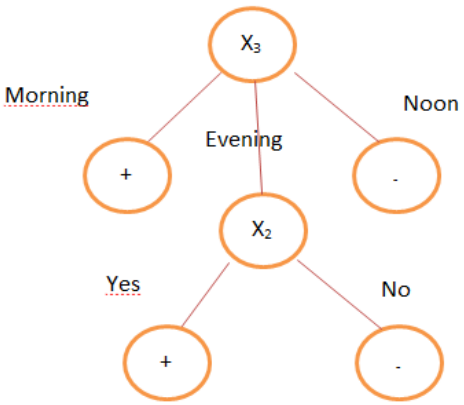
x_1 : Y: 3+, 2-, N: 2+, 3-
x_2 : Y: 4+, 1-, N: 1+, 4-
x_1 : M: 3+, 0-, E: 2+, 2-, N: 0+, 3-

$$Imp(x_1) = \frac{5}{10}.2.\frac{3}{5}.\frac{2}{5} + \frac{5}{10}.2.\frac{3}{5}.\frac{2}{5} = 0.48$$

$$Imp(x_2) = \frac{5}{10}.2.\frac{4}{5}.\frac{1}{5} + \frac{5}{10}.2.\frac{4}{5}.\frac{1}{5} = 0.32$$

$$Imp(x_3) = \frac{3}{10}.2.\frac{3}{3}.\frac{0}{3} + \frac{4}{10}.2.\frac{2}{4}.\frac{2}{4} + \frac{3}{10}.2.\frac{0}{3}.\frac{3}{3} = 0.2$$

$Imp(x_3)$ is the smallest impurity value. Therefore x_3 would be the root.



3b) Are there any irrelevant features?

x_1 is irrelevant

Name and Student ID:

Q4)[15pts]
The probability of a single observation x with rate parameter θ follows the following Poisson distribution:

$$\boldsymbol{P}(\boldsymbol{x}|\theta) = \frac{\theta^x e^{-\theta}}{x!}, for\ x = 0,1,2 \dots n$$

You are given the data points x_1 x_2,… x_n that are drawn independently from Poisson distribution with parameter θ.

Write down the log-likelihood of the data:

$$\boldsymbol{L}(\theta|X) = \prod_{i=1}^{n} \frac{\theta^{x_i} e^{-\theta}}{x_i!}, \log\big(\boldsymbol{L}(\theta|X)\big) = \sum_{i=1}^{n} (x_i log\theta - \theta - \log(x_i!)),$$

Find the maximum likelihood estimate of the parameter θ:

$$\frac{\partial \log\big(\boldsymbol{L}(\theta|X)\big)}{\partial \mathrm{x}} = \frac{1}{\theta}\sum_{i=1}^{n} x_i - n = 0 \xrightarrow{yields} \theta = \frac{1}{n}\sum_{i=1}^{n} x_i$$

Q5)[10pts]
Describe briefly the following dimensionality reduction methods:
Principal Component Analysis (PCA):

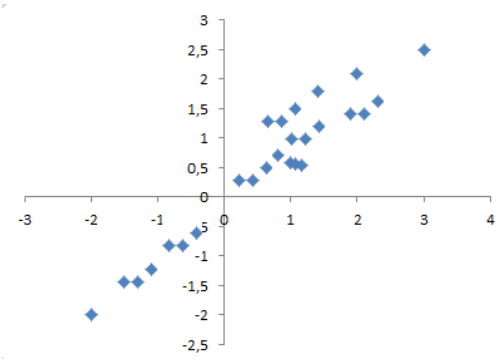
See lecture notes

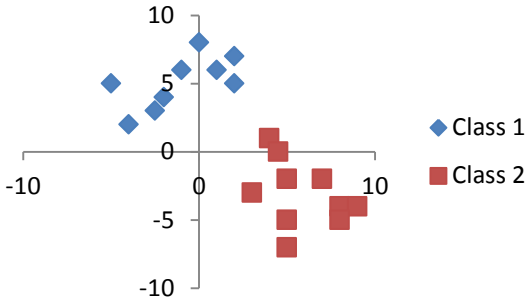
Linear Discriminant Analysis (LDA):

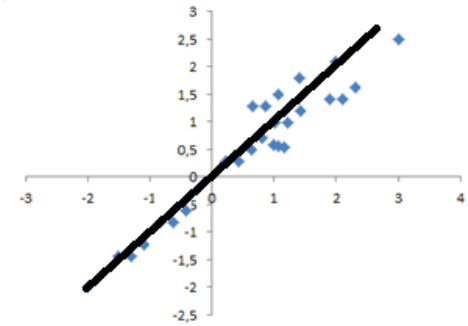
See lecture notes

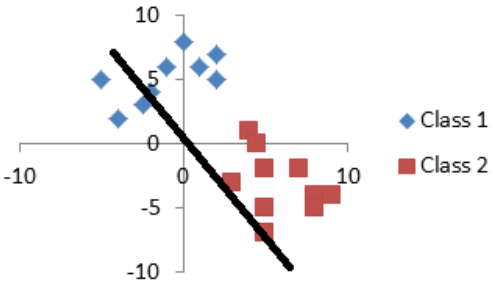
Plot the directions of the first PCA and LDA components of the following figures.

Name and Student ID:









Name and Student ID:

If you need more space, use this page only for PART A (Q1-Q5).

Name and Student ID:

Machine Learning BLG527E, Nov 6, 2014, 120mins, Midterm Exam (PART B)
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!

Q1	*Q2*	*Q3*	*Q4*	*Q5*	**Q6**	**Q7**	**Q8**	**Q9**	TOTAL
10	*6*	*14*	*15*	*10*	**10**	**10**	**10**	**15**	100

Q6)[10pts]

Briefly explain the K-Means algorithm:

K-means is a clustering algorithm. It clusters the given unlabeled dataset X={x^t},t=1,…N, into K clusters with means m_i, i=1,..K. K-means algorithm assumes circular clusters and discrete cluster membership. If the closest cluster center for the the t'th data point is the ith one, then x^t belongs to cluster i, shown as $b_i^t = 1$, otherwise, $b_i^t = 0$. The algorithm aims to minimize the reconstruction error

J=1/N $\frac{1}{N}\sum_{i=1}^{K}\sum_{t=1}^{N}\|m_i - x^t\|^2 b_i^t$. K is a parameter that needs to be specified. The algorithm may converge to a local minimum, so it should be run a number of times and the best solution should be used.

Algorithm Kmeans
Input: X_{Nxd},K,
Output: cluster centers $m_i, i = 1,..K$, cluster memberships, $b_i^t, i = 1,..K, t = 1,..N$
Initialize cluster centers m_i randomly (e.g. initialize to random points from X)
repeat
Determine the closest cluster center for each data point x^t:

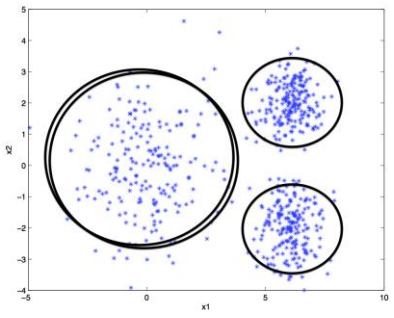
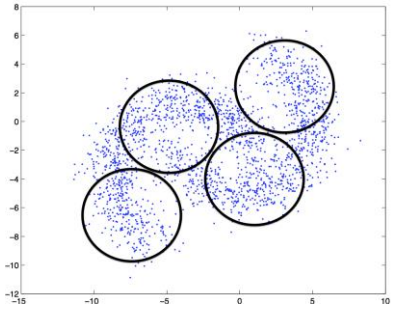
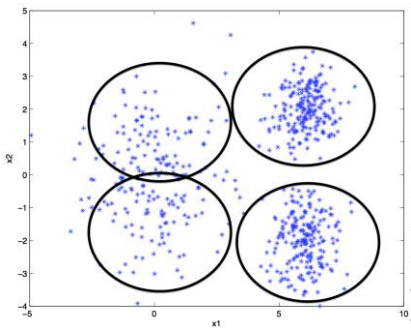
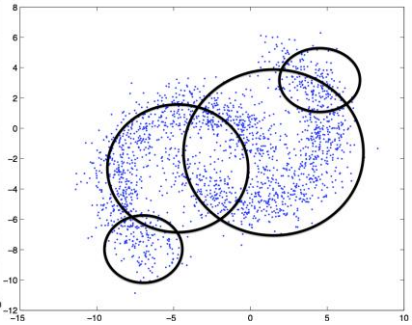
Set $b_j^t = 0$, for j=1,..K, let i=argmin$_{j=1..K}\|m_j - x^t\|^2$, set $b_i^t = 1$

Recompute the cluster centers:

$$m_i = \frac{\sum_{t=1}^{N} x^t b_i^t}{\sum_{t=1}^{N} b_i^t}$$

until cluster centers (or cluster memberships) do not change.

What are two possible clusterings of the following datasets using K-Means with K=4. Only draw the final clusters. (There is no unique answer for this)

Considering these datasets, does the K-Means algorithm have any drawbacks? Can you suggest alternative clustering methods?

For the left hand side, a fixed value of K=4; for the right hand side, circular clusters are problems of K-means. Hierarchical clustering could fix both problems, GMM with EM could be helpful for the LHS.

Name and Student ID:

Q7)[10pts]
What is the VC dimension of a line in 2 dimensional space? Explain your answer.

VC dimension h of a hypothesis class H is the maximum number of specifically chosen data points whose all 2^h possible labelings can be correctly classified by a certain classifier in H.
The VC dimension of a line in 2 dimensional space is 3.
Because for 3 datapoints, all 8 possible labelings can be correctly classified by a line.
For 4 datapoints, the following labeling can not be correctly labeled by any line, therefore the maximum number of points that can be shattered by a line in 2d is 3.

Q8)[10pts]
Assume that g is a quadratic model and for input x, z=[x_1^2 x_2^2 …x_d^2]

which outputs the following:

$$g(\mathrm{x}, \mathrm{v}, \mathrm{w}, w_0) = \mathrm{v}^{\mathrm{T}}\mathrm{z} + \mathrm{w}^{\mathrm{T}}\mathrm{x} + w_0$$

You need to make the parameters v of g take smaller values and hope that it will result in better generalization. Given a dataset $X = \{\mathrm{x}^t, r^t\}_{t=1}^N$, how would you obtain the solution for v,w and w_0 in this situation.

Hint: Modify the sum of squares error function to incorporate the need of smaller v values, and derive the solution for v,w and w_0 analytically.

The sum of squares error function is: $E(v, w, w_0|X) = \frac{1}{N}\sum_{t=1}^{N}\left[r^t - g(x^t, v, w, w_0)\right]^2 = \frac{1}{N}\sum_{t=1}^{N}\left[r^t - \left(v^T z^t + w^T x^t + w_0\right)\right]^2$

We need to modify it to incorporate the constraint that $v^T v$ should be as small as possible:

$$E_\lambda(\mathrm{v}, \mathrm{w}, w_0|X) = \mathrm{E}(\mathrm{v}, \mathrm{w}, w_0|X) + \lambda \mathrm{v}^T \mathrm{v}$$

In order to compute the solution for v, w, and w_0 we need to take the derivative of E_λ with respect to each of these variables and equate it to 0 to find the solution:

$$\frac{d\, E_\lambda(\mathrm{v}, \mathrm{w}, w_0|X)}{d\, w_0} = \frac{d\, \mathrm{E}(\mathrm{v}, \mathrm{w}, w_0|X)}{d\, w_0} = \frac{2}{N}\sum_{t=1}^{N} \mathrm{v}^{\mathrm{T}} z^t + \mathrm{w}^{\mathrm{T}} x^t + w_0 - r^t$$

$$w_0 = \frac{1}{N}\sum_{t=1}^{N} r^t - \mathrm{v}^{\mathrm{T}} z^t - \mathrm{w}^{\mathrm{T}} x^t$$

Similarly, we need to solve for w and v, plugging in the solution we already computed for w_0.

$$\frac{d\, E_\lambda(\mathrm{v}, \mathrm{w}, w_0|X)}{d\, \mathrm{w}} = \frac{d\, \mathrm{E}(\mathrm{v}, \mathrm{w}, w_0|X)}{d\, \mathrm{w}} = \frac{2}{N}\sum_{t=1}^{N} (\mathrm{v}^{\mathrm{T}} z^t + \mathrm{w}^{\mathrm{T}} x^t + w_0 - r^t)\, x^t = 0$$

$$\frac{d\, E_\lambda(\mathrm{v}, \mathrm{w}, w_0|X)}{d\, \mathrm{v}} = \frac{2}{N}\sum_{t=1}^{N} (\mathrm{v}^{\mathrm{T}} z^t + \mathrm{w}^{\mathrm{T}} x^t + w_0 - r^t)\, z^t + 2\lambda \mathrm{v} = 0$$

Name and Student ID:

Q9)[15pts]
Suppose you became the head of the department in the future and students have complained that amount of cheese (kaşar) in the sandwiches of the canteen has significantly reduced. You know that according to the regulations the cheese should be 50grams. After the complaints you go to the canteen and buy 10 sandwiches and measure the amount of cheese in them as follows:

51	51	51	51	51	47	47	47	47	47

Can you decide with 95% confidence that amount of cheese is significantly reduced? Justify your decision.

Hint:
If variance is known, with m average of x_i , we accept that x is less than p_0 with 100(1-α) confidence if

$$\frac{\sqrt{N}(m - m_0)}{s} \sim z_a \text{ is less than } z_\alpha$$

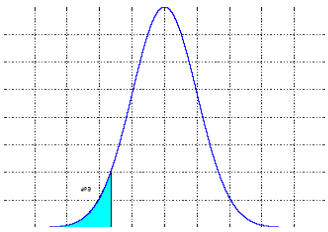
If variance is not known, with m and s^2 average and var of x_i , we accept that x is less than p_0 with 100(1-α) confidence if

$$\frac{\sqrt{N}(m - p_0)}{S} \sim t_{N-1} \text{ is less than } t_{\alpha,N-1}$$

$t_{0.05,9} = 1.83$, $t_{0.025,9} = 2.26$, $z_{0.05} = 1.64$, $z_{0.025} = 1.96$, choose wisely.

In order to make your computations easier, when computing the standard deviation, divide by N (not by N-1). Take $\sqrt{10}$ =3.16.

For the given sample, we can compute the mean m = 49, sample standard deviation s = 2
Since the std is not known and it is computed based on the data, we need to use the t-test.
The mean we assume the data came from is: p_0=50
We should accept, H_0: $m = p_0$ if $\frac{\sqrt{N}(m - p_0)}{S} >= -t_{\alpha,N-1}$
3.16*(49-50)/2 = -1.58 >=-1.85
Therefore, we have to accept that the cheese is NOT significantly reduced.



Nam nd Student ID:

If you need more space, use this page only for PART B (Q6-Q9).