CIS5200: Machine Learning Spring 2025

Homework 0

Release Date: January 16, 2025 *Due Date: January 28, 2025*

Name: Haoze Wu

PennKey: haozewu

Check out shorthands.sty for some convenient shortcuts

Reminder to show all work for full credit!

1 Written Questions

A1

Solution:

1. Proof:

   Considering one property of the determinant is that the determinant of a matrix product is the product of the corresponding determinants, that is, $\mathbf{det}(AB) = \mathbf{det}(A)\,\mathbf{det}(B)$, and the fact that $AA^{-1} = I$.

   Thus, we have: $\mathbf{det}(I) = \mathbf{det}(AA^{-1}) = \mathbf{det}(A)\,\mathbf{det}(A^{-1}) = 1$.

   Thus, we have $\mathbf{det}(A) = \frac{1}{\mathbf{det}(A^{-1})}$ for any invertible real-valued square matrix A.

2. Partial Proof:

   Proving the statement in the question is hard for me, but I can prove that for a symmetric, yet square and real valued matrix A this property holds.

   According to the Spectral Theorem, for the real-valued matrix $A \in R^{n\times n}$, all its eigenvalues are real and, moreover, we have $A = \sum_{i=1}^{N} \lambda_i v_i v_i^T$ where λ_i is the i-th eigenvalue of A and $v_1, v_2, \cdots, v_n$ are a set of orthonormal unit vectors.

   By definition, we have $tr(A) = \sum_{i=1}^{n} a_{ii}$, which is equivalent to $tr(A) = tr(\sum_{i=1}^{n} \lambda_i v_i v_i^T) = \sum_{i=1}^{n} \lambda_i tr(v_i v_i^T)$. And thus, we have $tr(v_i v_i^T) = \sum_{j=1}^{n} (v_i v_i^T)_{jj} = 1$. Thus, we can reach the conclusion that $tr(A) = tr(\sum_{i=1}^{N} \lambda_i v_i v_i^T) = \sum_{i=1}^{n} \lambda_i$.

3. This statement is incorrect. Consider the following matrix: $\begin{bmatrix}1 & 0 & 0\\ 0 & 1 & 0\\ 0 & 0 & 0\end{bmatrix}$, its rank is 2 while it only has one non-zero eigenvalue 1, which is contradicted to the statement.

A2

Solution:

1. To calculate the nullspace of matrix A, we may solve the following equations: $\begin{bmatrix} 2 & -1 \\ 4 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ $= \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, which is equivalent to

$$\begin{aligned} 2x_1 - x_2 &= 0 \\ 4x_1 - 2x_2 &= 0 \end{aligned} \tag{1}$$

   which leads to the relationship between x_1 and x_2: $x_2 = 2x_1$. Thus, the nullspace of matrix A is spanned by the vector $\begin{bmatrix} 1 \\ 2 \end{bmatrix}$, which is a subspace of R^2.

2. We first perform the row redution on the matrix A: $\begin{bmatrix} 2 & -1 \\ 4 & -2 \end{bmatrix} \xrightarrow{R_2 - 2R_1} \begin{bmatrix} 2 & -1 \\ 0 & 0 \end{bmatrix}$ Thus, the row space of matrix A is spanned by the vector $\begin{bmatrix} 2 \\ -1 \end{bmatrix}$, which is a subspace of R^2.

   To test whether the vector $[1, 1]^T$ is in the row space of matrix A, is equivalent to test whether there exists a linear combination of the row vectors of matrix A that equals to the vector $[1, 1]^T$, which is equivalent to solve the following equations: $\alpha \cdot \begin{bmatrix} 2 \\ -1 \end{bmatrix} = \begin{bmatrix} 1 \\ 1 \end{bmatrix}$, which has no solution for the variable α.

   Thus, the vector $[1, 1]^T$ is not in the row space of matrix A.

A3

Solution:

1. By definition, we have $\mathbf{det}(A - \lambda I) = 0$ for the eigenvalues of matrix A. Thus, we have

$$\begin{aligned} \mathbf{det}(A - \lambda I) &= \begin{vmatrix} 2-\lambda & 1 \\ 1 & 2-\lambda \end{vmatrix} = 0 \\ (2-\lambda)^2 - 1 &= 0 \\ \lambda^2 - 4\lambda + 3 &= 0 \\ (\lambda - 3)(\lambda - 1) &= 0 \end{aligned} \tag{2}$$

   Thus, the eigenvalues of matrix A are $\lambda_1 = 3$ and $\lambda_2 = 1$. Then, we can use the fact that $Ax = \lambda x$ to find the eigenvectors corresponding to an eigenvalue of matrix A: For $\lambda_1 = 3$, we have:

$$\begin{aligned} (A - 3I)x &= 0 \\ \begin{bmatrix} -1 & 1 \\ 1 & -1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{aligned} \tag{3}$$

   which leads to the relationship between x_1 and x_2: $x_1 = x_2$. Thus, the eigenvector corresponding to the eigenvalue 3 is $\begin{bmatrix} 1 \\ 1 \end{bmatrix}$. For the second eigenvalue $\lambda_2 = 1$, we have:

$$\begin{aligned} (A - I)x &= 0 \\ \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} &= \begin{bmatrix} 0 \\ 0 \end{bmatrix} \end{aligned} \tag{4}$$

   which leads to the relationship between x_1 and x_2: $x_1 = -x_2$. Thus, the eigenvector corresponding to the eigenvalue 1 is $\begin{bmatrix} 1 \\ -1 \end{bmatrix}$.

2. According to the property of the PSD matrix, a matrix A is PSD iff all its eigenvalues are non-negative. Thus, the matrix A is PSD since all its eigenvalues are non-negative as we have calculated in the previous question.

3. From the definition of the SVD, we have $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T$, where u_i and v_i are the left and right singular vectors of matrix A and σ_i is the singular value of matrix A. And by the definition of the Spectral Theroem, we have $A = \sum_{i=1}^{n} \lambda_i v_i v_i^T$, where λ_i is the i-th eigenvalue of matrix A and $v_1, v_2, \cdots, v_n$ are a set of orthonormal unit vectors. Thus, we may rewrite the SVD in terms of the eigenvalues and eigenvectors of matrix A: $A = \sum_{i=1}^{r} \sigma_i u_i v_i^T = \sum_{i=1}^{n} \lambda_i v_i v_i^T$ Combined with the eigenvalues and eigenvectors we have calculated in the previous question, after normalization, we have: $A = 3 \begin{bmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{bmatrix} + 1 \begin{bmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{bmatrix}$

   Here, the singular values of matrix A are equal to the eigenvalues of matrix A.

A4

Solution:

1. Let $y = -w^T x$, thus we have $\nabla_x f(x) = \frac{\partial(f(y))}{\partial(y)} \frac{\partial(y)}{\partial x}$ For the first term, we have $\frac{\partial(f(y))}{\partial(y)} = \frac{\partial(\frac{1}{1+e^y})}{\partial(y)} = -\frac{e^y}{(1+e^y)^2} = -\frac{e^{-w^T x}}{(1+e^{-w^T x})^2}$. For the second term, we have $\frac{\partial(y)}{\partial x} = \frac{\partial(-w^T x)}{\partial x} = -w$. Thus, we have $\nabla_x f(x) = -\frac{e^{-w^T x}}{(1+e^{-w^T x})^2} w$.

2. Rewrite $f(x) = ||Ax-b||_2^2$, we have $f(x) = (Ax-b)^T(Ax-b) = x^T A^T Ax - 2b^T Ax + b^T b$. Thus, for the gradient of $f(x)$, we have $\nabla_x f(x) = \nabla_x(x^T A^T Ax - 2b^T Ax + b^T b) = 2A^T Ax - 2A^T b$.

A5

Solution:

1. When the hyperplane passes through the origin, we have $w^T x_0 + b = 0$ for the point x_0 on the hyperplane and x_0 is the origin, which indicates that the scalar $b = 0$.

2. To calculate the distance from any given point x_0 to the hyperplane, we may use the definition of the distance:

$$\begin{aligned} &\frac{|w^T x_0 + b|}{||w||} \\ &= \frac{|w^T x_0|}{||w||} \end{aligned} \tag{5}$$

A6

Solution:

1. Given that $||x||_\infty = 1$, which is equivalent to $\max |x_i| = 1$, we have:

$$||x||_2 = \sqrt{\sum_{i=1}^{n} x_i^2} \leq \sqrt{\sum_{i=1}^{n} \max |x_i|^2} = n \tag{6}$$

2. Given that $||x||_2 = 1$, which is equivalent to $\sqrt{\sum_{i=1}^n x_i^2} = 1$, and we then have $\sqrt{\sum_{i=1}^n x_i^2} \leq \sum_{i=1}^n \sqrt{x_i^2} = \sum_{i=1}^n |x_i|$. Thus, we have $||x||_1 = \sum_{i=1}^n |x_i| \geq \sqrt{\sum_{i=1}^n x_i^2} = 1$.

A7

Solution:

1. Consider the fact that a function is a convex function iff its second derivative is non-negative, we may calculate the second derivative of the function $f(x) = x^3$:

$$\begin{aligned} f'(x) &= 3x^2 \\ f''(x) &= 6x \end{aligned} \tag{7}$$

which is negative when $x < 0$. Thus, the function $f(x) = x^3$ is not a convex function over R.

2. We start with calculating the second derivative of the function $f(x) = x^4 + \alpha x^2$:

$$\begin{aligned} f'(x) &= 4x^3 + 2\alpha x \\ f''(x) &= 12x^2 + 2\alpha \end{aligned} \tag{8}$$

which should be non-negative for all x over R. Thus, we have $12x^2 + 2\alpha \geq 0$ for all x over R, which leads to $\alpha \geq 0$.

A8

Solution:We may use the conditional probability here:

$$\begin{aligned} P(A|B) &= \frac{P(A \cap B)}{P(B)} \\ &= \frac{P(A)P(B|A)}{P(B)} \end{aligned} \tag{9}$$

where event A is the the email is actually a spam while event B is that the email is flagged as a spam by the system, where $P(A) = 1 - 0.8 = 0.2$ and $P(B) = 0.90 * (1 - 0.80) + (1 - 0.95) * 0.80 = 0.22$. Thus, we have:

$$\begin{aligned} P(A|B) &= \frac{0.2 * 0.90}{0.22} \\ &= 0.8182 \end{aligned} \tag{10}$$

A9

Solution:

1. Y is also a random variable following normal distribution with the mean as $E[Y] = E[\sum_{i=1}^{n} a_i X_i] = \sum_{i=1}^{n} a_i E[X_i] = \sum_{i=1}^{n} a_i \mu_i$ and the variance as $Var[Y] = Var[\sum_{i=1}^{n} a_i X_i] = \sum_{i=1}^{n} a_i^2 Var[X_i] = \sum_{i=1}^{n} a_i^2 \sigma_i^2$. Thus, $Y \sim N(\sum_{i=1}^{n} a_i \mu_i, \sum_{i=1}^{n} a_i^2 \sigma_i^2)$.

2. When $\mu_1 = 0$ and $\sigma_i^2 = 1$, X_i is vairable following the standard normal distribution. $\Pr[\max_{1\leq i\leq n} X_i > 2]$ is equivalent to $1 - \Pr[\max_{1\leq i\leq n} X_i \leq 2]$, which is equivalent to $1 - \Pr[X_1 \leq 2, X_2 \leq 2, \cdots, X_n \leq 2]$, which is equivalent to $1 - \Pr[X_1 \leq 2] \Pr[X_2 \leq 2] \cdots \Pr[X_n \leq 2]$, which is equivalent to $1 - (\Phi(2))^n$.

A10

Solution:

1. The probability of seeing any side among all six sides of this fair die is $\frac{1}{6}$. Thus, the expected number of rolls to see a 6 is $E[X] = \frac{1}{\frac{1}{6}} = 6$.

2. The probability of seeing a 6 after seeing another 6 is still $\frac{1}{6}$. Thus, the expected number of rolls is $E[X] = \frac{1}{\frac{1}{6\times 6}} = 36$.

2 Python Programming Questions

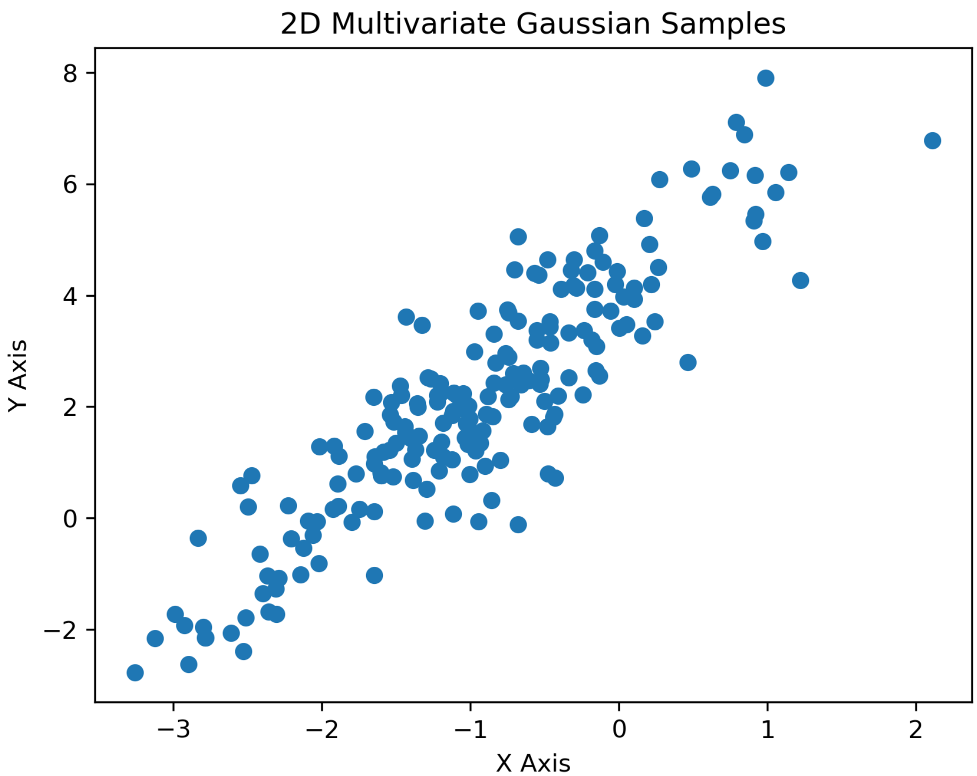

Figure 1: Figure for Q4 (MatplotLib)