The Chinese University of Hong Kong, Shenzhen
School of Data Science

DDA 3020 · Homework 1Written Part Solution
Due: 23:59, March 9th, 2024

Instructions:

- This assignment accounts for 14/100 of the final score.
- You must independently complete each assignment.
- Late submission will get discounted score: 20 percent discount on $(0, 24]$ hours late; 50 percent discount on $(24, 120]$ hours late; no score on late submission of more than 120 hours.

---

1 Written Problems (50 pts.)

Problem 1 (10pts) Linear Algebra.

1. A rotation in 3D by angle α about the z axis is given by the following matrix:

$$\mathbf{R}(\alpha) = \begin{pmatrix} \cos(\alpha) & -\sin(\alpha) & 0 \\ \sin(\alpha) & \cos(\alpha) & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

   Prove that $\mathbf{R}$ is an orthogonal matrix, i.e., $\mathbf{R}^T\mathbf{R} = \mathbf{I}$, for any α.

2. Prove that the eigenvalue of an orthogonal matrix must be 1 or -1.

Solution:

1. Let $c = \cos(\alpha)$ and $s = \sin(\alpha)$. Using the fact that $c^2 + s^2 = 1$, we have

$$\mathbf{R}^T\mathbf{R} = \begin{pmatrix} c & s & 0 \\ -s & c & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} c & -s & 0 \\ s & c & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} c^2 + s^2 + 0 & -cs + sc + 0 & 0 \\ -sc + sc + 0 & c^2 + s^2 + 0 & 0 \\ 0 & 0 & 1 \end{pmatrix} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{pmatrix}$$

2. Let A be an orthogonal matrix, and λ be the eigenvalue corresponding to the eigenvector x, then

$$Ax = \lambda x$$

   Since the transpose of a matrix has the same eigenvalue as the original matrix, we have

$$|\lambda|^2 x^T x = (Ax)^T(Ax) = x^T A^T A x = x^T x$$

   So we can conclude $|\lambda| = 1$

   Remark: we only consider the real eigenvalue.

Problem 2 (10pts) Optimization.
Prove that:

(1) $f(x) = |x|$ is convex;

(2) $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2$ is convex, where $\mathbf{A}$ is a matrix.

Solution:

(1)

$$\begin{aligned} f\left((1-\lambda)x_1 + \lambda x_2\right) &= \left|(1-\lambda)x_1 + \lambda x_2\right| \\ &\leq \left|(1-\lambda)x_1\right| + \left|\lambda x_2\right| \quad \text{by the triangle inequality} \\ &= (1-\lambda)\left|x_1\right| + \lambda\left|x_2\right| \quad \text{because } \lambda, 1-\lambda \geq 0 \\ &= (1-\lambda)f\left(x_1\right) + \lambda f\left(x_2\right) \end{aligned}$$

Therefore f is convex.

(2) $f(\mathbf{x}) = \|\mathbf{A}\mathbf{x} - \mathbf{b}\|^2 = ||\mathbf{A}\mathbf{x}||^2 + ||\mathbf{b}||^2 - 2\mathbf{b}^T\mathbf{A}\mathbf{x}$ $f(\mathbf{x})$ is twice differentiable and we want to get its second derivative (i.e., Hessian)

$$\frac{\partial f}{\partial \mathbf{x}} = -2\mathbf{b}^T\mathbf{A} + 2\mathbf{x}^T\mathbf{A}^T\mathbf{A}$$

$$\frac{\partial^2 f}{\partial \mathbf{x} \partial \mathbf{x}^T} = 2\mathbf{A}^T\mathbf{A}$$

which is a positive semi-definite matrix. Therefore f is a convex function.

Remark: This is actually the least square problem. Because the least square problem is convex, we can get the global solution by some optimization methods like gradient descent.

Problem 3 (10pts) Information Theory.
Proof that cross-entropy is not smaller than entropy, i.e., $H_{P,Q}(\mathcal{X}) \geq H_P(\mathcal{X})$, and the equality holds only when $P = Q$.
Solution: Given two distributions P and Q. Cross entropy is: $H_{P,Q}(\mathcal{X}) = -\sum_x p(x) \log q(x)$. First, you'll manipulate it to obtain the very well-known form: $H_{P,Q}(\mathcal{X}) = H_P(\mathcal{X}) + D_{KL}(p\|q)$, where $D_{KL}(p\|q)$ is the KL distance. Then, it only remains to prove that $D_{KL}(p,q) \geq 0$. And when $P = Q$, KL divergence is 0.
Remark: these 2 properties have been proved in tutorial 2.

Problem 4 (10pts) Linear Regression.
Suppose we have training data $\{(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_N, y_N)\}$, where $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}, i = 1, 2, \ldots, N$. Consider $f_{\mathbf{w},b}(\mathbf{x}_i) = \mathbf{x}_i^T\mathbf{w} + b$, where $\mathbf{w} = [w_1, w_2, \ldots, w_d]^T$.

(1) Find the closed-form solution of the following problem

$$\min_{\mathbf{w},b} \sum_{i=1}^{N} (f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)^2 + \lambda \bar{\mathbf{w}}^T \bar{\mathbf{w}}, \tag{1}$$

where $\bar{\mathbf{w}} = \hat{\mathbf{I}}_d \mathbf{w} = [0, w_1, w_2, \dots, w_d]^T$. Note that $\hat{\mathbf{I}}_d = \begin{bmatrix} 0 & 0 & 0 & \dots & 0 \\ 1 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ & & & \vdots & \\ 0 & \dots & 0 & 0 & 1 \end{bmatrix} \in \mathbf{R}^{(d+1)\times d}$

(2) Show how to use gradient descent to solve the problem.

Solution:

1. $$\begin{aligned} \min\nolimits_{w,b}(Xw - y)^T(Xw - y) + \lambda \bar{w}^T \bar{w} \\ \frac{\partial}{\partial w}(Xw - y)^T(Xw - y) + \lambda \bar{w}^T \bar{w} = 0 \\ 2X^t Xw - 2X^T y + 2\lambda \hat{I}_d w = 0 \\ X^T Xw + \lambda \hat{I}_d w = x^T y \\ \left(X^T X + \lambda \hat{I}_d\right) w = x^T y \\ w = \left(X^T X + \lambda \hat{I}_d\right)^{-1} X^T y \end{aligned}$$

2. You will get points for the second question as long as your answer is reasonable.

Problem 5 (10pts) MLE.
Consider a linear regression model with a 2-dimensional response vector $\boldsymbol{y}_i \in \mathbb{R}^2$. Suppose we have some binary input data, $x_i \in \{0, 1\}$. The training data is as follows:

x	y
0	$(-1, -1)^T$
0	$(-1, -2)^T$
0	$(-2, -1)^T$
1	$(1, 1)^T$
1	$(1, 2)^T$
1	$(2, 1)^T$

Let us embed each x_i into 2 d using the following basis function:

$$\phi(0) = (1, 0)^T, \quad \phi(1) = (0, 1)^T$$

The model becomes

$$\hat{\boldsymbol{y}} = \mathbf{W}^T \boldsymbol{\phi}(x)$$

where $\mathbf{W}$ is a 2×2 matrix. Compute the MLE for $\mathbf{W}$ from the above data.

Solution: In this exercise, we have 2 independent responses, the MLE for $\mathbf{W}$ can be considered separately as $\hat{\mathbf{W}} = [\hat{\mathbf{w}}_1, \hat{\mathbf{w}}_2]$. To find the parameters that assign the highest probability to the data, we can find the one which minimizes **RSS** for the proposed linear regression model $\hat{\mathbf{y}} = \mathbf{W}^T \phi(x)$. Take $\mathbf{y} = \begin{bmatrix} y_1^T \\ y_2^T \end{bmatrix}$, where $y_1^T = [-1, -1, -2, 1, 1, 2]$ and $y_2^T = [-1, -2, -1, 1, 2, 1]$

$$\hat{\mathbf{w}}_1 = \left(X^T X\right)^{-1} X^T y_1$$
$$\hat{\mathbf{w}}_2 = \left(X^T X\right)^{-1} X^T y_2$$

Since $\phi(0) = (1,0)^T, \quad \phi(1) = (0,1)^T$, we denote $X^T = \begin{pmatrix} 1 & 1 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 1 & 1 \end{pmatrix}$ By solve the equation above, we get $\hat{\mathbf{W}} = \begin{pmatrix} -\frac{4}{3} & -\frac{4}{3} \\ \frac{4}{3} & \frac{4}{3} \end{pmatrix}$.