Name: Haoze He
NYUID: N12537

# Midterm
# Group A

## Problem 1 (20 points)

Recall D-dimensional regression problem when your model performs label

prediction for the i-th example $x_i$ in the training data set using linear function:

$$f(x_i; \theta_0, \theta_1, \theta_2, \dots, \theta_D) = \sum_{d=1}^{D} \theta_d x_i(d) + \theta_0.$$

In this case $x_i$ is D-dimensional. Let $y_i$ denote the true label of the i-th example and let N be the total number of training examples. Parameters of the model $(\theta_0, \theta_1, \theta_2, \dots, \theta_D)$ are obtained by minimizing the empirical risk provided below:

$$R(\theta) = \frac{1}{2N} ||y - X\theta||_2^2 + \theta^T H \theta + \theta^T \theta + a^T \theta,$$

where a is a vector and H is a matrix that satisfies the condition: $H = H^T$. Both a and H are given. Write what is y, X, and $\theta$ in the formula above. Compute the optimal setting of parameters by setting the gradient of the risk to 0. Explain all steps in your derivations.

## Problem 2 (15 points)

A kernel is an efficient way to write out an inner product between two feature vectors computed from a pair of input vectors as follows:

$$K(x, y) = \phi(x)^\top \phi(y).$$

Assume that both inputs are 2-dimensional and write out the explicit mapping $\phi$ that mimics the kernel value for a 3rd-order polynomial kernel as follows:

$$K(x, y) = (x^\top y + 1)^3.$$

## Problem 3 (15 points)

The exponential distribution has density given as

$$p_\lambda(x) = \lambda e^{-\lambda x}$$

Find the maximum likelihood estimator for $\lambda$. Calculate an estimate using this estimator when $x_1 = 1$, $x_2 = 2$, $x_3 = 4$, $x_4 = 2$.

## Problem 4 (15 points)

Consider 2d family of classifiers given by an origin-centered circles $f(x) = sign(ax^\top x + b)$. What is the VC dimension of this family? Prove it.

## Problem 5 (15 points)

Using the principle of Lagrange multipliers, find the maximum and minimum values of $f(x, y) = x^2 - y^2$ subject to the constraint, $x^2 + y^2 = 1$.

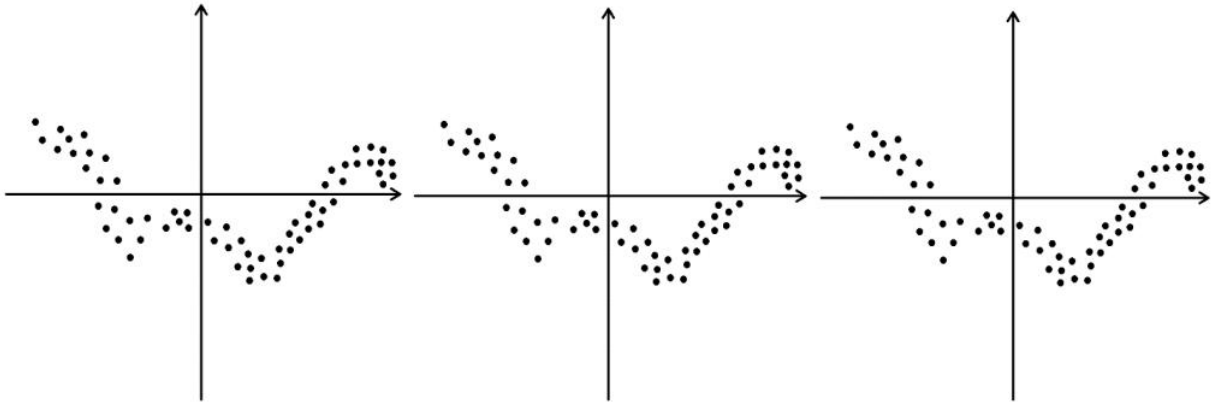
## Problem 6 (10 points)

What is the maximum likelihood estimator of the mean $\mu$ and covariance matrix $\Sigma$ of the following 2-dimensional data set $\mathcal{X}$? Justify your answer.

$$\mathcal{X} = \left\{ \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \begin{bmatrix} 3 \\ 4 \end{bmatrix}, \begin{bmatrix} 5 \\ 6 \end{bmatrix}, \begin{bmatrix} 7 \\ 8 \end{bmatrix}, \begin{bmatrix} 9 \\ 10 \end{bmatrix} \right\}$$

## Problem 7 (10 points)

Explain the difference between overfitting and underfitting. For a picture given below, show and example of underfitting on the left plot, proper fit in the middle plot, and overfitting on the right plot.



Q1.

Step 1: explain the parameters.

$\vec{y} \in R^{N*1}$: matrix contains all the true labels

$\vec{X} \in R^{N*D}$: matrix contains all the training data.

$\vec{\theta} \in R^{D*1}$: parameters use to fit the training data and give the result.

Step2.

$R(\theta) = \frac{1}{2N} \|\vec{y} - \vec{X}\vec{\theta}\|_2^2 + \vec{\theta}^T H \vec{\theta} + \vec{\theta}^T \vec{\theta} + \vec{a}^T \vec{\theta}$ // Original formula

$$= \frac{1}{2N}\left[(\vec{y}-\vec{X}\vec{\theta})^T\cdot(\vec{y}-\vec{X}\vec{\theta})\right]+\vec{\theta}H\vec{\theta}+\vec{\theta}\vec{\theta}+\vec{a}^T\vec{\theta} \quad \text{// Append 1st term}$$

$$= \frac{1}{2N}\left[\vec{y}^T\vec{y}-2\vec{y}^T\vec{X}\vec{\theta}+\vec{\theta}^T\vec{X}^T\vec{X}\vec{\theta}\right]+\vec{\theta}^TH\vec{\theta}+\vec{\theta}^T\vec{\theta}+\vec{a}^T\vec{\theta} \quad \text{// Append 1st term}$$

$$\frac{\partial R(\theta)}{\partial\theta} = \frac{1}{2N}\left[-2\vec{X}^T\vec{y}+2\vec{X}^T\vec{X}\vec{\theta}\right]+(H+H^T)\vec{\theta}+2\vec{\theta}+\vec{a}^T=0 \quad \text{// Start Derivation}$$

$$-2\vec{X}^T\vec{y}+2\vec{X}^T\vec{X}\vec{\theta}+2N(H^T\vec{\theta}+2\theta+a) \quad =0 \quad \text{// } H^T=H \text{ according to Question}$$

$$-\vec{X}^T\cdot\vec{y}+\vec{X}^T\vec{X}\vec{\theta}+2NH\vec{\theta}+2N\vec{\theta}+Na=0 \quad \text{// simplify}$$

$$\vec{X}^T\vec{X}\theta+2NH\vec{\theta}+2N\vec{\theta}=\vec{X}^T\vec{y}-Na \quad \text{// simplify}$$

$$(\vec{X}^T\vec{X}+2NH+2NI)\vec{\theta}=\vec{X}^T\vec{y}-Na$$

$$\vec{\theta}=(\vec{X}^T\vec{X}+2NH+2NI)^{-1}(\vec{X}^T\vec{y}-Na) \quad \text{// solution}$$

Q2.

$$K(x,y)=(x^Ty+1)^3=(x^Ty+1)(x^Ty)^2+2x^Ty+1)$$

$$=(x^Ty)^3+1+3(x^Ty)^2+3x^Ty$$

$$=(x_1y_1+x_2y_2)^3+1+3(x_1y_1+x_2y_2)^2+3(x_1y_1+x_2y_2)$$

$$=x_1^3y_1^3+x_2^3y_2^3+3x_1^2y_1^2x_2y_2+3x_1y_1x_2^2y_2^2+1+$$

$$3(x_1^2y_1^2+x_2^2y_2^2+2x_1y_1x_2y_2)+3(x_1y_1+x_2y_2)$$

$$=x_1^3y_1^3+x_2^3y_2^3+3x_1y_1x_2^2y_2^2+3x_1^2y_1^2x_2y_2+3x_1^2y_1^2+3x_2^2y_2^2+$$

$$6x_1y_1x_2y_2+3x_1y_1+3x_2y_2+1$$

Hence:

$$\Phi(x)=\left[x_1^3 \quad x_2^3 \quad \sqrt{3}x_1^2x_2 \quad \sqrt{3}x_1x_2^2 \quad \sqrt{3}x_1^2 \quad \sqrt{3}x_2^2 \quad \sqrt{6}x_1x_2 \quad \sqrt{3}x_1 \quad \sqrt{3}x_2 \quad 1\right]$$

s.t.

$$\Phi(x)^T\Phi(y)=(x^Ty+1)^3$$

Q3.

$$L(\lambda)=\prod_{i=1}^{N}\lambda\cdot e^{-\lambda x_i}$$

Instead of directly maximizing the likelihood. We instead maximize the log-likelihood. the log-likelihood.

$$\log L(\lambda)=\log\prod_{i=1}^{N}\lambda\cdot e^{-\lambda x_i}=\sum_{i=1}^{n}\log\lambda+(-\lambda x_i)=N\log\lambda+\sum_{i=1}^{n}(-\lambda x_i)$$

$$\frac{\partial\log L(\lambda)}{\partial\lambda}=\frac{N}{\lambda}+\sum_{i=1}^{n}(-x_i)$$

Set the derivative equal to 0:

$$\lambda = \frac{N}{\sum_{i=1}^{n} x_i}$$

Using the given data.

$$\lambda = \frac{N}{\sum_{i=1}^{n} x_i} = \frac{4}{1+2+4+2} = \frac{4}{9}$$

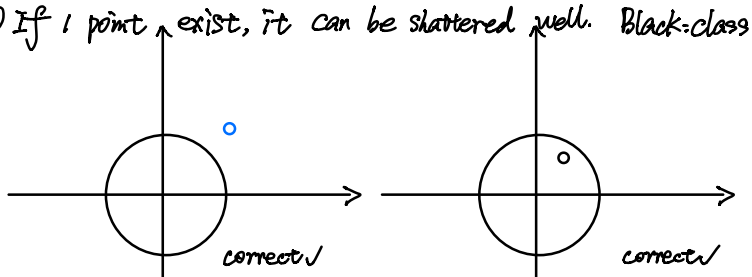
Then, verify it's maximum by taking second derivative

$$\frac{\partial(\log L(\lambda))}{\partial^2(\lambda)} = \frac{-N}{\lambda^2} < 0$$

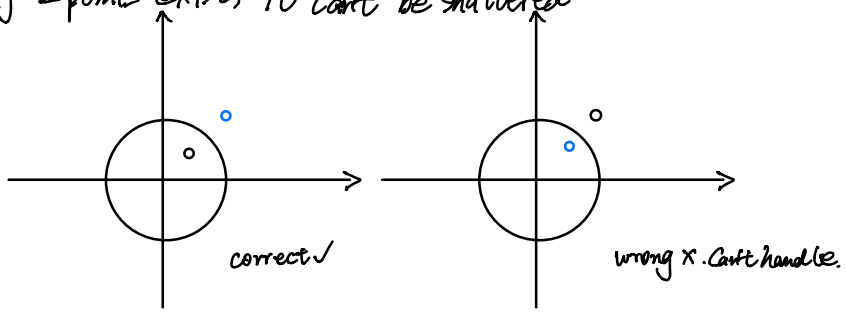
Hence, proved.

Q4. VC Dimension = 1

① If 1 point exist, it can be shattered well. Black: class 1 Blue: class 2




② If 2 points exist, it can't be shattered




Hence, VC Dimension = 1

Q5

$$f(x,y) = x^2 - y^2 \quad ①$$
$$g(x,y) = x^2 + y^2 - 1 = 0 \quad ②$$

$$F(x) = x^2 - y^2 + \lambda(x^2 + y^2 - 1)$$

$$\frac{\partial F(x)}{\partial x} = 2x + 2\lambda x = 0 \Rightarrow 2xy + 2\lambda xy = 0 \quad ①$$

$$\frac{\partial F(x)}{\partial y} = -2y + 2y\lambda = 0 \Rightarrow -2xy + 2\lambda xy = 0 \quad ②$$

$$\Rightarrow 4xy = 0 \Rightarrow x = 0 \text{ or } y = 0$$

$$\frac{\partial F(x)}{\partial \lambda} = x^2+y^2-1=0 \quad (3)$$

Assume $x=0 \Rightarrow$ According to (3), $y=\pm 1$

$y=0 \Rightarrow$ According to (3) $x=\pm 1$

Hence, $Max = x^2-y^2 = 1^2-0=1$

$Min = x^2-y^2 = 0^2-1^2=-1$

Q6. To get $u$ and $\Sigma$, we try to maximum likelihood:

$$p(\vec{X}|\theta) = \prod_{i=1}^{N} p(\vec{x_i}|\vec{u_i},\Sigma_i) = \prod_{i=1}^{N} p(\vec{x_i}|\vec{u},\Sigma) \quad \text{i.i.d.}$$

$$\log L(\theta) = \sum_{i=1}^{N} \log p(\vec{x_i}|\vec{u},\Sigma) = \sum_{i=1}^{N} \log \frac{1}{(2\pi)^{D/2}\sqrt{|\Sigma|}} \exp\left[-\frac{1}{2}(\vec{x_i}-\vec{u})^T\cdot\Sigma^{-1}\cdot(\vec{x_i}-\vec{u})\right]$$

① Max over $\vec{u}$:

$$\frac{\partial \log L(\theta)}{\partial \vec{u}} \rightarrow \sum_{i=1}^{N}(x_i-\vec{u})^T\cdot\Sigma^{-1}=\vec{0} \Rightarrow \vec{u}=\frac{1}{N}\cdot\sum_{i=1}^{N}\vec{x_i} = \begin{bmatrix}5\\6\end{bmatrix}$$

② Max over $\Sigma$:

$$\log L(\theta) = \sum_{i=1}^{N} -\frac{D}{2}\log 2\pi - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(\vec{x_i}-\vec{u})^T\Sigma^{-1}\cdot(\vec{x_i}-\vec{u})$$

$$= -\frac{ND}{2}\log 2\pi + \frac{N}{2}\log|\Sigma^{-1}| - \frac{1}{2}\sum_{i=1}^{N} tr\left|(\vec{x_i}-\vec{u})^T(\vec{x_i}-u)\Sigma^{-1}\right|$$

We set $A=\Sigma^{-1}$

$$= -\frac{ND}{2}\log 2\pi + \frac{N}{2}\log|A| - \frac{1}{2}\sum_{i=1}^{N} tr\left[(\vec{x_i}-\vec{u})^T(\vec{x_i}-\vec{u})A\right]$$

$$\frac{\partial \log L(\theta)}{\partial A} = \frac{N}{2}\Sigma - \frac{1}{2}\sum_{i=1}^{N}(\vec{x_i}-\vec{u})^T\cdot(\vec{x_i}-\vec{u}) = 0$$

$$\Sigma = \frac{1}{N}\sum_{i=1}^{N}(x_i-u)^T\cdot(\vec{x_i}-\vec{u})$$

$$= \frac{1}{5}\cdot[32+8+0+8+32]$$

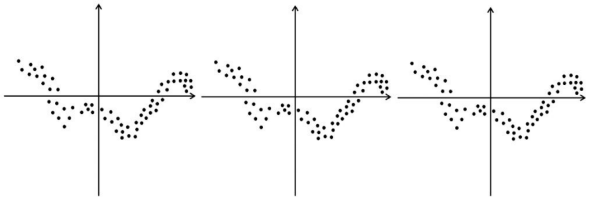
$$= 16$$

Hence, $u=\begin{bmatrix}5\\6\end{bmatrix}$ $\Sigma=16$ maximum the likelihood.

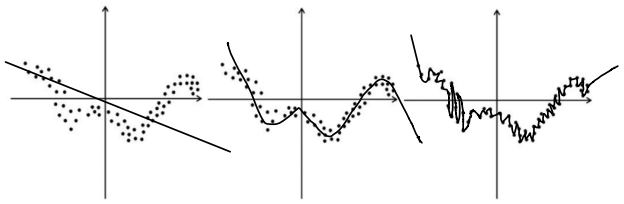
Q7.

Problem 7 (10 points)

Explain the difference between overfitting and underfitting. For a picture given below, show and example of underfitting on the left plot, proper fit in the middle plot, and overfitting on the right plot.



Difference - ① Underfitting means a machine learning model can not properly present the trend and feature of data.
Overfitting means a machine learning model describe the noise and random error of training set instead of the proper trend of the data.
② Overfitting occurs when the machine learning model is over-complex.
underfitting occurs when the model is not complex enough or the model is not converged.



underfitting proper-fit over-fitting