ML Written Homework 1

Student ID: b10201054

September 23, 2025

1 Linear Algebra Recap

(a) (i) $\Rightarrow$ (ii) $\mathbf{A}$ is positive semi-definite implies $\mathbf{A}$ is symmetric. Since $\mathbf{A}$ is symmetric, $\mathbf{A}$ is diagonalizable. Suppose $\lambda_1, \lambda_2, ..., \lambda_n$ are the eigenvalues of $\mathbf{A}$, there are eigenvectors $v_1, v_2, ..., v_n$ such that $\mathbf{A}v_k = \lambda_k v_k$.

By the definition of positive semi-definite, we have

$$\langle \mathbf{A}v_k, v_k \rangle = \langle \lambda_k v_k, v_k \rangle = \lambda_k \|v_k\|^2 \geq 0$$

Hence, all eigenvalues $\lambda_k \geq 0$.

(ii) $\Rightarrow$ (iii) $\mathbf{A}$ is symmetric and all of eigenvalues λ_k are non negative. Since $\mathbf{A}$ is symmetric, $\mathbf{A}$ is diagonalizable. That is, there are P, D such that $\mathbf{A} = P^T DP$, where D = $diag(\lambda_1, \lambda_2, ..., \lambda_n)$.

Since $\lambda_k \geq 0$, we can find $e_k = \sqrt{\lambda_k}$ and $E = diag(e_1, e_2, ..., e_n)$ so that $D = E^2$. Now we have

$$\mathbf{A} = P^T DP = P^T EEP = (P^T E)(EP) = (EP)^T (EP)$$

Hence, $\mathbf{A} = \mathbf{B}^T\mathbf{B}$ where $\mathbf{B} = EP$

(iii) $\Rightarrow$ (i) $\mathbf{A} = \mathbf{B}^T\mathbf{B}$, then we have

$$u^T\mathbf{A}u = u^T\mathbf{B}^T\mathbf{B}u = (\mathbf{B}u)^T(\mathbf{B}u) = \|\mathbf{B}u\|^2 \geq 0$$

for all $u \in V$. Hence, $\mathbf{A}$ is positive semi-definite.

(b) Note that $\langle \mathbf{A}^T\mathbf{A}x, x \rangle = \langle \mathbf{A}x, \mathbf{A}x \rangle = \|\mathbf{A}x\|^2 \geq 0$ for all $x \in V$. By (a), $\mathbf{A}^T\mathbf{A}$ is positive semi-definite. Hence the eigenvalues of $\mathbf{A}^T\mathbf{A}$ are nonnegative.

(c) For $x \in Null(\mathbf{A})$, we have $\mathbf{A}^T\mathbf{A}x = 0$, so that $x \in Null(\mathbf{A}^T\mathbf{A})$.

For $x \in Null(\mathbf{A}^T\mathbf{A})$,

$$\langle \mathbf{A}^T\mathbf{A}x, x \rangle = \langle \mathbf{A}x, \mathbf{A}x \rangle = \|\mathbf{A}x\|^2 = 0$$

which implies $\mathbf{A}x = 0$. Hence $x \in Null(\mathbf{A})$. By the above, we can find that $Null(\mathbf{A}) = Null(\mathbf{A}^T\mathbf{A})$.

(d) $\langle \mathbf{A}v_i, \mathbf{A}v_j \rangle = \langle \mathbf{A}^T\mathbf{A}v_i, v_j \rangle = \langle \lambda_i v_i, v_j \rangle = 0$ for any $i \neq j$. Hence $\{\mathbf{A}v_1, ..., \mathbf{A}v_r\}$ is an orthogonal set.

(e) Suppose V is a $n \times n$ matrix whose i-th column is v_i, then V is an orthogonal matrix where $\mathbf{A}^T\mathbf{A} = VDV^T$. Let U be a $m \times m$ matrix with its i-th column u_i, where $u_i = \frac{\mathbf{A}v_i}{\sigma_i}$, and Σ be a $m \times n$ matrix with its diagonal entries σ_i.

Then we have the relation:

$$U\Sigma = \mathbf{A}V \Rightarrow \mathbf{A} = U\Sigma V^T$$

By (d), $\{\mathbf{A}v_1, ..., \mathbf{A}v_r\}$ is an orthogonal set, we have $\{\mathbf{u}_1, ..., \mathbf{u}_r\}$ is also an orthogonal set. Hence U is also an orthogonal matrix.

2 Definition of Derivative as Linear Operator

(a) Suppose $\mathbf{x} = (x_1, x_2, ..., x_n)^T, \mathbf{a} = (a_1, a_2, ..., a_n)^T$. $\|\mathbf{x} - \mathbf{a}\|_2 = \sqrt{\sum_{i=1}^n (x_i - a_i)^2}$.

$$\frac{\partial \sqrt{\sum_{i=1}^n (x_i - a_i)^2}}{\partial x_k} = \frac{1}{2} \cdot \frac{2(x_k - a_k)}{\sqrt{\sum_{i=1}^n (x_i - a_i)^2}} = \frac{x_k - a_k}{\|\mathbf{x} - \mathbf{a}\|_2}$$

Then we have

$$\frac{\partial \|\mathbf{x} - \mathbf{a}\|_2}{\partial \mathbf{x}} = (\frac{x_1 - a_1}{\|\mathbf{x} - \mathbf{a}\|_2}, ..., \frac{x_n - a_n}{\|\mathbf{x} - \mathbf{a}\|_2})^T = \frac{\mathbf{x} - \mathbf{a}}{\|\mathbf{x} - \mathbf{a}\|_2}$$

(b) Suppose $\mathbf{a} = (a_1, ..., a_m)^T, \mathbf{b} = (b_1, ..., b_n)^T, \mathbf{X} = (x_{ij})$.

$$\mathbf{a}^T\mathbf{X}\mathbf{T} = \sum_{j=1}^n \sum_{i=1}^m a_i x_{ij} b_j$$

$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{T}}{\partial x_{ij}} = \frac{\sum_{j=1}^n \sum_{i=1}^m a_i x_{ij} b_j}{\partial x_{ij}} = a_i b_j$$

Then we have

$$\frac{\partial \mathbf{a}^T\mathbf{X}\mathbf{T}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{b}^T$$

(c) Suppose $\mathbf{x} = (x_1, ..., x_n)^T, \mathbf{A} = (A_{ij})$.

$$\mathbf{x}^T\mathbf{A}\mathbf{x} = \sum_{j=1}^n \sum_{i=1}^n x_i A_{ij} x_j$$

$$\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial x_k} = \frac{\partial \sum_{j=1}^n \sum_{i=1}^n x_i A_{ij} x_j}{\partial x_k} = \sum_{i=1}^n x_i A_{ik} + \sum_{j=1}^n A_{kj} x_j = (\mathbf{A}^T\mathbf{x})_k + (\mathbf{A}\mathbf{x})_k$$

Hence, we have

$$\frac{\partial \mathbf{x}^T\mathbf{A}\mathbf{x}}{\partial \mathbf{x}} = (\mathbf{A} + \mathbf{A}^T)\mathbf{x}$$

3 Matrix Calculus

(a) We have $\det(\mathbf{A}) = \sum_{j=1}^{n} \sum_{i=1}^{n} A_{ij} * C_{ij}$, where $C_{ij} = (-1)^{i+j} M_{ij} = adj(\mathbf{A})_{ji}$. Thus we have

$$\frac{\partial \det(\mathbf{A})}{\partial A_{ij}} = C_{ij} = adj(\mathbf{A})_{ji}$$

Hence,

$$\frac{\partial \det(\mathbf{A})}{\partial \mathbf{A}} = adj(\mathbf{A})^T = \det(\mathbf{A})(\mathbf{A}^{-1})^T$$

(b) By chain rule:

$$\frac{\partial \log(\det(\mathbf{A}))}{\partial \mathbf{A}} = \frac{\partial \log(\det(\mathbf{A}))}{\partial \det(\mathbf{A})} \frac{\partial \det(\mathbf{A})}{\partial \mathbf{A}} = \frac{1}{\det(\mathbf{A})} \det(\mathbf{A})(\mathbf{A}^{-1})^T = (\mathbf{A}^{-1})^T$$

4 Closed-Form Linear Regression Solution

(a) Suppose $\mathbf{K} = diag(\kappa_1, ..., \kappa_m)$

$$\sum_{i}^{m} \kappa_i (y_i - \mathbf{X}_i \theta)^2 + \lambda \sum_{j}^{m} \omega_j^2 = (y - \mathbf{X}\theta)^T \mathbf{K} (y - \mathbf{X}\theta) + \lambda \theta^T \theta$$

The optimal solution happens when the gradient = 0

$$\frac{\partial (\sum_i^m \kappa_i (y_i - \mathbf{X}_i \theta)^2 + \lambda \sum_j^m \omega_j^2)}{\partial \theta} = -2\mathbf{X}^T \mathbf{K} (y - \mathbf{X}\theta) + 2\lambda \mathbf{I}\theta = 0$$

This gives us

$$\theta^* = (\mathbf{X}^T \mathbf{K} \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{K} y$$

(b)

5 Noise and Regularization

(a)

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

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

$$= \frac{1}{2N} \sum_{i=1}^{N} (f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)^2 + (\mathbf{w}^T \eta_i)^2 + 2(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)(\mathbf{w}^T \eta_i) \tag{3}$$

Since expect value $\mathbb{E}$ is linear, we can seperate the formula into three part.

$$\mathbb{E}((f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)^2) = (f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)^2$$

$$\mathbb{E}((\mathbf{w}^T\eta_i)^2) = \sigma^2 \left\|\mathbf{w}\right\|^2$$

$$2(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)(\mathbf{w}^T\eta_i) = 2(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)\sum_{j=1}^{k} w_j\eta_{i,j}$$

$$\mathbb{E}(2(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)\sum_{j=1}^{k} w_j\eta_{i,j}) = 2(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)\sum_{j=1}^{k} w_j\mathbb{E}(\eta_{i,j}) = 0$$

$$\tilde{L}(\mathbf{w}, b) = \mathbb{E}(\frac{1}{2N}\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x}_i + \eta_i) - y_i)^2)$$

$$\tilde{L}_{ss}(\mathbf{w}, b) = \mathbb{E}(\frac{1}{2N}\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x}_i + \eta_i) - y_i)^2) \tag{4}$$

$$= \frac{1}{2N}\sum_{i=1}^{N}((f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)^2 + \sigma^2 \left\|\mathbf{w}\right\|^2) \tag{5}$$

$$= \frac{1}{2N}\sum_{i=1}^{N}(f_{\mathbf{w},b}(\mathbf{x}_i) - y_i)^2 + \frac{\sigma^2}{2} \left\|\mathbf{w}\right\|^2 \tag{6}$$