ANSWER SHEET

Name and Student ID:

Machine Learning BLG527E, March 22, 2017, 120mins, Midterm Exam
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	TOTAL
20	20	20	20	20	100

QUESTIONS

Q1) [20pts]
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.

Prior Prob.

$P(C_+) = \frac{2}{5}$

$P(C_-) = \frac{3}{5}$

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	-

j=	1	2	3
$P(x_j=0 \mid C_+)$	1/2	1/2	1
$P(x_j=1 \mid C_+)$	1/2	1/2	0
$P(x_j=0 \mid C_-)$	2/3	2/3	1/3
$P(x_j=1 \mid C_-)$	1/3	1/3	2/3

1a) [15pts] Construct the Naïve Bayes classifier for the given training dataset.

$$P(C_+ \mid \underline{x}) = \frac{P(\underline{x} \mid C_+) . P(C_+)}{P(\underline{x})}$$, Similarly for C_-

Naive Bayes assumes: $p(x \mid C_+) = \prod_{j=1}^{3} p(x_j \mid C_+)$

Since there are only 2 classes we can decide C_+ if

$$P(\underline{x} \mid C_+) . P(C_+) > P(\underline{x} \mid C_-) . P(C_-)$$

$$\prod_{j=1}^{3} P(x_j \mid C_+) . P(C_+) > \prod_{j=1}^{3} P(x_j \mid C_-) . P(C_-)$$

1b))[5pts] Classify the instance ($x_1 = 1$, $x_2 = 1$, $x_3 = 0$) using your classifier.

Circled in the table above are the class likelihoods for the given input.

$$\frac{1}{2} . \frac{1}{2} . 1 . \frac{2}{5} \overset{?}{>} \frac{1}{3} . \frac{1}{3} . \frac{1}{3} . \frac{3}{5}$$

$$\frac{1}{10} > \frac{1}{45}$$, since $P(\underline{x} \mid C_+) . P(C_+) > P(\underline{x} \mid C_-) . P(C_-)$

$\underline{x} = \begin{bmatrix} 1 \\ 1 \\ 0 \end{bmatrix}$ belongs to C_+

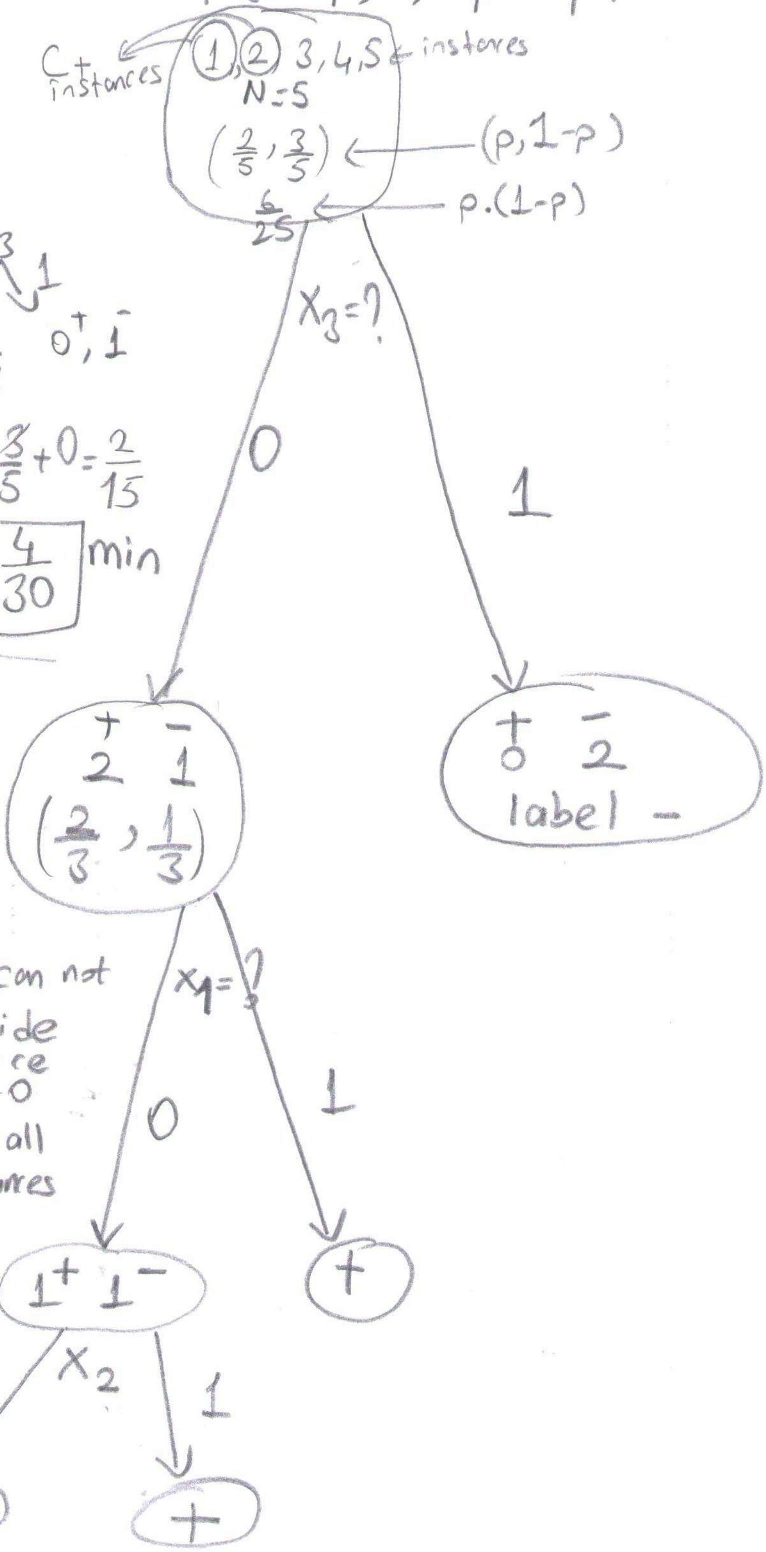
Name and Student ID:

Q2) [20pts]

Generate a decision tree for this dataset using Gini index (2p(1-p)) as the impurity measure.

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	-

For ease of computation, I will use $p(1-p)$, $p = p_+$



x_1: 0 → $\frac{1}{3}^+, \frac{2}{3}^-$; 1 → $\frac{1}{2}^+, \frac{1}{2}^-$

$\frac{2}{9}\cdot\frac{3}{5} + \frac{1}{4}\cdot\frac{2}{5} = \frac{2}{15} + \frac{1}{10} = \frac{7}{30}$

x_2: 0 → $\frac{1}{3}^+, \frac{2}{3}^-$; 1 → $\frac{1}{2}^+, \frac{1}{2}^-$

$\frac{2}{9}\cdot\frac{3}{5} + \frac{1}{4}\cdot\frac{2}{5} = \frac{7}{30}$

x_3: 0 → $\frac{2}{3}^+, \frac{1}{3}^-$; 1 → $0^+, 1^-$

$\frac{2}{9}\cdot\frac{3}{5} + 0 = \frac{2}{15}$ $\boxed{= \frac{4}{30}}$ min

Id	x_1	x_2	x_3	y
1	1	0	0	+
2	0	1	0	+
4	0	0	0	−

x_1: 0 → $\frac{1}{2}^+, \frac{1}{2}^-$; 1 → $1^+, 0^-$

$\frac{1}{4}\cdot\frac{2}{3} + 1\cdot0\cdot\frac{1}{3}$ $\boxed{= \frac{1}{6}}$

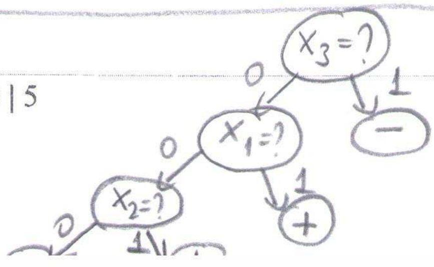
x_2: 0 → $\frac{1}{2}^+, \frac{1}{2}^-$; 1 → $1^+, 0^-$

$\frac{1}{4}\cdot\frac{2}{3} + 0$ $\boxed{= \frac{1}{6}}$

same

Choose x_1 or x_2

x_3 can not divide since $x_3 = 0$ for all instances



Name and Student ID:

Q3)[20pts]

The probability of a single observation x with mean rate parameter μ and variance 1 follows the following normal distribution:

$$P(x|\mu) = \frac{1}{\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2}}$$

You are given the data points $x_1\ x_2, \ldots\ x_n$ that are drawn independently from this distribution.

[5pts] Write down the log-likelihood of the data:

$$\log L = \log P(X|\underline{\mu}) = \log \prod_{i=1}^{n} p(x_i|\mu) = \sum_{i=1}^{n} \frac{1}{2} \log 2\pi - \frac{1}{2} \sum_{i=1}^{n} (x_i - \mu)^2$$

[15pts] Find the maximum likelihood estimate of the parameter μ:

$$\frac{d \log L}{d\mu} = -\frac{1}{2} \sum_{i=1}^{n} \frac{d(x_i-\mu)^2}{d\mu} = -1 \sum_{i=1}^{n} (x_i - \mu) = n\mu - \sum_{i=1}^{n} x_i$$

$$\frac{d \log L}{d\mu} = 0 \implies n\mu = \sum_{i=1}^{n} x_i$$

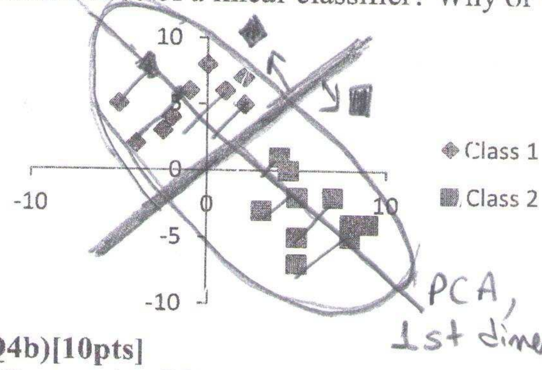
$$\boxed{\mu_{ML} = \frac{1}{n} \sum_{i=1}^{n} x_i}$$

Name and Student ID:

Q4) [20pts]

Q4a)[5pts]

Would using PCA as a preprocessing method on the following dataset reduce the performance of a linear classifier? Why or why not?



PCA, 1st dimension

PCA would not consider the class labels & project all instances on the given axis on the left. Since the instances in the reduced 1d space are separable by a poi in the projected space, PCA would not harm the performan of a linear classifier.

Q4b)[10pts]

What are the differences and similarities between the following clustering algorithms:

assumes $\Sigma_i = \sigma_i I_d$ for all clusters, and normal distributio for all cluster instances with mean $= \mu_i =$ cent of clust

3 K-means clustering:

- May not convergence, but fast (+)
- hard cluster membership

4 GMM clustering:

- May not convergence, slower than k-means faster than agglomerative clustering.

Assumes each cluster instances $\sim N(\mu_i, \Sigma_i)$

- Soft cluster membership

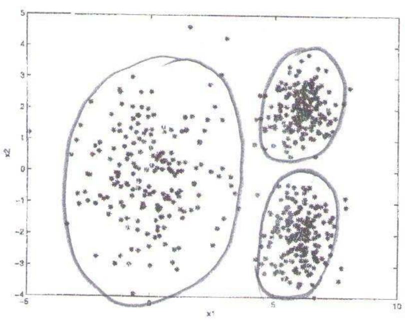
4 Hierarchical clustering using average link distance:

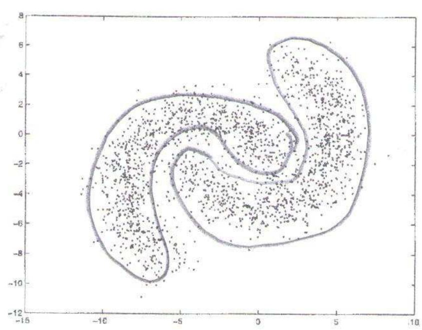
Slower than the other two methods. Can cluster data eve if it doesn't obey a certain distribution such as Gaussia

Take avg pointwise dist. as the distance between two clusters

Hard cluster membership.

Q4c)[5pts] Underneath each dataset, write down the clustering algorithm that you think is the most appropriate for the dataset, indicate the data clusters that would be obtained for appropriate clustering algorithm parameters.





Algorithm: GMM

since Σ_i's are different and not of the form $\sigma_i I_d$ for each cluster $\begin{bmatrix}\sigma_1^2 & \\ & \sigma_2^2\end{bmatrix}$

σ_2^2 is larger for each one.

Algorithm: Hierarchical clustering with average link distance or isomap

Name and Student ID:

Q5)[20pts]
Assume that g is a linear model and for input $x=[x_1\ x_2 \ldots x_d]$ which outputs:

$$g(x, w, w_0) = w^T x + w_0$$

You need to make the parameters w, w_0 as small as possible to avoid overfitting.

Given a dataset $X = \{x^t, r^t\}_{t=1}^N$, how would you obtain the solution for w, w_0?

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

Simplify notation $\underline{x} = [1\ x]$ and

$g(\underline{x}, \underline{w}) = \underline{w}^T \underline{x}$

new $\underline{w} = [\text{old}\,\underline{w}\ \ \underline{w}_0]$

$$E_\lambda = \frac{1}{N}\sum_{t=1}^{N}(g(\underline{x}^t, \underline{w}) - r^t)^2 + \lambda \underline{w}^T \underline{w}$$

$$E_\lambda = \frac{1}{N}\sum_{t=1}^{N}(\underline{w}^T x^t - r^t)^2 + \lambda \underline{w}^T \underline{w}$$

$$\frac{dE\lambda}{d\underline{w}} = \frac{2}{N}\sum_{t=1}^{N}(\underline{w}^T x^t - r^t).\underline{x}^t + 2\lambda \underline{w} = 0$$

$$\left(\frac{\cancel{2}}{N}\sum_{t=1}^{N} x^{t^T} x^t + \cancel{2}\lambda.I\right)\underline{w} = \frac{\cancel{2}}{N}\sum x^t r^t$$

$$\underline{w} = \left(\frac{1}{N}\sum_{t=1}^{N} x^{t^T} x^t + \lambda I\right)^{-1}\left(\frac{1}{N}\sum x^t r^t\right)$$

$\underline{x}^t$ is the old $\underline{x}^t$

$$\boxed{\begin{aligned} w &= \left(\frac{1}{N}\sum_{t}^{N} \underline{x}^{t^T} x^t + \lambda I\right)^{-1}\left(\frac{1}{N}\sum x^t r^t\right) \\ w_0 &= (I + \lambda I)^{-1}\left(\frac{1}{N}\sum r^t\right) \end{aligned}}$$

if $\frac{d(\ldots)}{dw_0} = 0$