practice midterm

Data Mining, Spring Quarter

Duration: 1 hours 45 minutes

Name: Pooran Singh Negi DU ID:

1. **This is closed book/notes exams**
2. **Please write your name and DU ID before starting the exam.**
3. **Show all the step of your answer and justify you answer/steps**

Problem 1.(1 points each.)

1a. Give an example of machine learning algorithm which is used for regression. linear regression

1b. is LDA supervised dimentionality reduction technique yes/no? y

1d. is median a robust estimate of central tendency(yes/no)? yes

1e. If there are k class in a dataset. Using LDA, what is the maximum number of dimensions you can project data into. k−1

Problem 1.(2 points). Given that all the N scalar observations x_i are in a vector $\mathbf{x}_{N\times 1} = \begin{bmatrix} x_1 \\ \vdots \\ x_N \end{bmatrix}$. Let $\mathbf{1}_{N\times 1}$ be a vector of ones. Show that magnitude of projection of $\mathbf{x}_{N\times 1}$ onto $\mathbf{1}_{N\times 1}$ is the mean of N observations x_i.

$$\frac{x^T \mathbb{1}}{\mathbb{1}^T \mathbb{1}} = \frac{x_1 + x_2 + \cdots x_N}{1 + 1 + \cdots 1} = \frac{\sum_{i=1}^{N} x_i}{N}$$

Problem 2.(6 = (1+2)+2+1 points.)

2a. What does mercer theorm gaurantees given that kernel is positive definite? Prove that if $k(x_i, x_j)$ is mercer kernel then $ck(x_i, x_j)$ is also mercer kernel for $c > 0$.

2b Prove that $k(\boldsymbol{x_1}, \boldsymbol{x_2}) = \tanh(\gamma \boldsymbol{x_1}^T \boldsymbol{x_2})$ is not a mercer kernel, where $\tanh(\boldsymbol{z}) = \frac{\exp(2\boldsymbol{z})-1}{\exp(2\boldsymbol{z})+1}$ choose (N=1) one data point i.e. $x_1 = x_2 =$ zero vector, then 1×1 Gram matrix $[K(x_1, x_2)] = [0]$ not positive definite or positive (scalar (1×1 matrix))

Problem 3.(2+2 points.)

3a. Given that decision function is $f(\boldsymbol{x}) = \boldsymbol{w}^T x + b$. If after kenelizing one can show that $\boldsymbol{w} = \sum_{i=1}^{m} {}_i \Phi(\boldsymbol{x_i})$ with associated kernel $k(x_i, x_j) = \Phi(\boldsymbol{x_i})^T \Phi(\boldsymbol{x_j})$. Write down decision rule in kernel format. Plug w into f

3b. What is kernel trick? Let's say a classifier depends on computing distance $\|x_1 - x_2\|_2^2$. If we want to kernelize this classifier, then expand this distance and write down the kernelized version for distance. $\|x_1-x_2\|_2^2 = (x_1-x_2)^T(x_1-x_2) = (x_1^T - x_2^T)(x_1-x_2) =$ $x_1^Tx_1 - x_1^Tx_2 - x_2^Tx_1 + x_2^Tx_2 = x_1^Tx_1 - 2x_1^Tx_2 + x_2^Tx_2$ Kernelizing $K(x_1,x_1) - 2K(x_1,x_2) + K(x_2,x_2)$

Problem 4 (4= 1+1+2) Which norm does feature selection in linear regresssion? Write this norm definition and explain the geometrically why using this norm does feature selection. see lecture notes or book links