CS 445/545: Machine Learning Name:_____Parth Parashar___________________

Portland State University

Midterm (43 pts. Possible), Fall 2021

Rhodes

Please show a sufficient amount of work for full credit on exercises. If the question prompts you to simply provide an answer, it is acceptable to simply submit the answer (without explanation, if none is necessary). The exam is open book and open notes. **However, you are not allowed to confer with fellow students or, nor are you permitted to use "le Google" to seek out a solution**.

*Email your exam solutions to our grader by the assigned due date. You can submit typed or hand-written solutions (or a combination of both if this is preferred); please make an effort to ensure that your solutions are clear and legible.

1. (1 pt.) Soft-margin SVMs, defined with slack variables ξ_i always admit of a solution. **True** **False**

Answer: - True

2. (1 pt.) A zero training set error necessarily indicates good generalization performance. **True** **False**

Answer: - False

3. (1 pt.) Recall the general expression for backprop weight updates: $\Delta w_{ji}^{t} = \eta \delta_j x_i + \alpha \Delta w_{ji}^{t-1} - \lambda w_{ji}^{t-1}$. Explain the role of the following terms (just in a sentence):

$\alpha \Delta w_i^{t-1}$:

λw_i^{t-1}:

Answer: - In the screenshot below

NAME - PARTH PARASHAR
PSUID - 923928157

Ans 3) According to the question, we have,

$$\Delta w^{t}_{ji} = \eta \delta_j x_j + \alpha \Delta w_{ji}^{t-1} - \lambda w_{ji}^{t-1}$$

(A) $\alpha \Delta w_{ji}^{t-1}$:→ This is term in the formula presented above which is responsible for the weight changes to move in the same direction.

(B) $-\lambda w_{ji}^{t-1}$:→ To reduce the overfitting of model, this weight decay term encourages the weights in the network to be small, and thereby providing the below benefits as well,

① Smoothens the decision boundary
② It does not allow the weights to get too large and in the process, stabilizes, the back propagation.

4. (1pt.) The PLA (perceptron learning algorithm) always converges in a finite number of steps. **True** **False**

Answer:- False

5. (1 pt.) Given a finite data set, there exists a finite dimensional vector space for which the data is linearly separable. **True** **False**

Answer: - True

6. (1 pt.) Given a dataset $\{\mathbf{x}_i\}_{i=1}^{N}$ of N datapoints, where $\mathbf{x}_i \in \mathbb{R}^d$, if $d >> N$, we say that these data are: **Low-Dimensional** **High-Dimensional**

Answer: - High Dimensional Data

7. (1 pt.) A model with infinite VC dimension can have a finite number of parameters.
True **False**

Answer: - True

8. (1 pt.) Suppose we wish to calculate $P(H|X_1, X_2)$ and we have no conditional independence information. Which of the following sets of numbers are sufficient for the calculation? (circle)

(i) $P(X_1, X_2), P(H), P(X_1|H), P(X_2|H)$

(ii) $P(X_1, X_2), P(H), P(X_1, X_2|H)$

(iii) $P(H), P(X_1|H), P(X_2|H)$

Answer: - (ii) $P(X1, X2), P(H), P(X1, X2|H)$

9. (1 pt.) Consider the sigmoid function: $f(x) = \dfrac{1}{1+e^{-x}}$. Which expression is equal to $f'(x)$? (circle)

(i) $f(x)\log(1-f(x))$

(ii) $f(x)(1-f(x))$ **Answer: - (ii)**

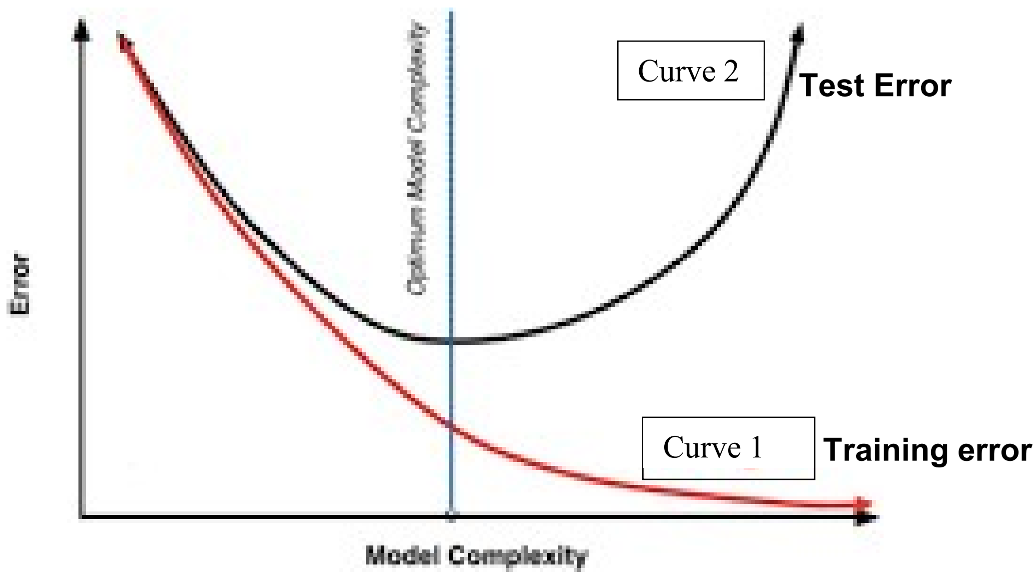
(iii) $\dfrac{1}{f(x)}$

(iv) $\aleph_0$

10. (1 pt.) Suppose you are given a dataset of cellular images from patients with and without cancer and that you are required to train a classifier that predicts the probability that the patient has cancer. The dataset has 900 cancer-free images and 100 images from cancer patients. If I train a classifier which achieves 85% accuracy on this dataset, should we consider this to be a good classifier? Explain your answer in a sentence or two.

Answer: - This is not considered to be a good classifier. It's because even if the classifier makes wrong predictions while predicting the outcomes and all the images are classified as cancer-free, the model will still have an accuracy of 90%. which is better than the 85% accuracy of the other model making it to not be a good classifier.

11. (1 pt.) The plots of training and test error are shown as a function of model complexity below. Appropriately identify the plots.



Training Error: Curve 1 Curve 2 (circle)

Answer:- For training error:- Curve 1 (RED)

Test Error: Curve 1 Curve 2

Answer:- For test error:- Curve 2 (BLACK)

12. (2 pts.) Suppose that the prevalence of a disease is 1%. This disease can be screened by a medical test that is 90% accurate. This means that the test result is positive about 90% of the times when it is applied on patients who have the disease and that the test result is negative about 90% of the time when it is applied on patients who do not have the disease. Suppose that you take the test and the test shows a positive result. How likely is it that you have the disease?

NAME- PARTH PARASHAR

PSU ID- 923928157

Q12 ⇒ from the question, we have,

Prevalence of disease = 1%.
let us assume that there are 100 people.
⇒ 1 in every 100 people would be affected by the disease

⇒ $TP = 1$

Now, positive Rate of test Results = 90%.
⇒ when applied on 1 of 100 ⇒ 90% of (1 of 100)

$\Rightarrow (0.9)$

$\Rightarrow TN = (1 - 0.9) = (0.1)$

$\Rightarrow FP = 99 * 0.9$ ⇒ People not having disease show (+ve)
$FP = 89.1$

⇒ To calculate the I have disease if test is (+ve) is:-

$$\left(\frac{TP}{TP + FP}\right) = \frac{1}{1 + 89.1}$$

$$\Rightarrow \frac{1}{90.1} \Rightarrow 0.011$$

13. (2pts.) Recall that the with LDA dimensionality reduction we wish to solve the following optimization problem: $\arg\max_{w} \frac{w^T S_B w}{w^T S_W w}$. In a few simple sentences, explain the meaning of this expression with respect to LDA.

NAME - PARTH PARASHAR

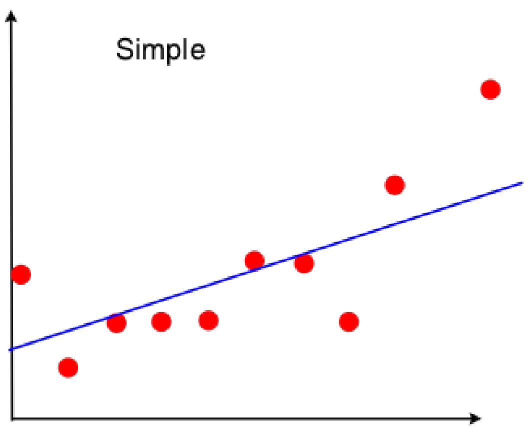
PSU ID - 923928157

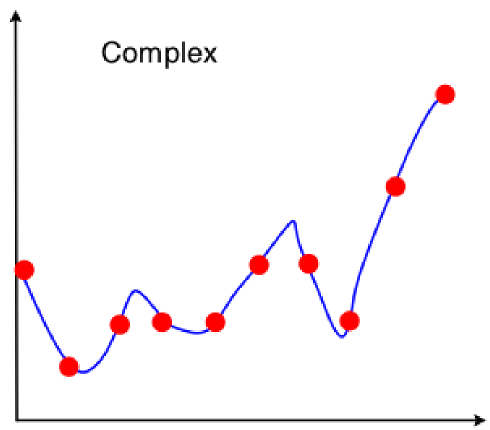
Q13 ⇒ optimization problem :- $\arg\max \frac{w^T S_B w}{w^T S_W w}$

Here, for the above ratio, we have to find a maximum value of the ratio with respect to w.
This has to be done with respect to w and LDA.
⇒ we have to find a vector 'w' onto which the projection of data point of each class has to be made.
This projection has to be made in such a way that it maximizes the ratio of scatter between classes and scatter within classes as well.

This in turn will give us a resultant vector 'w' which will preserve the classification integrity of the original dataset.

14. (2 pts.) Using the following mathematical models (i.e. the curves shown), explain briefly the idea of the "*bias-variance tradeoff*" in machine learning, how it relates to model *complexity* and the notions of *overfitting* and *underfitting*. In relation to these ideas, elaborate on what is means to have a "good" predictive model.





Answer: -

NAME - PARTH PARASHAR

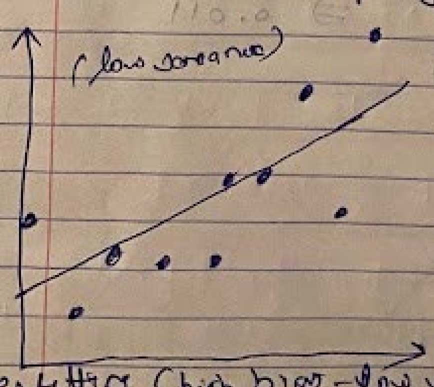
PSU ID - 923928157

Q14 ⇒ The bias - variance trade-off is the property of a model that, if the variance of the parameters (model) across the model have to be reduced, then it is achieved by increasing the bias among the estimated parameters.

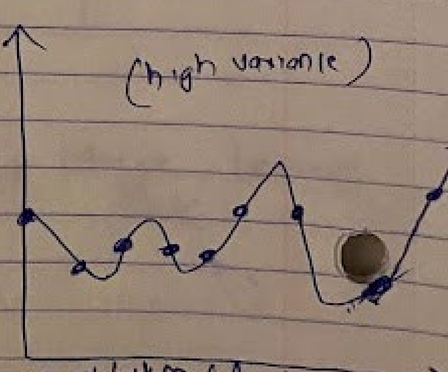
This can also be defined as a conflict b/w two major types of error where if one is reduced, the other will increase and vice-versa.
Both these errors (bias and variance) prevent a supervised learning algorithms from generalizing beyond their training set.

(A) BIAS ERROR :- Error from erronous assumptions in the learning algorithm.
High bias can cause an algorithm to miss the relevant relations b/w features and target outputs. This condition is underfitting

(B) VARIANCE :- Error from sensitivity to small fluctuations in the training set.
High variance results from noise in the algorithm
This is overfitting.



underfitting (high bias - low variance)



overfitting (low bias)

NAME — PARTH PARASHAR
PSUID - 923928157

A "good" predictive model represents that the model should be able to generate predictions for new unseen data as well as the model should have low bias and low variance.

15. (1 pt.) "t-SNE" is an example of which type of general ML algorithm: (circle)

(i) classification (ii) regression (iii) dimensionality reduction (iv) backpropagation

Answer:- (iii) Dimensionality Reduction

16. (2 pts.) Let $\mathbf{x} = (x_1, x_2)$. Using the feature mapping

$$<\lambda(\mathbf{x}) = \left(x_1^2,\ \sqrt{2} \times x_1 x_2, x_2^2\right)$$

show that

$$<\lambda\big((2,3)\big) \times <\lambda\big((4,4)\big) = \big((2,3) \times (4,4)\big)^2$$

NAME - PARTH PARASHAR
PSU ID - 923928157

Ans 16) Given $\phi(x) = (x_1^2, \sqrt{2} \cdot x_1 x_2, x_2^2)$

$\phi((2,3)) \cdot \phi((4,4)) = ((2,3) \cdot (4,4))^2$

$\Rightarrow [((2)^2, 2.3.\sqrt{2}, (3)^2) \cdot ((4)^2, 4.4.\sqrt{2}, (4)^2)] = ((8+12))^2$

$\Rightarrow [4, 6\sqrt{2}, 9] \cdot [16, 16\sqrt{2}, 16] = ((8+12))^2$

$\Rightarrow [64 + 192 + 144] = (20)^2$

$\Rightarrow [400] = (400)$

$\Rightarrow$ Hence Proved.

17. (5 pts.) ***Gradient Descent***. Consider the multivariate function: $f(x, y) = x^2 + y^2$

Devise an iterative rule using gradient descent that will iteratively move closer to the minimum of this function. Assume we start our search at an arbitrary point: (x_0,y_0). Give your update rule in the conventional form for gradient descent, using η for the learning rate.

(i) Write the explicit x-coordinate and y-coordinate updates for step (i+1) in terms of the x-coordinate and y-coordinate values for the ith step.

$x^{(i+1)} \leftarrow$

$y^{(i+1)} \leftarrow$

(ii) Briefly explain how G.D. works, and the purpose of the learning rate.

(iii) Is your algorithm guaranteed to converge to the minimum of f (you are free to assume that the learning rate is sufficiently small)? Why or why not?

(iv) Re-write your rule from part (i) with a momentum term, including a momentum parameter α.

NAME - PARTH PARASHAR
PSU ID - 923928157

Ans 17) (i) $f(x,y) = x^2 + y^2$

Now, as we know that $\theta_{t+1} = \theta_t - \eta \nabla_\theta R|\theta_t$

$$\Rightarrow x^{(i+1)} = x^i - \eta \frac{d f(x,y)}{dx}$$

$$\Rightarrow x^{(i+1)} = x^i - \eta \frac{d(x^2+y^2)}{dx}$$

Now, by using the partial derivative Rule, we have,

$$\Rightarrow x^{(i+1)} = x^i - \eta (2x)^i \quad \left\{ \text{as } \frac{d f(x,y)}{dx} = 2x \right\}$$

Doing the same steps for $f(x,y)$ and taking partial derivative for 'y', we have,

$$\Rightarrow y^{i+1} = y^i - \eta \frac{d f(x,y)}{dy}$$

Putting the values of $f(x,y)$ in the eq^n above, we have,

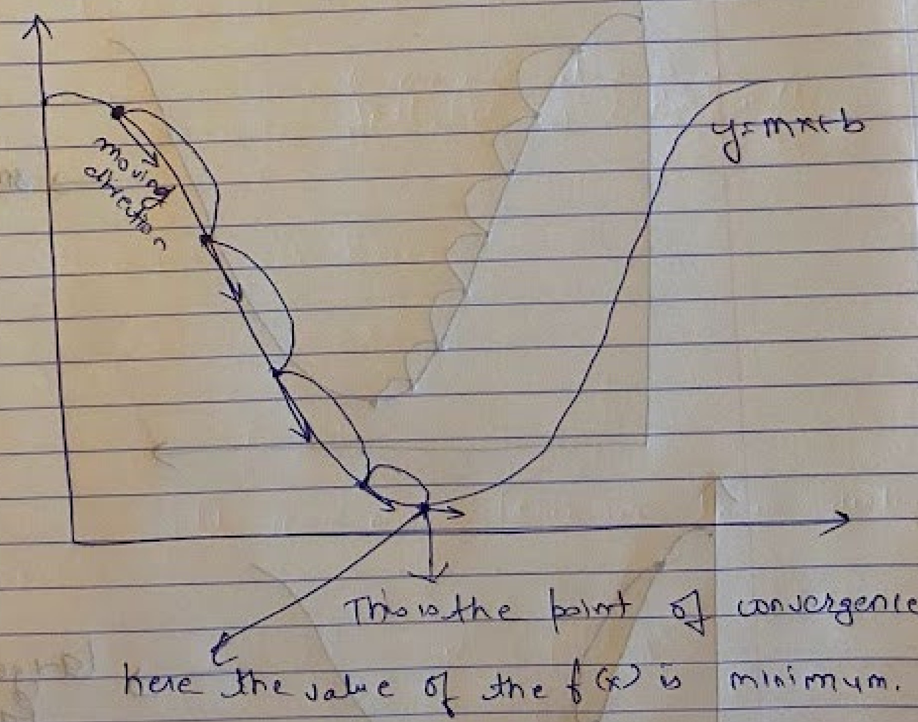
$$\Rightarrow y^{i+1} = y^i - \eta \frac{d(x^2+y^2)}{dy}$$

$$\Rightarrow y^{i+1} = y^i - \eta (2y)^i \quad \text{as } \frac{d(x^2+y^2)}{dy} = 2y$$

NAME - PARTH PARASHAR
PSU ID - 923928157

(ii) Gradient Descent is commonly used to train machine learning models and neural networks as well.
It is a geometrical term which can be understood by the below :-

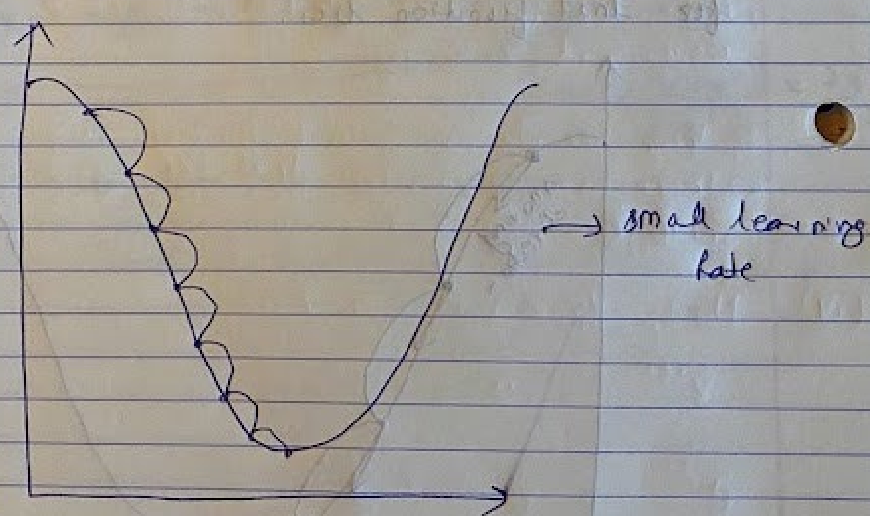
Consider the line with the equation $y = mx + b$
This is a function which represents a line.
for traversing the line, we take a point on the line and traverse along the tangent of the line to reach the point where the minima for that function lies.

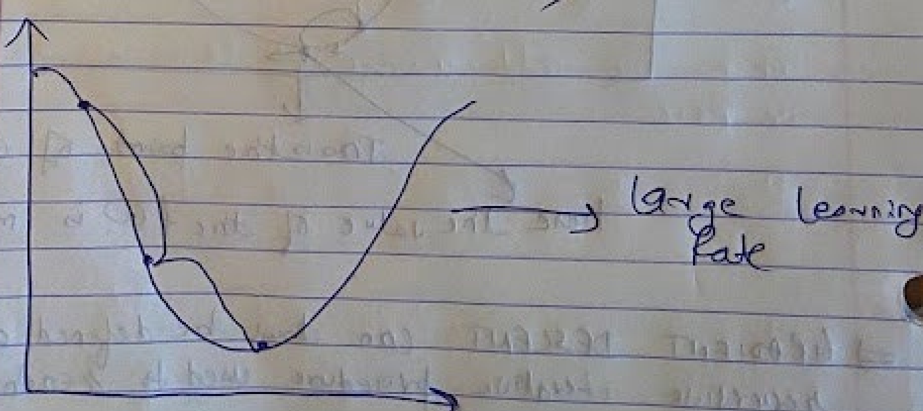


⇒ GRADIENT DESCENT can thus be defined as an recursive iterative procedure used to reach the minimum value of the function.

⇒ It is a vector that points in the Direction of maximum ascent ($\nabla_\theta R$) or maximum descent ($-\nabla_\theta R$)

$\boxed{\text{LEARNING RATE}}$ → It is the size of the steps for each recursive iteration to reach the minimum.





NAME - PARTH PARASHAR

PSU ID - 923928157

(iii) Now, as we know that this is a Quadratic function with the form $f(x,y)$ coordinate displacement,

$$f(x,y) = x^2+y^2$$

$$\frac{f'(x,y)}{dx} = 2x \quad , \quad \frac{f'(x,y)}{dy} = 2y$$

$$\frac{f''(x,y)}{dx} = 2 \quad , \quad \frac{f''(x,y)}{dy} = 2$$

Now, we can see that this $f(x,y)$, has a global minima at $x=0$, $y=0$

$\Rightarrow$ There exists a global minima at $x=0, y=0$

$\Rightarrow$ there is a guarantee that the algorithm will converge

(iv) Now, from (i), we have,

$$x^{i+1} = x^i - \eta\,(2x)^i$$

$$y^{i+1} = y^i - \eta\,(2y)^i$$

when we add momentum terms in these equations we have,,

$$x^{i+1} = x^i - \eta\,(2x)^i - \alpha \frac{d f(x,y)}{dx}$$

$$\Rightarrow \boxed{x^{i+1} = x^i - \eta\,(2x)^i - \alpha\,(2x)^{i-1}} \quad \left\{ \frac{d(x^2+y^2)}{dx} = 2x \right\}$$

NAME - PARTH PARASHAR

PSU ID - 923928157

Similarly, adding momentum term to the y, we have,

$$y^{i+1} = y^i - \eta (2y)^i - \alpha \frac{\partial f(x,y)}{\partial y}$$

$$\Rightarrow y^{i+1} = y^i - \eta (2y)^i - \alpha \frac{\partial (x^2+y^2)}{\partial y}$$

$$\boxed{\Rightarrow y^{i+1} = y^i - \eta (2y)^i - \alpha (2y)^{i-1}}$$

$$\left\{ \text{as } \frac{\partial (x^2+y^2)}{\partial y} = 2y \right\}$$

18. (2 pts.) Consider the confusion matrix given below for a hypothetical 3-class classifier.

Predicted Class →

True Class ↓

Class	A	B	C
A	50	20	30
B	20	30	30
C	30	50	40

What is the accuracy of this classifier?

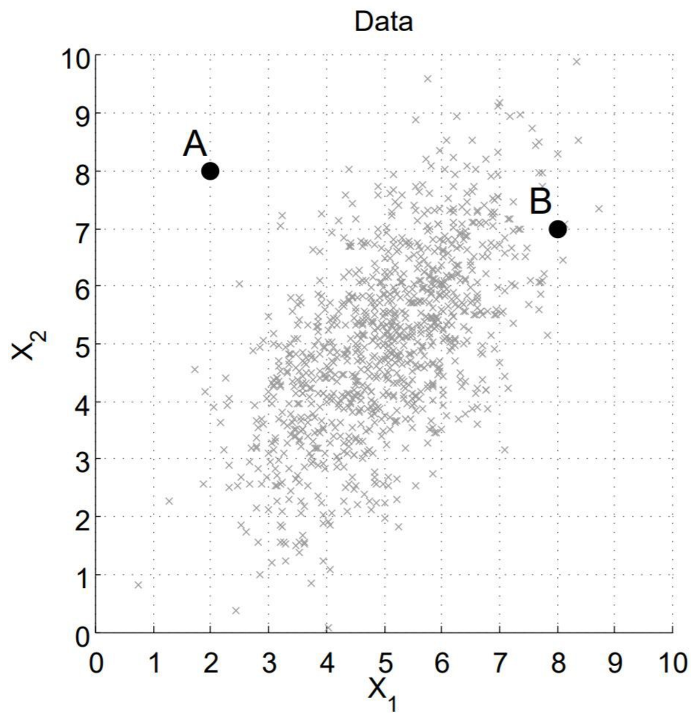
NAME - PARTH PARASHAR
PSU ID - 923928157

Ans 15) Accuracy $= \frac{(TP + TN)}{Total}$

$$= \frac{12\cancel{0}}{30\cancel{0}} \quad \frac{\cancel{4}}{\cancel{10}} \quad \frac{2}{5}$$

$$= 0.4$$

19. (3 pts.) *PCA.* The plot below shows a sample drawn from a two dimensional multivariate Normal (Gaussian) distribution. Define vectors $\mathbf{v}_1$ and $\mathbf{v}_2$ as the directions of the first and second principal components, after applying PCA to the dataset, where $\|\mathbf{v}_1\| = \|\mathbf{v}_2\| = 1$.



(i) Sketch and label $\mathbf{v}_1$ and $\mathbf{v}_2$ in the figure above. The arrows should originate from the mean of the distribution. You do not need to compute the actual PCA procedure, instead simply visually estimate the directions of the arrows.

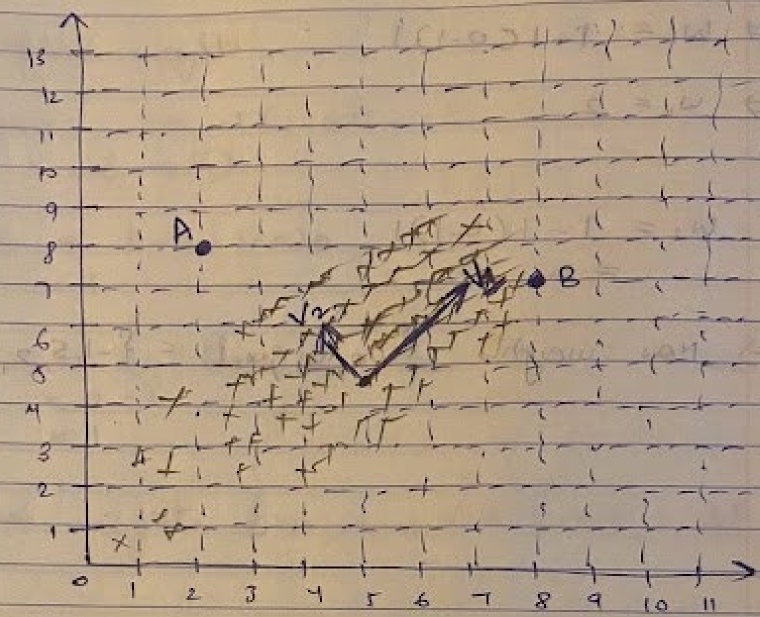
(ii) Which point (A or B) would have the higher reconstruction error after projecting onto the first principal component direction v1? Circle one:

Point A   Point B

NAME- PARTH PARASHAR
PSU ID - 923928157

Ans 19)

(i)



(ii) Point A $\longrightarrow$ higher reconstruction error after projecting onto the first principal component v_1.

20. (4 pts.) By hand, iterate the PLA (Perceptron Learning Algorithm) for the following training dataset.

Training Datum	$\boldsymbol{X_1}$	$\boldsymbol{X_2}$	**Class**
(i)	0	1	0
(ii)	2	0	0
(iii)	1	1	1

Use initial weights: $\boldsymbol{w_0 = -1.5, w_1 = 0, w_2 = 2}$ and learning rate η=1; **iterate the PLA for 2 epochs** (6 total steps); clearly indicate at each step the new weight values and the predicted class.

NAME- PARTH PARASHAR

PSU ID - 923928157

Ans 20)

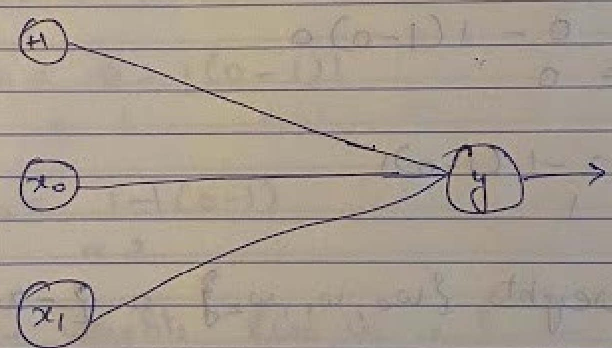
Training data	X_1	X_2	Class
(i)	0	1	0
(ii)	2	0	0
(iii)	1	1	1

Initial weights $w_0 = (-1.5)$

$w_1 = 0$

$w_2 = 2$

learning Rate $\eta = 1$



Activation function :-

$$z = \sum_{i}^{n} (w_i x_i) \; ; \; (z) = \begin{cases} 1 & z > 0 \\ 0 & z \leq 0 \end{cases}$$

weight updation will take place as follows:-

$w_i \leftarrow w_i - \Delta w_i \quad , \; w_i \leftarrow w_i - \eta (y^k - t^k) x_i^k$

NAME - PARTH PARASHAR
PSU ID - 923928157

for Epoch-1 :

(i) $(x_0, x_1) = (0,1)$; $t=0$

$y = z(-1.5 + 0 + 2)$

$y = z(0.5) \Rightarrow 1$

update weights:

$w_0 = (-1.5) - 1(1-0)1$
$= (-2.5)$

$w_1 = 0 - 1(1-0)0$
$= 0$

$w_2 = 2 - 1(1-0)1$
$= 1$

New weights $\{w_0, w_1, w_2\} = \{-2.5, 0, 1\}$

(ii) $(x_0, x_1) = (2,0)$; $t=0$

$y = z(-2.5 + 0 + 1)$

$y = z(-1.5)$
$y = 0$

$\Rightarrow y = t$

NAME - PARTH PARASHAR
PSU ID - 923928157

(iii) $(x_0, x_1) = (1,1)$; $t=1$

$y = Z(-2.5+0+1)$

$y = Z(-1.5) \Rightarrow y=0$

$y \neq t$

$\Rightarrow$ we need to update weights,

$w_0 = -2.5 - 1(0-1)$
$= -2.5+1$
$= (-1.5)$

$w_1 = 0 - 1(0-1)1$
$= 1$

$w_2 = 1 - 1(0-1)$
$= 2$

$\Rightarrow$ New weights $\{w_0, w_1, w_2\} = \{-1.5, 1, 2\}$

Epoch - 2 :→

(i) $(x_0, x_1) = (0,1)$; $t=0$

$y = Z(-1.5+0+2)$
$y = Z(0.5)$

$\Rightarrow y = 1 \qquad \Rightarrow y \neq t$

NAME – PARTH PARASHAR
PSU ID – 923928157

update weights

$w_0 = -1.5 - 1(1-0)1$
$= (-2.5)$

$w_1 = 1 - 1(1-0)0$
$= 1$

$w_2 = 2 - 1(1-0)1$
$= 2 - 1$
$= 1$

$\Rightarrow$ New weights $\{w_0, w_1, w_2\} = \{-2.5, 1, 1\}$

(ii) $(x_0, x_1) = (2, 0)$; $t = 0$

$y = Z(-2.5 + 2 + 0)$
$y = Z(-0.5)$

$\Rightarrow y = 0$ $\Rightarrow y = t$

(iii) $(x_0, x_1) = (1, 1)$; $t = 1$

$y = Z(-2.5 + 1 + 1) = Z(-0.5)$

$\Rightarrow y = 0$ $\Rightarrow y \neq t$

update weights,
$w_0 = -2.5 - 1(0-1)1$
$= (-1.5)$

NAME- PARTH PARASHAR
PSU ID - 923928157

$\Rightarrow w_1 = 1 - 1(0-1)1$

$\Rightarrow w_1 = 2$

$w_2 = 1 - 1(0-1)1$
$= 2$

$\Rightarrow$ new weights $\{w_0, w_1, w_2\} = \{-1.5, 2, 2\}$

21. (2 pts.) Suppose you have the following short DNA sequences in your training set:

s_1 = ACCGT s_2 = GTTGT s_3 = CGCCT

Suppose you are using these to train an SVM with a "match-count" kernel, where $k(s_i, s_j)$ returns the number of locations strings s_1 and s_2 at which the symbols match.

Give the Kernel matrix K for this training set and kernel function**.**

Name- Parth Parashar
PSU ID - 923928157

Ans 21) The given DNA sequences in the training set,

s_1 = ACCGT
s_2 = GTTGT
s_3 = CGCCT

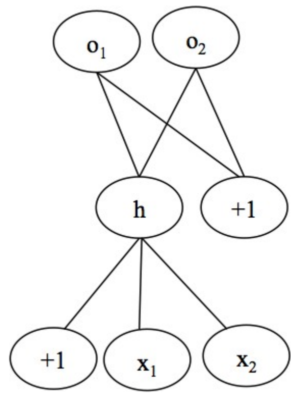
having (s_i, s_j) = Number of sites where strings match.

Therefore, the Kernel matrix (K) is given below:-

K	s_1	s_2	s_3
s_1	5	2	2
s_2	2	5	1
s_3	2	1	5

→ This matrix is positive semi-definite
⇒ defined K is a kernel function.

22. (4 pts.) Consider the multilayer neural network given below.



Suppose all the weights are initialized to 0.1. Assume the sigmoid activation function for hidden and output nodes:

$$o = \sigma(\boldsymbol{w} \cdot \boldsymbol{x}), where\ \sigma(z) = \frac{1}{1+e^{-z}}.$$

(a) Given input $\mathbf{x}$ = (1,2), what is the activation at output node o_1 ?

(b) Recall that the weight update rule for backpropagation is

$$\triangle\, w_{ji} = \eta \delta_j x_{ji} + \textit{momentum-term}$$

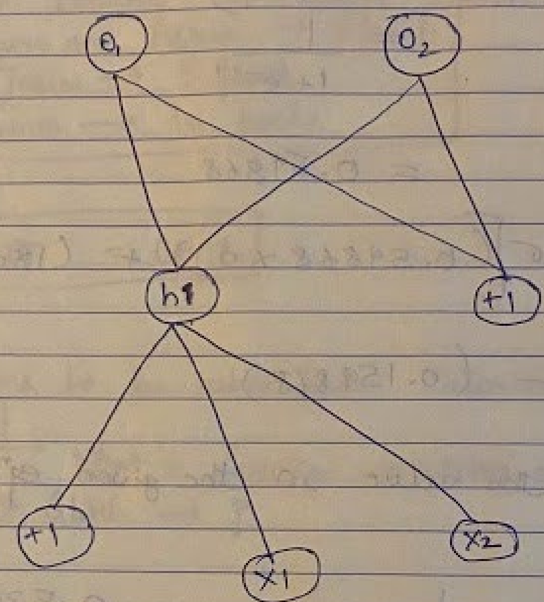
Let $\eta = 1$ and *momentum* = 0.

Suppose that after the weights are initialized to 0.1 and the input in part (a) is given, the error term for o_1 is calculated to be δ_{o_1} = 0.2.

What is the new value of $w_{o_1 h}$, the weight from the hidden unit to o_1 ?

Name- PARTH PARASHAR
PSU ID- 923928157

Ans 2.2)



From the question, it is given that all the weights are initialised to 0.1.
Also, the sigmoid function is given as:-

$$O = \sigma(w.x), \text{ where } \sigma(z) = \frac{1}{1+e^{-z}}$$

(a) for $x = (1,2)$, we have,

$$h = \sigma((1\times0.1) + (1\times0.1) + (2\times0.1))$$
$$h = \sigma(0.1+0.1+0.2)$$

$$h = \sigma(0.4)$$

Putting this value in the sigmoid, we have,

NAME- PARTH PARASHAR
PSU ID- 923928157

$h = \sigma(0.4)$

$= \frac{1}{1+e^{-0.4}}$

$= 0.59868$

$O_1 = \sigma\left[(0.59868 \times 0.1) + (1 \times 0.1)\right]$

$O_1 = \sigma(0.159868)$

Putting this value in the given eq^n, we have,

$O_1 = \frac{1}{1+e^{-0.159868}} = 0.53988$

$\Rightarrow O_1 = 0.53988$

(b) $\Delta w_{ji} = \eta \delta_j x_{ji}$ + momentum-term

$\eta = 1$
momentum $= 0 \Rightarrow \alpha = 0$
$\delta_{O_1} = 0.2$

Now, as we know that,

$\Delta w_{ji} = \eta \delta_j x_{ji} + \alpha \Delta w_{ji}^{t-1}$

NAME- PARTH PARASHAR

PSU ID - 923128157

$\Delta W_{0,h} = (1 \times 0.2 \times 0.59868) + 0$

$\Delta W_{0,n} = 0.119736$

$\Delta W_{0,h} = W_{0,h} + \Delta W_{0,h}$

$\Delta W_{0,h} = 0.1 + 0.119736$

$= 0.2196$

23. (3 pts.) **Definition**. A real-valued symmetric matrix A is positive semi-definite if: $\mathbf{x}^T A\mathbf{x} \geq 0$ for all real-valued vectors $\mathbf{x} \neq \mathbf{0}$.

(i) Let's consider a simple matrix as an example; let $A = \begin{bmatrix} 1 & -1 \\ -1 & 1 \end{bmatrix}$. Prove that A in this case is positive semi-definite. In other words, you need to show that the necessary condition holds in all cases.

(ii) Briefly, explain the significance of positive semi-definite matrices in relation to the *Mercer's Theorem* in ML.

Name - PARTH PARASHAR

PSU ID - 923928157

Ans 23) (i) $A \rightarrow$ positive semi-definite if $x^T A x \geq 0$ for $x \neq 0$

Now, the given matrix is:-

$$A = \begin{bmatrix} 1 & (-1) \\ (-1) & 1 \end{bmatrix}$$

let us suppose that $\hat{x} = (x_1, x_2)$, then,

for A to be a positive semi-definite, it should satisfy the condition mentioned below:-

$$\hat{x}^T A x \geq 0$$

$$\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} 1 & (-1) \\ (-1) & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \geq 0$$

Now, by solving this equation, we have,

$$\begin{bmatrix} (x_1 - x_2) & (x_2 - x_1) \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \geq 0$$

$\Rightarrow x_1 (x_1 - x_2) + x_2 (x_2 - x_1) \geq 0$

$\Rightarrow x_1^2 - \cancel{x_1 x_2} + x_2^2 - \cancel{x_2 x_1} \geq 0$

$\Rightarrow x_1^2 + x_2^2 - 2x_1 x_2 \geq 0$

$\Rightarrow (x_1 - x_2)^2 \geq 0 \quad \Rightarrow$ always true $\Rightarrow$ Hence proved

Name - PARTH PARASHAR

PSU ID - 923928157

Ans 23) (ii) Mercers theorum determines which functions can be used as a Kernel function.

Mercer's theorum also states that a symmetric and positive - definite matrix can be represented as a sum of a convergent sequence of product functions.

Also, from Mercer's theorum, a matrix is a Gram matrix if and only if it is positive and semi-definite. i.e. it is an inner product matrix in some space.

In other words, if a gram matrix M is a positive semi - definite ($x^T M x \geq 0$) then original function K defines a Kernel / a new inner product in a new feature space.