2

Name and Student ID: ANSWERS

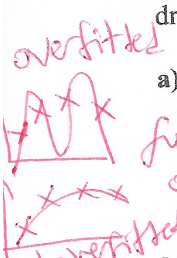
Machine Learning, BLG527E, October 26, 2017, Midterm Exam.

1 (20)	2 (20)	3 (20)	4 (15)	5 (15)	6 (10)	Total

Name:
Number:
Signature:
Duration: 120 minutes.
Write your answers neatly in the space provided for them.
Write your name on each sheet.
Books and notes are closed. Good Luck!

QUESTIONS

Q1) [20 points, 4 points each] What is (use at most three sentences per question, you can use drawings, formulas, etc. also):



a) overfitting Let training data X_N be produced by an underlying function $f \in F$. Let the model $g \in G$ learn X_N. Overfitting occurs when $E(g, X_N)$ is small but $E(g, X_M^V)$ on a validation set X_M^V is large. Overfitting can be due to N being too small or G being more complex than F.

b) VC dimension a classifier

Given N data points with 2^N possible labelings, $VC(H)$ of hypothesis class H is the maximum number of instances N for which all 2^N possible labelings of a data set X can be correctly classified by a classifier from H.

c) Bayes rule

$$\underbrace{P(C_i|x)}_{\text{posterior}} = \frac{\overbrace{p(x|C_i)}^{\text{likelihood}}.\overbrace{P(C_i)}^{\text{prior}}}{\underbrace{P(x)}_{\text{evidence}}} = \frac{p(x|C_i).P(C_i)}{\sum_{C_j} p(x|C_j).P(C_j)}$$

d) GMM (Gaussian Mixture Model)

GMM models an input prob. distribution $p(\underline{x})$ as a mixture of Gaussians: $p(\underline{x}) = \sum_{i=1}^{K} P(G_i).p(\underline{x}|G_i)$ where $p(\underline{x}|G_i) = N(\mu_i, \Sigma_i)$ and $P(G_i)$ are mixture probabilities. GMM clustering allows multiple and non identity covariances for each cluster and also soft cluster membership.

e) Mahalabonis Distance

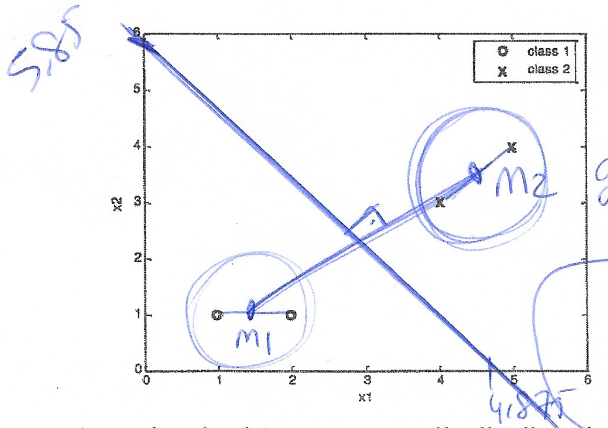
A distance measure that normalizes distances in each dimension according to the covariance matrix Σ:

$$d_M(\underline{x},\underline{y}) = (\underline{x}-\underline{y})^T \Sigma^{-1} (\underline{x}-\underline{y})$$

Name and Student ID:

$P(C_1) = P(C_2) = \frac{1}{2}$

Q2) [20 points] Consider the labeled data points: $X = \left\{ \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} 2 \\ 1 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} 4 \\ 3 \end{bmatrix}, 2 \right), \left(\begin{bmatrix} 5 \\ 4 \end{bmatrix}, 2 \right) \right\}$



4a) $g_i(\underline{x}) = \ln P(\underline{x}|C_i) + \cancel{\ln P(C_i)}$ since $P(C_1) = P(C_2)$

$g_i(\underline{x}) = -(\underline{x} - \underline{m}_i)^T(\underline{x} - \underline{m}_i)$

$g_i(\underline{x}) = 2\underline{m}_i^T \underline{x} - \underline{m}_i^T \underline{m}_i$ since $\underline{x}^T\underline{x}$ is the same for both C_1 & C_2

Assuming that inputs are normally distributed with class covariance matrices are as follows:

$S_1 = S_2 = s^2 I = s^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

4a) [20 pts] Compute the discriminant functions for both classes, $g_1(\underline{x})$ and $g_2(\underline{x})$.

4b) [5 pts] Compute and draw the discriminant function that separates the two classes.

Hint1: If $\mathbf{x} \sim N_d(\boldsymbol{\mu}, \Sigma)$, then the pdf for x is given by:

$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1}(\mathbf{x} - \boldsymbol{\mu})\right] \rightarrow \ln p(\underline{x}) = \underbrace{\ln 2\pi s}_{\text{const}} - \underbrace{\frac{1}{2s^2}}_{\text{const}}(\underline{x} - \mu)^T(\underline{x} - \mu)$$

$2\pi \quad |S| \quad \exp\left(-\frac{1}{2s^2}(\underline{x} - \mu)^T(\underline{x} - \mu)\right)$

Hint2: Use the log likelihood for the discriminant function.

Hint3: $(1.5)^2 = 2.25$, $(2.5)^2 = 6.25$, $(3.5)^2 = 12.25$, $(4.5)^2 = 20.25$

4a) $\underline{m}_1 = \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix} + \begin{bmatrix} 2 \\ 1 \end{bmatrix}\right)/2 = \begin{bmatrix} 3/2 \\ 1 \end{bmatrix}$ $\quad \underline{m}_2 = \begin{bmatrix} 9/2 \\ 7/2 \end{bmatrix}$

$g_1(\underline{x}) = 2\begin{bmatrix} 3/2 \\ 1 \end{bmatrix}^T \underline{x} - \begin{bmatrix} 3/2 \\ 1 \end{bmatrix}^T \begin{bmatrix} 3/2 \\ 1 \end{bmatrix}$ $\quad g_2(\underline{x}) = \begin{bmatrix} 9 \\ 7 \end{bmatrix}^T \underline{x} - \left(\frac{81}{4} + \frac{49}{4}\right)$

$g_1(\underline{x}) = \begin{bmatrix} 3 \\ 2 \end{bmatrix}^T \underline{x} - \left[\frac{9}{4} + 1\right] = \begin{bmatrix} 3 \\ 2 \end{bmatrix}^T \underline{x} - 13/4$ $\quad = \begin{bmatrix} 9 \\ 7 \end{bmatrix}^T \underline{x} - 130/4$

4b) $g(\underline{x}) = g_1(\underline{x}) - g_2(\underline{x}) = 0$ Discriminant function.

$= \begin{bmatrix} -6 \\ -5 \end{bmatrix}^T \underline{x} + 117/4$

$\Rightarrow -6x_1 - 5x_2 + 117/4 = 0$

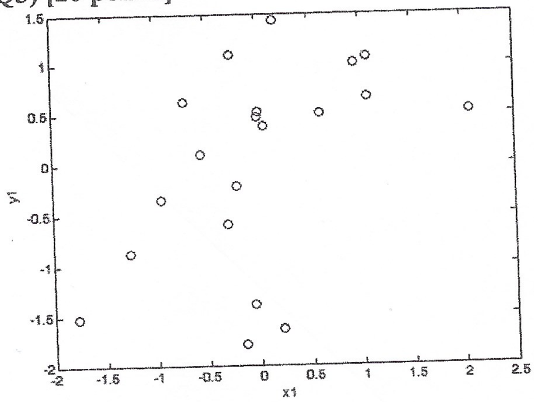
$x_1 = 0 \Rightarrow x_2 = \frac{117}{20} = 5,85$

$x_2 = -\frac{6}{5}x_1 + \frac{117}{20}$

$x_2 = 0 \Rightarrow x_1 = \frac{117}{20} \times \frac{5}{6} = 4,875$

Name and Student ID:

Q3) [20 points]



Data points shown in the figure above have been normalized to have zero mean $\mu = [0\ 0]^T$
The covariance matrix Σ, eigen vectors D of the covariance matrix are:

$$\Sigma = \begin{bmatrix} 0.7697 & 0.4255 \\ 0.4255 & 0.9905 \end{bmatrix} \quad D = \begin{bmatrix} -0.7909 & 0.6119 \\ 0.6119 & 0.7909 \end{bmatrix},$$

and the corresponding eigen values are $\lambda_1 = 0.4405$ and $\lambda_2 = 1.3197$.

Given this information, reduce the dimensionality of the following vector to one dimension using PCA:

$x = [1\ 2]^T$

***Hint:** Determinant of the matrix:* $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ *is ad-bc. Inverse of* $\begin{bmatrix} a & b \\ c & d \end{bmatrix}$ *is* $\frac{1}{ad-bc}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix}$

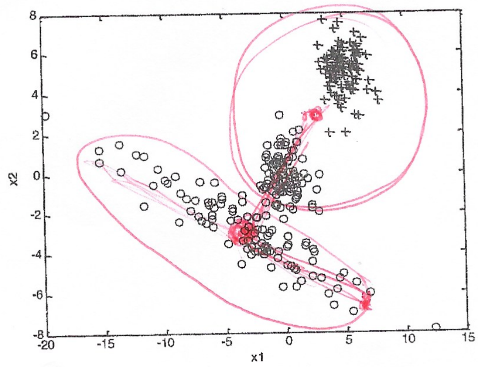
eigenvector corresponding to the largest eigenvalue

$$z_1 = [\underline{w}_1]^T \underline{x}$$

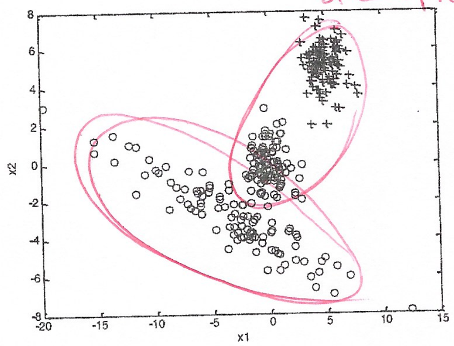
$$= \begin{bmatrix} 0,6119 \\ 0,7909 \end{bmatrix}^T \begin{bmatrix} 1 \\ 2 \end{bmatrix} = 2,1937$$

Name and Student ID:

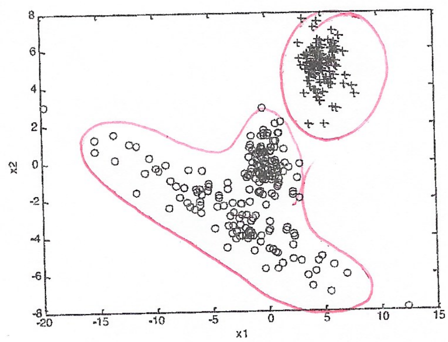
Q4) [15pts] Show (on the figure) the clusterings that would be obtained on the dataset using the clustering algorithm underneath (ignore the labels in the data). I should have specified what K was. Let's say K=2



K-means clustering



GMM clustering

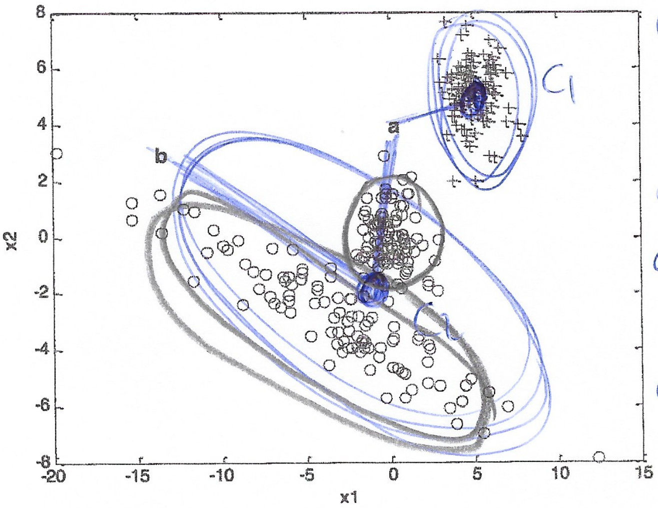


Hierarchical clustering using average link distance let C_1: + C_2: o

Name and Student ID:

let S_1 & S_2 be covariance matrices for C_1 & C_2, m_1, m_2 be means.

Q5) [15 points]
Given the following labeled data with two classes (+, o), assume that there are same number of + and o instances and o instances are equally separated between the two groups.
5a) Assume that you used multivariate classification, where each class can have a different and non-diagonal covariance matrix. How would you classify points a and b for this case?
5b) Could you do better classification by assuming some other distribution for class o inputs?



We would compute the Mahalabonis distance between a point and the class center:

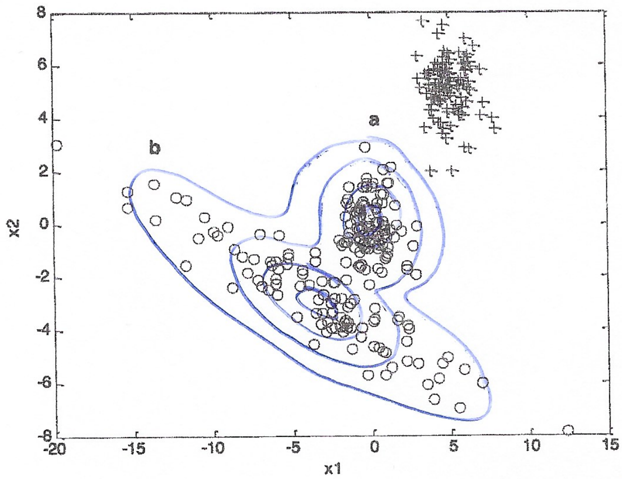
$$d_i(\underline{x}) = (\underline{x}-m_i)^T \Sigma_i^{-1}(\underline{x}-\underline{m_i})$$

b would be classified as C_1

a could be classified as C_1 or C_2 based on the Σ_1 and Σ_2 matrices.

5d) Draw the approximate equiprobable contours for class o if you were to use Parzen window density estimator.

5b) we would use GMM with 2 Gaussians for C_2. Then we would have a belong to C_1, because C_1 has more instances than the upper cluster of C_2



5c) Parzen window density estimator computes $p(\underline{x})$ as

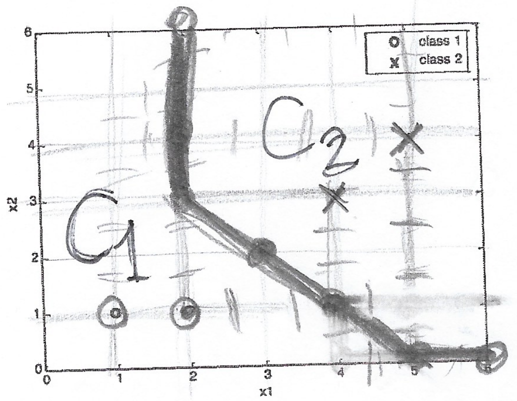
$$p(\underline{x}) = \frac{1}{N}\sum_{k=1}^{N} K(\underline{x}^k, \underline{x})$$

where $K(\underline{x}^k, \underline{x}) = \frac{1}{(2\pi)^{d/2}} \exp\left[-\frac{1}{2s^2}\|\underline{x}^k - \underline{x}\|^2\right]$

Name and Student ID:

Q6) [10 points]

Consider the labeled data points given as follows: $X = \left\{ \left(\begin{bmatrix} 1 \\ 1 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} 2 \\ 1 \end{bmatrix}, 1 \right), \left(\begin{bmatrix} 4 \\ 3 \end{bmatrix}, 2 \right), \left(\begin{bmatrix} 5 \\ 4 \end{bmatrix}, 2 \right) \right\}$



Compute distance between any point $\underline{x}$ and all training data points $\underline{x}^t$ as

$d(\underline{x}, \underline{x}^t) = |x_1 - x^t_1| + |x_2 - x^t_2|$

Show the decision regions for class 1 and class 2 when 1-nearest-neighbor classifier with L1 (sum of the absolute values of the differences at each coordinate) distance measure is used.