CS273A Final Exam

Introduction to Machine Learning: Winter 2021

Thursday March 18th, 2021

Your name: Brooke Ryan

Row/Seat Number: N/A

Your ID #(e.g., 123456789) 51809437

UCINetID (e.g.ucinetid@uci.edu) brooken

- **Declaration of Honor.** By submitting this exam you are making the following declaration: I hereby declare, upon my Honor, that this work is my own, and that in reaching an answer I have not assisted any other person, nor have been assisted by any other person in any form. I acknowledge that, had I cheated, I would have been the kind of person who cheats, and I do not wish to be the kind of person who cheats.
- Please put your name and ID **on every page**.
- Total time is 2 hours. READ THE EXAM FIRST and organize your time; don't spend too long on any one problem.
- Please **write clearly** and **show all your work**.
- Please ensure your final answer is contained in the space provided. We will not consider or grade anything beyond that space.
- If you need any clarification, please ask in our zoom room: `https://uci.zoom.us/j/94903054276`
- You may use **one** sheet containing handwritten notes for reference, and a (basic) calculator.
- Turn in your notes and any scratch paper with your exam.

Problems 30 mins/question

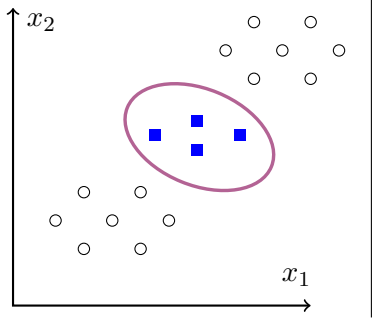
8 ~~**Dimensionality reduction,**~~ ***(10 points.)*** if time look @ end **22**

Total, *(100 points.)*

This page is intentionally blank, use as you wish.

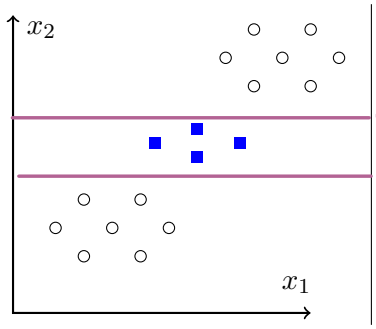
Problem 1 Separability, *(16 points.)*

For each of the following examples of training data and classifiers, state whether there exists a set of parameters that can separate the data and justify your answer briefly (∼1 sentence+plotting).



Perceptron with quadratic features, $[x_1\ x_2\ x_1x_2\ x_1^2\ x_2^2]$: *(4 points.)*

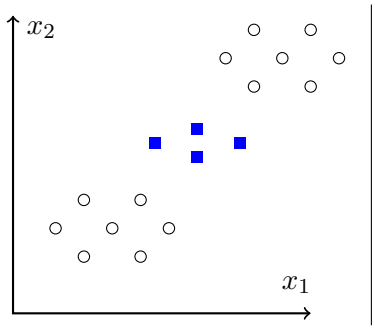
yes, this sketched circular graph requires quadratic form, due to the equation of a circle being degree 2.



Depth two decision tree: *(4 points.)*

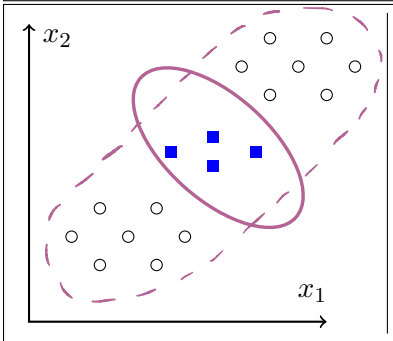
depth 1
depth 2

Yes, using a 2-depth decision tree we can split the data



Gaussian Bayes Classifier with equal covariance: *(4 points.)*

NO.
A gaussian bayes classifier with equal covariance would be a linear classifier, and since the data is not linearly separable, this isn't possible.



Gaussian Bayes Classifier with unequal covariance: *(4 points.)*

yes, the data can be separated with an unequal covariance, this can be achieved through sharply peaked Gaussian on the squares.

This page is intentionally blank, use as you wish.

Problem 2 True or False statements, *(10 points.)*

For each of the following statements, choose whether the statement is true or false.

Statement	True	False
In PCA, if we double the number of features from n to $2n$ by including two exact copies of each original feature, we need to also increase the number of principal components from k to $2k$ to get the same latent representation.	☐	☒
If there exists a set of h instances that cannot be shattered by any $f_\theta(x)$, this implies that the VC dimension of f_θ is less than h.	☐	☒
In Random Forests, increasing the number of trees that are trained and averaged over usually reduces the variance without increasing the bias.	☐	☒
One advantage of the dual form of the SVM over the primal (linear) form is its ability to work in infinite dimensional feature spaces.	☒	☐
The expectation maximization (EM) algorithm for Gaussian mixture models is guaranteed to converge to the global maximum of the data log likelihood.	☐	☒
Given a fixed policy for a Markov decision process, the value of each state can be computed by solving a linear system of equations.	☒	☐
A good k in k-Means clustering can be selected by minimizing the mean square error of held out validation data.	☐	☒
In boosting, the different classifiers in the ensemble can be learned in parallel to reduce the time it takes to finish learning.	☐	☒
Both boosting and bagging usually increase the gap in complexity between individual classifiers and the entire ensemble.	☐	☒
In k-armed bandits, once we are sufficiently confident that action a is the best, we can minimize our long-term regret by always thereafter taking the action a.	☒	☐

This page is intentionally blank, use as you wish.

Problem 3 Factorization and Gradient Descent, *(16 points.)*

We will consider a matrix factorization problem, where we have a 2×2 binary matrix X that we want to factorize using 1-dimensional factors, U' and V (i.e. U' and V are 2×1 matrices). We want to make sure that U' is a positive matrix, which can be achieved by substituting $U' = \exp(U)$, where U is real-valued and has the same dimensions as U'.

Since X is a binary matrix, we will use a *logistic sigmoid*, σ to ensure the predictions are always between $(0, 1)$, i.e. $\hat{X}_{ij} = \sigma(\exp(U_i)V_j)$. In the following, you may need to use the facts that $\sigma(x) = \frac{1}{1+\exp(-x)}$ and $\sigma(-x) = 1 - \sigma(x)$.

(1) Let's use the logistic loss function[1] as our loss. Write the equation for this loss when $X_{ij} = 1$ (J_1) and when $X_{ij} = 0$ (J_0) in terms of U_i and V_j. Put final answers in the boxes. *(4 points.)*

$P(X_{ij}=1 \mid U_i, V_j) = \log\left(\sigma(\exp(U_i)V_j)\right)$

$J_1(U_i, V_j)$, when $X_{i,j} = 1$ $\quad J_1 = \log\left(\sigma(\exp(U_i)V_j)\right)$

$P(X_{ij}=0 \mid U_i, V_j) = \log\left(\sigma(-\exp(U_i)V_j)\right) = \log\left(1 - \sigma(\exp(U_i)V_j)\right) =$

$J_0(U_i, V_j)$, when $X_{i,j} = 0$ $\quad J_0 = \log\left(-\sigma(\exp(U_i)V_j)\right)$

(2) Derive the gradient of these losses, with respect to U_i and V_j. Put final answers in the box. *(6 points.)* $\frac{d}{dx}\log(x) = \frac{1}{x}$

Sigmoid deriv. =

$J_1 \frac{d}{dU_i} = \frac{\log(-\sigma(\exp(U_i)V_j))}{-\sigma(\exp(U_i)V_j))} \cdot \frac{\partial}{\partial U_i}(-\sigma(\exp(U_i)V_j))$

$= -V_j\left[\sigma(\exp(U_i)V_j))(1-\sigma(\exp(U_i)V_i))\right]$

$\frac{\partial}{\partial U_i} J_1(U_i, V_j) =$ $\quad V_j \log\left(\sigma(\exp(U_i)V_j)\right)(1-\sigma(\exp(U_i)V_j))$

$\frac{\partial}{\partial V_j} J_1(U_i, V_j) =$ $\quad U_i \log\left(\sigma(\exp(U_i)V_j)\right)(1-\sigma(\exp(U_i)V_j))$

$\frac{\partial}{\partial U_i} J_0(U_i, V_j) =$ $\quad -V_j \log\left(-\sigma(\exp(U_i)V_j)\right)(1-\sigma(\exp(U_i)V_j))$

$\frac{\partial}{\partial V_j} J_0(U_i, V_j) =$ $\quad -U_i \log\left(-\sigma(\exp(U_i)V_j)\right)(1-\sigma(\exp(U_i)V_j))$

Continued on next page

[1]Hint: if you don't remember the loss function, it can be derived as the negative log-likelihood of the probabilistic model $P(X_{ij} = 1|U_i, V_j) = \sigma(\exp(U_i)V_j)$, and $P(X_{ij} = 0|U_i, V_j) = \sigma(-\exp(U_i)V_j)$.

This page is intentionally blank, use as you wish.

(3) Consider the following dataset and parameters:

i	j	X_{ij}
0	0	1
0	1	0
1	0	0
1	1	1

i	U_i
0	0
1	0

j	V_j
0	0
1	0

Compute the loss, J on this data, i.e. $J_1(U_0, V_0) + J_0(U_0, V_1) + J_0(U_1, V_0) + J_1(U_1, V_1)$. Also compute the gradient of this loss. Put your answers in the boxes. *(4 points.)*
You may need to use $\log(0.5) = -0.693$.

$J = -4(-0.693)$
$= +2.772$

$J =$ 2.772

b/c everything subtracted?

$\frac{\partial}{\partial U_0} J =$ 0

$\frac{\partial}{\partial U_1} J =$ 0

$\frac{\partial}{\partial V_0} J =$ 0

$\frac{\partial}{\partial V_1} J =$ 0

(4) Would taking gradient descent steps on the loss J to update the parameters U and V will result in better U and V, i.e. ones that will achieve a lower loss? Put "Yes" or "No" in the box. *(2 points.)*

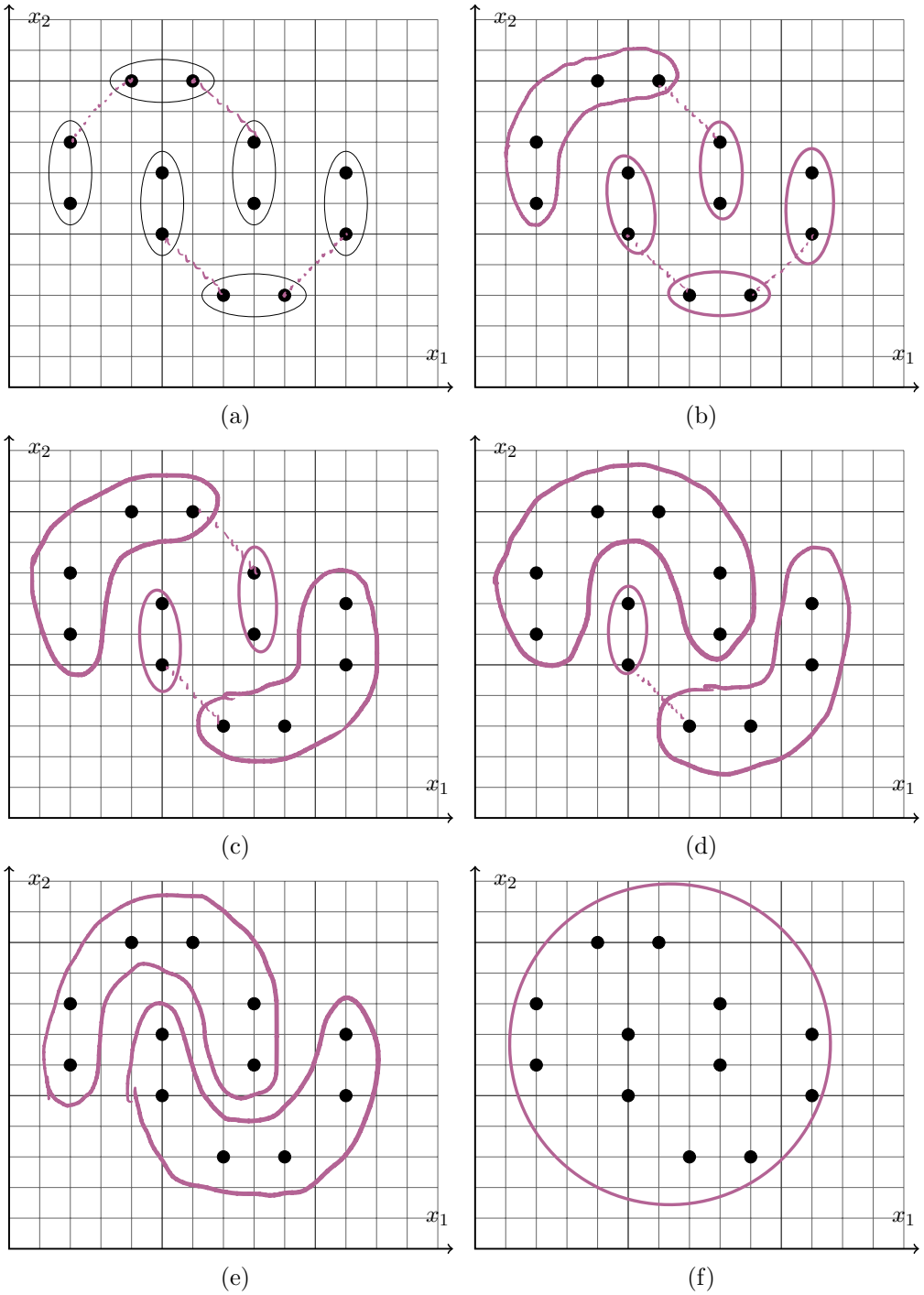
No

This page is intentionally blank, use as you wish.

Problem 4 Single-linkage Agglomerative Clustering, *(6 points.)*

For *twelve* points on a two-dimensional space, we have performed the first 6 steps of agglomerative clustering for you, which has produced the first figure below.

In the next five figures, show each step of single-linkage agglomerative clustering, continued from the first figure. Break ties arbitrarily. *(6 points.)*



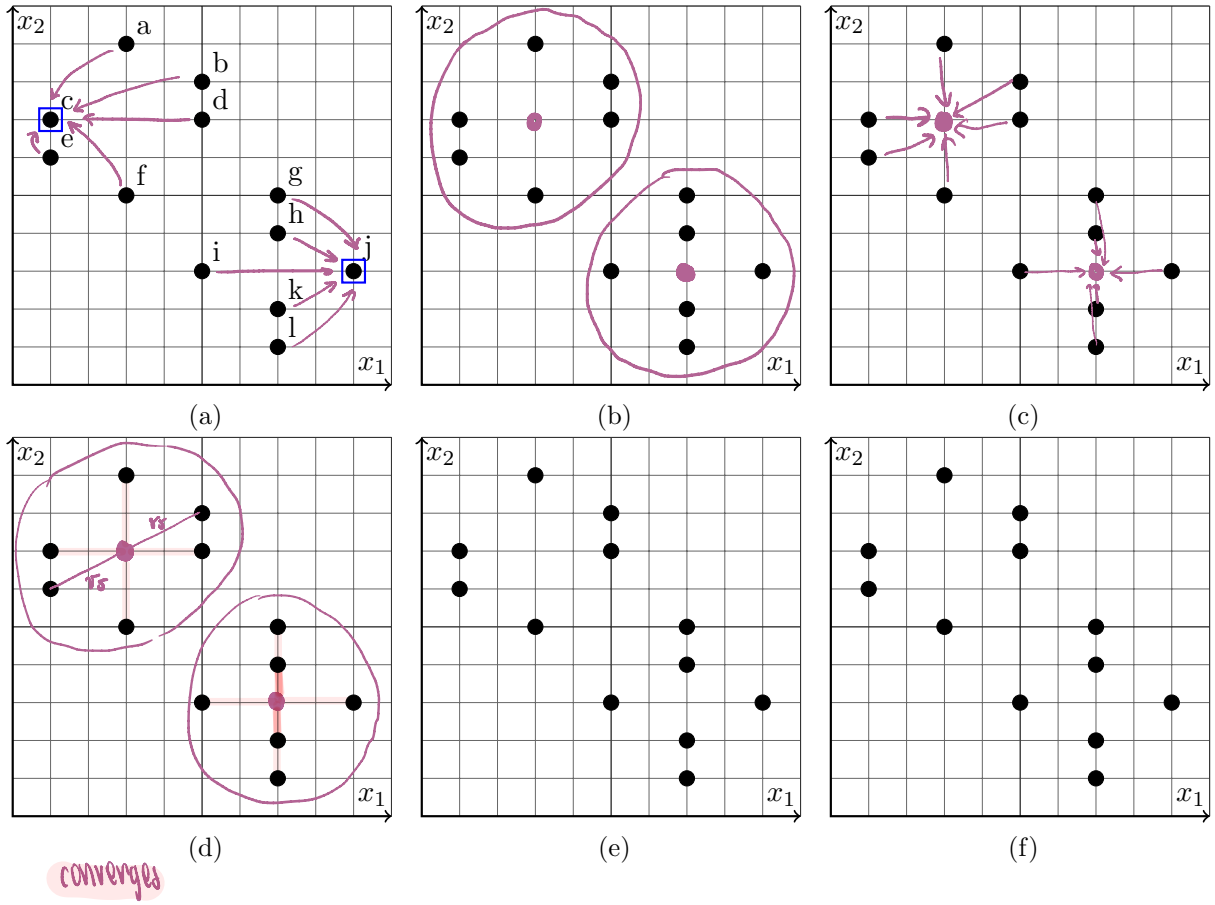
This page is intentionally blank, use as you wish.

Problem 5 K-Means Clustering, *(18 points.)*

Here are 12 other points in two dimensions, and we will now consider K-Means, $K = 2$.

(1) With two initial cluster shown as squares, you need to perform the steps of K-Means clustering, showing the cluster assignments as arrows in the odd figures (a, c, e), and the updated cluster means in the even figures (b, d, f). In case of technical difficulties, you may also write down the coordinates of cluster means, and a list of all points (letter-coded in figure (a)) and their assignments. Show the first 6 steps, or till convergence. *(4 points.)*

For all questions in this problem, round the cluster means to the nearest integer. Assume each grid point is of length 1, i.e. the initial clusters are at $(1, 7)$ and $(9, 3)$.



converges

(2) Compute the cost of the final clustering that K-Means minimizes, and put in the box. *(4 points.)*

$4 \times 2^2 + 2 \times 2^2 + 2 \times (\sqrt{5})^2 =$

$6 \times 2^2 + 2 \times 5$

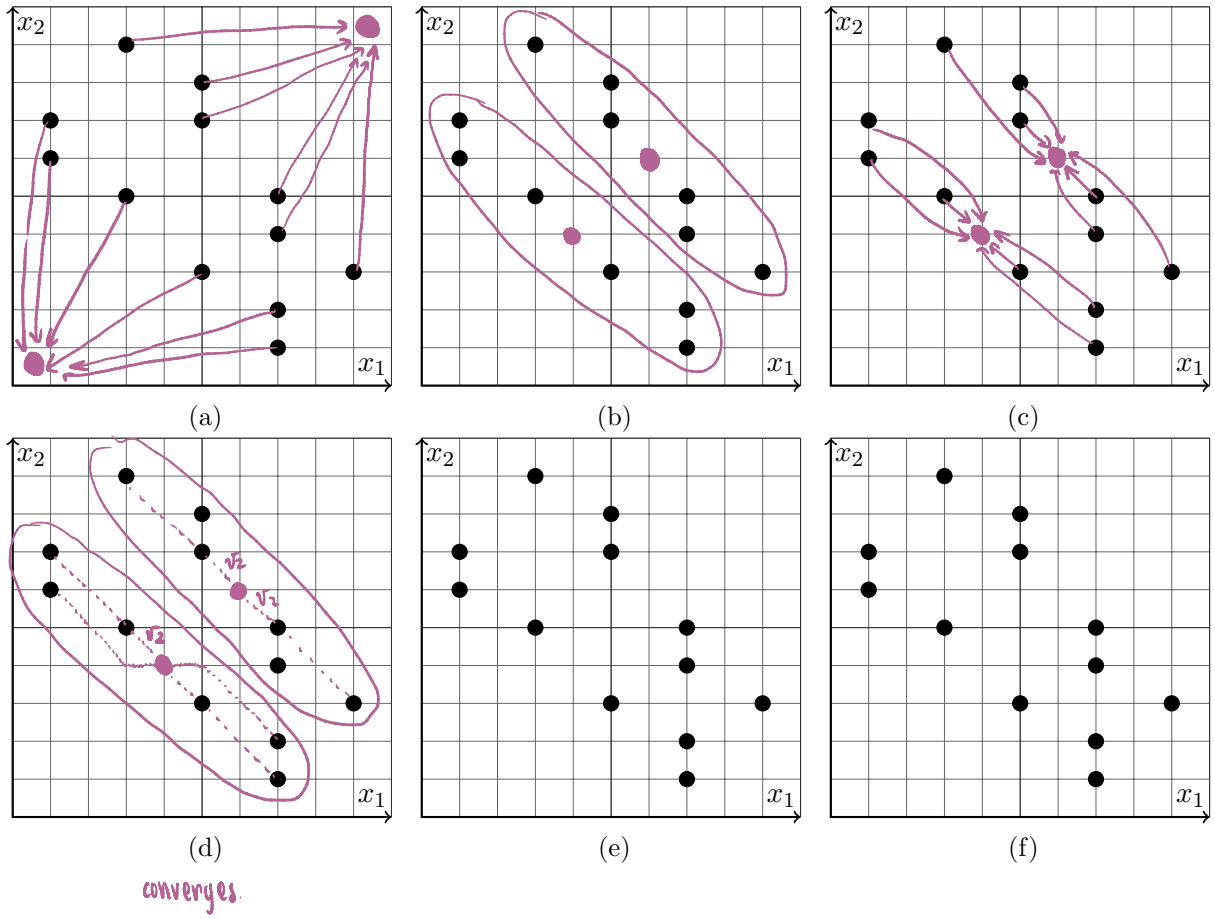
$= 24 + 10 = 34$

34

Continued on next page.

This page is intentionally blank, use as you wish.

(3) Provide two initial clusters in (a) below that you think will result in a different clusters when K-means converges. Walk through six steps of clustering, or till convergence, showing the cluster assignments as arrows in (a, c, e), and updated means in (b, d, f). *(4 points.)*



converges.

(4) Compute the cost of the final clustering above, and put your answer in the box. *(4 points.)*

$4(\sqrt{2})^2 + 4(3\sqrt{2})^2 + 4(\sqrt{13})^2$

$= 4(2) + 4(9)(2) + 4(13)$

$= 8 + 72 + 52$

132

(5) Which clustering do you think is better? Check/cross the corresponding box. *(2 points.)*

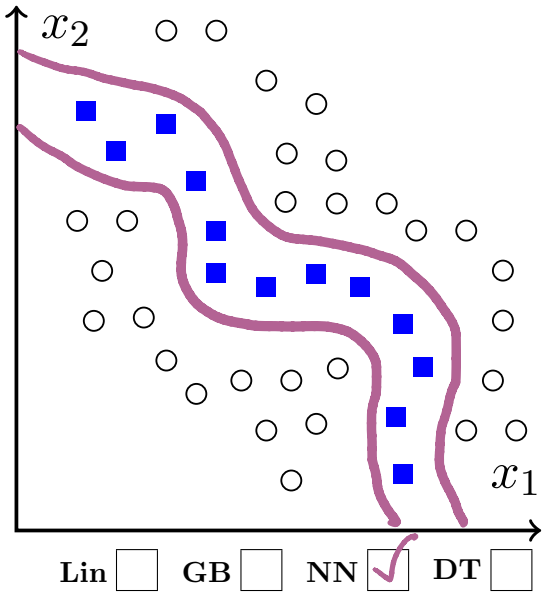
Previous clustering, from (1) [✓] This clustering, from (3) [ ]

This page is intentionally blank, use as you wish.

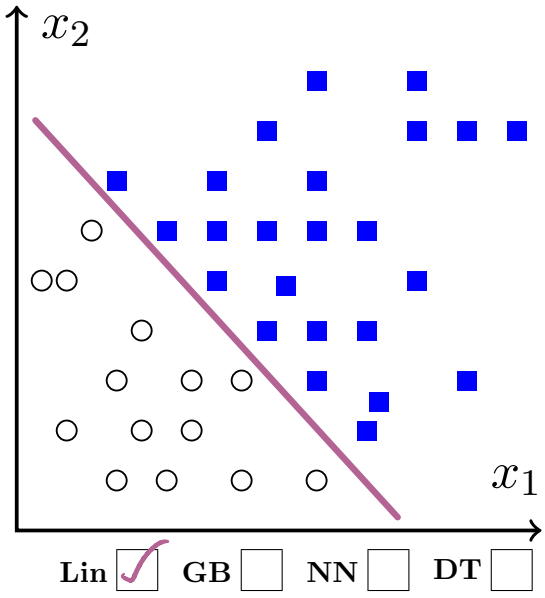
Problem 6 Classification, *(16 points.)*

For each of the datasets shown in two dimensions where the label is indicated by the shape, choose which classifier amongst the options would work *best*. For each question, **choose only one option** (by checking the corresponding box), and **sketch the boundary** that explains your choice. The options are:

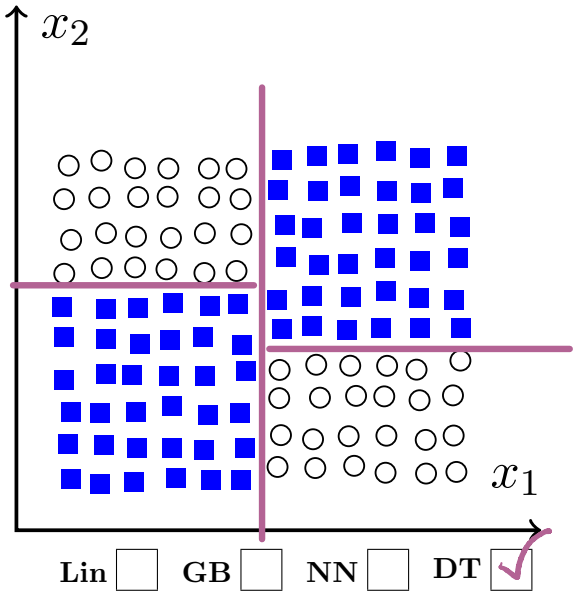
- **Lin:** Linear Classifier with three parameters, i.e. $f(\mathbf{x}) = \theta_0 + \theta_1 x_1 + \theta_2 x_2$
- **GB:** Gaussian Bayes Classifier, with different, non-spherical covariance allowed for each class
- **NN:** K-Nearest Neighbor classifier, with $K = 2$
- **DT:** Decision trees, limited to depth 2



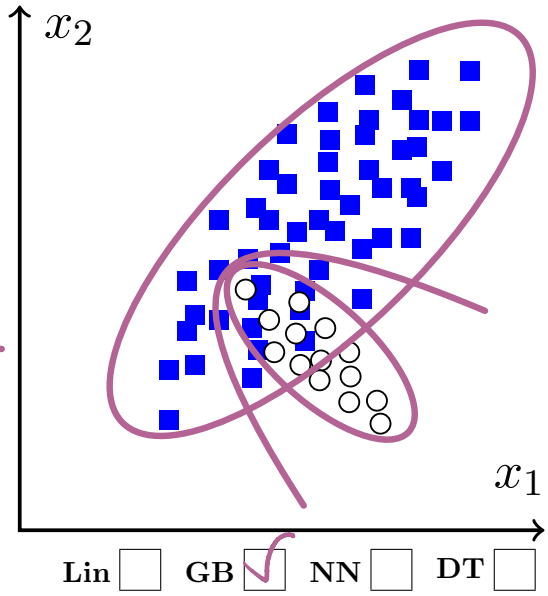
(a) *(4 points.)*



(b) *(4 points.)*



(c) *(4 points.)*

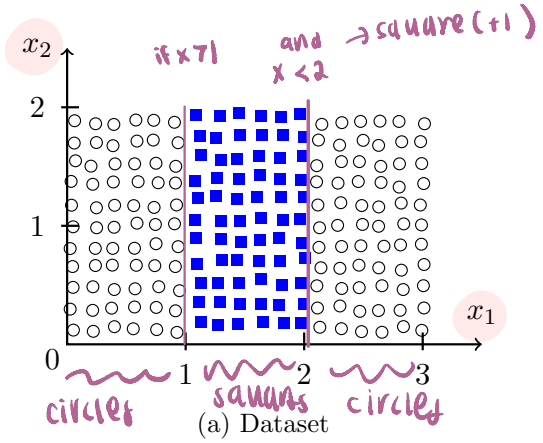


(d) *(4 points.)*

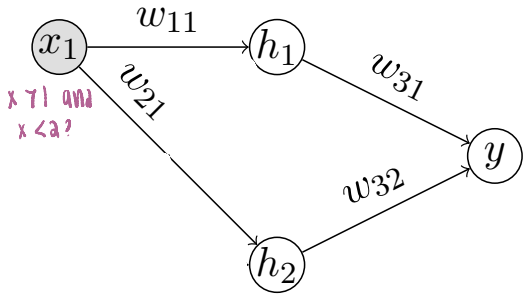
This page is intentionally blank, use as you wish.

Problem 7 Neural Networks, *(8 points.)*

In this problem, we will be considering classification in a simple dataset, shown below. The dataset is two-dimensional and consists of two classes, +1 as squares, and -1 as circles. Clearly, this dataset is not linearly separable, so let's try to learn a neural-network classifier.



(a) Dataset



(b) Neural model

The neural network we will use consists of one hidden layer, and uses the threshold function as the activation function, $T[z] = \begin{cases} +1 & z > 0 \\ -1 & \text{o.w.} \end{cases}$. Given this, we can write the neural network as:

$$h_1 = T[w_{10} + w_{11}x_1 + w_{12}x_2]$$
$$h_2 = T[w_{20} + w_{21}x_1 + w_{22}x_2]$$
$$y = T[w_{30} + w_{31}h_1 + w_{32}h_2]$$

Clearly, $h_1, h_2, y \in \{-1, +1\}$. We show this neural network above, with w_{10}, w_{20}, w_{30} implicit. Suggest values for all 9 weights so that the neural network achieves 0 loss.

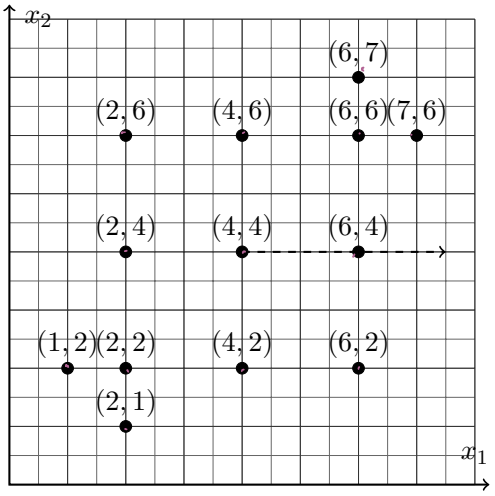
Hint: Think of each hidden unit as a linear classifier.

$w_{10} = -1$	$w_{11} = 1$	$w_{12} = 0$
$w_{20} = 2$	$w_{21} = -1$	$w_{22} = 0$
$w_{30} = 0$	$w_{31} = 1$	$w_{32} = 1$

This page is intentionally blank, use as you wish.

Problem 8 Dimensionality reduction, *(10 points.)*

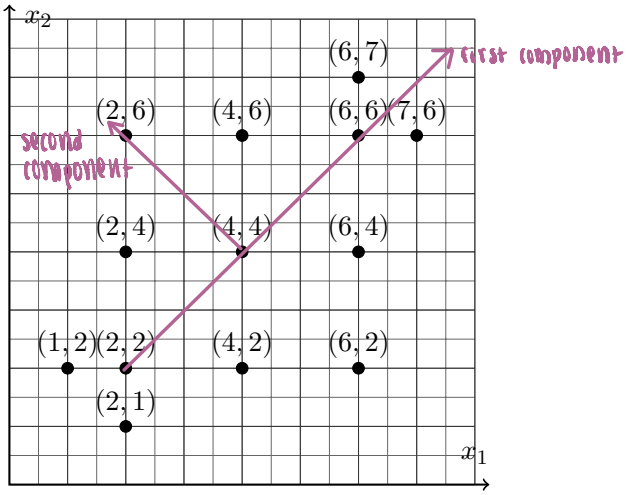
(1) For the following points in two dimentions, consider dimensionality reduction along the given vector (dashed line from $(4, 4)$ to the right). What is the reconstruction error, in MSE, when using this vector?
(Put answer in box) *(5 points.)*



$6(2)^2 + 2(2)^2 + 2(3)^2$
$= 6(4) + 2(4) + 2(9)$
$= 24 + 8 + 18$
$= \frac{50}{13}$ points

$\frac{50}{13}$

(2) On the figure below, draw directions of the first two principal components. *(5 points.)*



This page is intentionally blank, use as you wish.

This page is intentionally blank, use as you wish.

This page is intentionally blank, use as you wish.