Birla Institute of Technology and Science, Pilani

Work Integrated Learning Programmes Division

Cluster Programme - M.Tech. in Artificial Intelligence and Machine Learning

I Semester 2022-23

Course Number	AIMLC ZC416
Course Name	Mathematical Foundations for Machine Learning
Nature of Exam	Open Book
Weightage for grading	40%
Duration	150 minutes
Date of Exam	02/04/2023 (14:00 - 16:00)

# Pages	3
# Questions	8

Instructions

1. All questions are compulsory.
2. All parts of a question should be answered consecutively. Each answer should start from a fresh page.

(1) Let $(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \ldots (\boldsymbol{x}_N, y_N)$ be N points on which we find a SVM classifier using the dual SVM formulation which is given below:

$$\text{maximize} \sum_{i=1}^{i=N} \alpha_i - \sum_{i=1}^{i=N} \sum_{j=1}^{j=N} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i . \boldsymbol{x}_j \text{ subject to}$$

$$\sum_{i=1}^{N} \alpha_i y_i = 0$$

$$\alpha_i \geq 0 \forall i$$

Let O_A be the value of the objective function at the optimal solution returned by the dual SVM formulation for this problem. Now we add a new point $(\boldsymbol{x}_{N+1}, y_{N+1})$ and find a SVM classifier by solving the dual formulation again. Let O_B be the value of the objective function at the optimal solution for this problem. Considering the following three relationships (a) $O_A > O_B$, (b) $O_A = O_B$, (c) $O_A < O_B$ determine which of these relationships are possible and give a mathematical argument for your answer in each case.

[5 Marks]

Solution:

Let us denote the solution vector corresponding to O_A to be $< \alpha_1, \alpha_2, \ldots, \alpha_N >$. We note that $< \alpha_1, \alpha_2, \ldots \alpha_N, 0 >$ is a feasible solution to the problem B on $N + 1$ points since it satisfies all the constraints associated with this problem. The value of the objective function at this feasible solution for B is equal to O_A since all the extra terms in the objective function involve α_{N+1} which is equal to 0. Now the optimal solution for B, O_B is greater than or equal to this feasible solution by definition. Thus we have $O_B \geq O_A$. It is not possible to have $O_B < O_A$.

Thus (a) is not possible, (b) and (c) are possible.

Marking Scheme: 4 Marks for explanation, 1 for final answer.

(2) Let $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots \boldsymbol{x}_N$ be N points on which we perform Principle Components Analysis leading to the discovery of the principle component directions $\boldsymbol{b}_1, \boldsymbol{b}_2 \ldots \boldsymbol{b}_D$. The given data is now transformed to the points $\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots \boldsymbol{y}_N$ where $\boldsymbol{y}_i = \boldsymbol{Q}\boldsymbol{x}_i + \boldsymbol{\mu}$ where $\boldsymbol{Q}^T\boldsymbol{Q} = \boldsymbol{I}$ and $\boldsymbol{\mu}$ is a constant vector. Determine the principle components for the transformed set of points in terms of the old principle components. How much variance is accounted for by the first principal component for the transformed set of points in terms of the variance accounted for by the first principle component for the original set of points? Justify your answer with mathematical arguments.

[5 Marks]

Solution:

Before performing PCA on the points $\boldsymbol{y}_i$ we perform mean-centering so that the points become $\boldsymbol{y}_i - \boldsymbol{\mu}$, and the data matrix becomes $\boldsymbol{Y} = [\boldsymbol{y}_1 - \boldsymbol{u}, \boldsymbol{y}_2 - \boldsymbol{\mu}, \ldots \boldsymbol{y}_N - \boldsymbol{\mu}]$. The data-covariance matrix can then be computed as $E(\boldsymbol{Y}\boldsymbol{Y}^T) = \boldsymbol{Q}\boldsymbol{X}\boldsymbol{X}^T\boldsymbol{Q}^T$.

The PCA components $\boldsymbol{b}_1, \boldsymbol{b}_2, \ldots \boldsymbol{b}_D$ can be viewed as the normalized eigenvectors of the data-covariance matrix $E(\boldsymbol{X}\boldsymbol{X}^T)$ on the original set of points. Thus we have $E(\boldsymbol{X}\boldsymbol{X}^T) = \boldsymbol{B}\boldsymbol{\Lambda}\boldsymbol{B}^T$, and $E(\boldsymbol{Y}\boldsymbol{Y}^T) = \boldsymbol{Q}\boldsymbol{B}\boldsymbol{\Lambda}\boldsymbol{Q}\boldsymbol{B}^T$. The new set of eigenvectors becomes $\boldsymbol{c}_i = \boldsymbol{Q}\boldsymbol{b}_i$ and they have unit length since $\boldsymbol{c}_i^T\boldsymbol{c}_i = \boldsymbol{b}_i^T\boldsymbol{Q}^T\boldsymbol{Q}\boldsymbol{b}_i = \boldsymbol{b}_i^T\boldsymbol{b}_i = 1$. Thus the eigenvalues for the transformed problem remain the same as the original problem, and the variance accounted for by the first eigenvalue is the same as in the PCA of the original set of points.

Marking Scheme: Computation of the new covariance matrix $\rightarrow$ 2 Marks, Subsequent explanation $\rightarrow$ 3 Marks.

(3) Consider three linearly independent vectors in $\mathbb{R}^n$ named a_1, a_2 and a_3. Now construct three vectors $b_1 = a_2 - a_3$, $b_2 = a_1 - a_3$ and $b_3 = a_1 - a_2$. Now consider the set $\mathcal{Q} = \{b_1, b_2, b_3\}$. Prove or disprove that the set $\mathcal{Q}$ is linearly independent.

[5 Marks]

Solution

Consider $\beta_1 b_1 + \beta_2 b_2 + \beta_3 b_3 = 0$. This is same as $\beta_1(a_2 - a_3) + \beta_2(a_1 - a_3) + \beta_3(a_1 - a_2) = 0$. It can be shown that $\beta_1 = 1, \beta_2 = -1, \beta_3 = 1$ satisfies the equation. Hence the given set is linearly dependent.

Marking Scheme: 2 Marks $\rightarrow$ setting up linear independence equations, 3 Marks $\rightarrow$ final solution

(4) Consider two sets named $\mathcal{H}_1$ and $\mathcal{H}_2$. It is known that these two sets are convex sets.

(a) Prove or disprove that $\mathcal{H}_1 \cap \mathcal{H}_2$ is a convex set. Here $\cap$ represents the set intersection operation.

(b) Prove or disprove that $\mathcal{H}_1 \cup \mathcal{H}_2$ is a convex set. Here $\cup$ represents the set union operation.

[5 Marks]

Solution:

(a) Let $\mathcal{H} = \mathcal{H}_1 \cap \mathcal{H}_2$. Let $x, y \in \mathcal{H}$. This means $x, y \in \mathcal{H}_1$ and $x, y \in \mathcal{H}_2$ This means $\lambda x + (1-\lambda)y \in \mathcal{H}_1$ and $\lambda x + (1-\lambda)y \in \mathcal{H}_2$. Hence $\lambda x + (1-\lambda)y \in \mathcal{H}$. Hence $\mathcal{H}_1 \cap \mathcal{H}_2$ is a convex set

(b) It need not be convex always which is demonstrated using the following counter example. Let $\mathcal{H}_1 = [1\ \ 2]$ on x-axis and $\mathcal{H}_2 = [-2\ \ -1]$ on x axis. If you draw a line from $x_1 = -1.5$ which lies in $\mathcal{H}_1$ to $x_2 = 1.5$ which lies in $\mathcal{H}_2$. The point $x_3 = 0$ lies on this line segment , but its not part of $\mathcal{H}_1 \cup \mathcal{H}_2$.

Marking Scheme: 2.5 Marks $\rightarrow$ part (a), 2.5 Marks $\rightarrow$ part (b)

(5) A linear Algebra student arrived at an $n \times n$ real matrix given below

$$A = \begin{pmatrix} m_{11} & m_{12} & \cdots & m_{1n} \\ m_{21} & m_{22} & \cdots & m_{2n} \\ \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ m_{n1} & m_{n2} & \cdots & m_{nn} \end{pmatrix}$$

Help the student to find singular value decomposition of full rank matrix A if the columns of A are orthogonal.

[5 Marks]

Solution:

Since A has orthogonal columns

$$\begin{aligned} \sum_{i=1}^{n} m_{ij} m_{ik} &= 0 \text{ for } j \neq k \\ \text{Let } \sum_{i=1}^{n} m_{ij} m_{ik} &= \sigma_j^2 \text{ for } j = k \end{aligned}$$

$$A^T A = \begin{pmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ 0 & \sigma_2^2 & \cdots & 0 \\ \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ 0 & 0 & \cdots & \sigma_n^2 \end{pmatrix}$$

Clearly, eigenvalues of A^TA are σ_j^2 for $j = 1, \cdots n$
Solving $[A^TA - \sigma_i^2 I]\mathbf{x} = \mathbf{0}$, we get

$$v_i = \begin{pmatrix} 0 \\ \vdots \\ 1 \text{ (ith row)} \\ \vdots \\ 0 \end{pmatrix}$$

Now $Av_i = \sigma_i u_i$ for all $i = 1 \cdots n$. Thus

$$u_i = \frac{1}{\sigma_i} \begin{pmatrix} m_{1i} \\ m_{2i} \\ \vdots \\ m_{ni} \end{pmatrix}$$

Thus if $A = U\Sigma V$, then $U = [u_1, \cdots, u_n]$, $V = [v_1, \cdots, v_n] = \mathbf{I}$ and

$$\Sigma = \begin{pmatrix} \sigma_1 & 0 & \cdots & 0 \\ 0 & \sigma_2 & \cdots & 0 \\ \vdots & \vdots & \vdots\vdots\vdots & \vdots \\ 0 & 0 & \cdots & \sigma_n \end{pmatrix}$$

Marking Scheme 2 Marks $\rightarrow$ getting eignevalues of A^TA, 3 Marks $\rightarrow$ remaining argument

(6) A data analyst modeled the objective function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ as product of squares of n feature x_i s and $n \geq 2$. He has to maximize the objective function such that sum of squares of n features is less than or equal to c^2 where $c \in \mathbb{R}$. Write the mathematical formulation of the problem and solve it. Using the above result prove the inequality

$$(a_1 a_2 \cdots a_n)^{1/n} \leq \frac{a_1 + \cdots + a_n}{n} \text{ for any } a_i > 0, i = 1, \cdots, n$$

[5 Marks]

Solution:

The mathematical formulation is given by

$$\begin{aligned} & \max_{x_i \in \mathbb{R}, i=1\cdots,n} x_1^2 .....x_n^2 \\ \text{subject to constraints} \quad & x_1^2 + ... + x_n^2 \leq c^2 \end{aligned}$$

The Lagrangian is given by

$$L(\mathbf{x}, \lambda) = x_1^2 .....x_n^2 + \lambda(x_1^2 + ... + x_n^2 - c^2)$$

Equating gradient equal to 0, we get

$$\begin{aligned}
\frac{\partial L}{\partial x_i} &= \frac{2x_1^2....x_n^2}{x_i} + 2\lambda x_i = 0 \\
\frac{\partial L}{\partial \lambda} &= x_1^2 + ... + x_n^2 - c^2 = 0 \\
\Rightarrow x_1^2 + \cdots + x_n^2 &= c^2 \\
\Rightarrow x_1^2 \cdots x_n^2 &= -\lambda x_i^2 \ \forall i \\
\Rightarrow x_1^2 = \cdots = x_n^2 & \quad \text{for } \lambda \neq 0 \\
\Rightarrow x_1^2 = \cdots = x_n^2 &= \frac{c^2}{n} \\
\Rightarrow x_i &= \pm\frac{c}{\sqrt{n}} \ \forall i = 1, \cdots, n \\
\Rightarrow \max f = \frac{c^{2n}}{n^n}
\end{aligned}$$

Since $a_i > 0, i = 1, \cdots, n$, let $a_i = x_i^2$ for some x_i and let $x_1^2 + ... + x_n^2 = c^2$ for some c.
Then from part(1), we know

$$\begin{aligned}
(a_1 a_2 \cdots a_n)^{1/n} &= (x_1^2.....x_n^2)^{1/n} \\
&\leq \frac{c^2}{n} \\
&= \frac{a_1 + \cdots + a_n}{n} \quad \text{for any } a_i > 0, i = 1, \cdots, n
\end{aligned}$$

Marking Scheme: 3 Marks $\rightarrow$ setting up and solving the Lagrangian, 2 Marks $\rightarrow$ proving final inequality.

(7) You are given the quadratic polynomial $f(x, y, z) = 2x^2 - 2xy - 4xz + y^2 + 2yz + 3z^2 - 2x + 2z$:
   (a) Write $f(x, y, z)$ in the form $f(x, y, z) = x^T Ax - b^T x$ where $x = (x, y, z)$, A is a real symmetric matrix, and b is constant vector.
   (b) Find the point (x, y, z) where f(x, y, z) is at an extremum.
   (c) Is this point a minimum, maximum, or a saddle point of some kind?

[5 Marks]

Solution:

[2 MARKS]:

$$A = \begin{bmatrix} 2 & -1 & -2 \\ -1 & 1 & 1 \\ -2 & 1 & 3 \end{bmatrix}$$

and $b = [2, 0, -2]^T$ Find the point (x, y, z) where $f(x, y, z)$ is at an extremum.
Solution[1 Mark]:
$\frac{\partial f}{\partial x} = 4x - 2y - 4z - 2 = 0$
$\frac{\partial f}{\partial y} = -2x + 2y + 2z = 0$
$\frac{\partial f}{\partial z} = -4x + 2y + 6z + 2 = 0$

$(x, y, z) = (1, 1, 0)$ is extremum point. Is this point a minimum, maximum, or a saddle point of some kind? Solution[1 Mark]: Since Hessian $=2A$ which is positive definite and symmetric matrix., so $(1, 1, 0)$ is a point of minimum.

(8) Solve the System of equations by Gaussian elimination method
$2x_1 + 5x_2 + 2x_3 - 3x_4 = 3$
$3x_1 + 6x_2 + 5x_3 + 2x_4 = 2$
$4x_1 + 5x_2 + 14x_3 + 14x_4 = 11$
$5x_1 + 10x_2 + 8x_3 + 4x_4 = 4$

[5 Marks]

Solution:
Consider the augmented matrix [A/B]

$$[\mathbf{A}/\mathbf{B}] = \begin{bmatrix} 2 & 5 & 2 & -3 & | & 3 \\ 3 & 6 & 5 & 2 & | & 2 \\ 4 & 5 & 14 & 14 & | & 11 \\ 5 & 10 & 8 & 4 & | & 4 \end{bmatrix}$$

After applying row elementary transformations we get

$$[\mathbf{A}/\mathbf{B}] \approx \begin{bmatrix} 1 & 5/2 & 1 & -3/2 & | & 3/2 \\ 0 & 1 & -4/3 & -13/3 & | & 5/3 \\ 0 & 0 & 1 & -1/2 & | & 4 \\ 0 & 0 & 0 & 1 & | & 4 \end{bmatrix}$$

by back substitution, we get $x_4 = 4, x_3 = 6, x_2 = 27$ and $x_1 = -66$

Marking Scheme: 3 Marks $\rightarrow$ getting the final matrix, 2 Marks $\rightarrow$ solving for the variables by backsubstitution.