Introduction to Machine Learning (67577)

Exercise 1
Estimation Theory & Mathematical Background

2 Theoretical Part

2.1 Mathematical Background

2.1.1 Linear Algebra

1. Prove that orthogonal matrices are isometric transformations. That is, let $T : V \mapsto W$ be some linear transformation and A the corresponding matrix. Show that if A is an orthogonal matrix then $\forall x \in V \; ||Ax|| = ||x||$.

1 תהי מטריצה אורתוגונלית A ויהי $x \in V$, אזי

$$||Ax|| = \sqrt{\langle Ax, Ax \rangle} = \sqrt{x^T A^T A x} \underset{A^TA = I}{=} \sqrt{x^T x} = \sqrt{\langle x, x \rangle} = ||x||$$

$A^TA = I$ כי A אורתוגונלית

2. Calculate the SVD of the following matrix A. That is, find the matrices $U, \Sigma, V^\top$ where U, V are orthogonal matrices and Σ diagonal.

$$A = \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 2 \end{bmatrix}$$

Recall, that to find the SVD of A we can calculate $A^\top A$ to deduce V,Σ and then calculate $AA^\top$ to deduce U. Equivalently, once we deduced V, Σ we can fine U using the equality $AV = U\Sigma$.

$$A^TA = \begin{bmatrix} 1 & 1 \\ 1 & -1 \\ 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 1 & -1 & 2 \end{bmatrix} = \begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 2 & -2 & 4 \end{bmatrix}$$

נמצא ע"ע של A:

$$\det(A^TA - I\lambda) = \begin{vmatrix} 2-\lambda & 0 & 2 \\ 0 & 2-\lambda & -2 \\ 2 & -2 & 4-\lambda \end{vmatrix}$$

$$= (2-\lambda)\begin{vmatrix} 2-\lambda & -2 \\ -2 & 4-\lambda \end{vmatrix} + 2\begin{vmatrix} 0 & 2-\lambda \\ 2 & -2 \end{vmatrix}$$

$$= (2-\lambda)\left(4-6\lambda+\lambda^2\right) - 4\left(2-\lambda\right)$$

$$= (2-\lambda)\left(\lambda^2-6\lambda\right) = -\lambda(\lambda-2)(\lambda-6)$$

$$\boxed{\lambda = 0, 2, 6}$$

ע״ע:

$\lambda = 0$:

$$\begin{bmatrix} 2 & 0 & 2 \\ 0 & 2 & -2 \\ 2 & -2 & 4 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \bar{0} \to \begin{matrix} x = -z \\ y = z \\ x - y + 2z = 0 \end{matrix}$$

ו״ע:

$$\frac{1}{\sqrt{3}}\begin{bmatrix} 1 \\ -1 \\ -1 \end{bmatrix}$$

$\lambda = 2$

$$\begin{bmatrix} 0 & 0 & 2 \\ 0 & 0 & -2 \\ 2 & -2 & 2 \end{bmatrix}\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \bar{0} \to \begin{matrix} z = 0 \\ x = y \end{matrix}$$

$$\frac{1}{\sqrt{2}}\begin{bmatrix}1\\1\\0\end{bmatrix}$$ נ"ר לכן

$$\underline{\lambda=6}:\quad \begin{bmatrix}-4 & 0 & 2\\ 0 & -4 & -2\\ 2 & -2 & -2\end{bmatrix}\begin{bmatrix}x\\y\\z\end{bmatrix}=\bar{0}\ \rightarrow\ \begin{matrix}z=2x\\ z=-2y\end{matrix}$$

$$\frac{1}{\sqrt{6}}\begin{bmatrix}1\\-1\\2\end{bmatrix}$$ נ"ר לכן

$$V=\begin{bmatrix}\frac{1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{3}}\\ \frac{-1}{\sqrt{6}} & \frac{1}{\sqrt{2}} & \frac{-1}{\sqrt{3}}\\ \frac{2}{\sqrt{6}} & 0 & \frac{-1}{\sqrt{3}}\end{bmatrix}\quad D=\begin{bmatrix}6 & 0 & 0\\ 0 & 2 & 0\\ 0 & 0 & 0\end{bmatrix}$$ לכן

$$\sigma_1=\sqrt{6},\ \sigma_2=\sqrt{2},\ \sigma_3=0$$ לכן

נמצא את U:

$$AA^T=\begin{bmatrix}1 & 1 & 0\\ 1 & -1 & 2\end{bmatrix}\begin{bmatrix}1 & 1\\ 1 & -1\\ 0 & 2\end{bmatrix}$$

$= \begin{bmatrix} 2 & 0 \\ 0 & 6 \end{bmatrix}$ $\lambda_{1,2} = 2, 6$ לכן ע״ע הם

$\underline{\lambda = 2}$: $\begin{bmatrix} 0 & 0 \\ 0 & 4 \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} = \bar{0} \rightarrow$ $\begin{bmatrix} 1 \\ 0 \end{bmatrix}$:ו״ע

$\underline{\lambda = 6}$: $\begin{bmatrix} 0 \\ 1 \end{bmatrix}$:ו״ע

לכן

$$AA^T = U \begin{bmatrix} 6 & 0 \\ 0 & 2 \end{bmatrix} U^T$$

$U = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ כך ש

לכן

$$A = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix} \begin{bmatrix} \sqrt{6} & 0 & 0 \\ 0 & \sqrt{2} & 0 \end{bmatrix} \begin{bmatrix} \frac{1}{\sqrt{6}} & \frac{-1}{\sqrt{6}} & \frac{2}{\sqrt{6}} \\ \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} & 0 \\ \frac{1}{\sqrt{3}} & \frac{-1}{\sqrt{3}} & \frac{-1}{\sqrt{3}} \end{bmatrix}$$

3. In this question we prove the Power-Iteration algorithm for finding the SVD of a matrix. Let $A \in \mathbb{R}^{m\times n}$ and define $C_0 = A^\top A$. Denote $\lambda_1 \geq \ldots \geq \lambda_n$ the eigenvalues of C_0, with the corresponding normalized eigenvectors $\mathbf{v}_1, \ldots, \mathbf{v}_n$.

   Let us assume the $\lambda_1 > \lambda_2$. Define $b_k \in \mathbb{R}$ as follows:

$$b_0 = \sum_{i=1}^{n} a_i v_i, \quad b_{k+1} = \frac{C_0 b_k}{||C_0 b_k||}$$

   where $a_1 \neq 0$. Show that: $\lim_{k\to\infty} b_k = \pm v_1$.

$$b_{k+1} = \frac{C_0 b_k}{||C_0 b_k||} = C_0 \cdot \frac{C_0 b_{k-1}}{||C_0 b_{k-1}||} \cdot \frac{1}{\left|\left| \frac{C_0 \cdot C_0 b_{k-1}}{||C_0 b_{k-1}||} \right|\right|}$$

$$= C_0^2 \frac{b_{k-1}}{||C_0 b_{k-1}||} \cdot \frac{||C_0 b_{k-1}||}{||C_0^2 b_{k-1}||} = \frac{C_0^2 b_{k-1}}{||C_0^2 b_{k-1}||}$$

$$b_k = \frac{C_0^k b_0}{||C_0^k b_0||} = \frac{C_0^k \sum a_i V_i}{||C_0^k \sum a_i V_i||}$$

ולכן

$$= \frac{(V \Sigma^T \Sigma V^T)^k \sum a_i V_i}{||(V \Sigma^T \Sigma V^T)^k \sum a_i V_i||} = \frac{\sum V D^k V^T a_i V_i}{||\sum V D^k V^T a_i V_i||}$$

$$= \frac{\sum a_i \lambda_i^k Ve_i}{||\sum a_i \lambda_i^k Ve_i||} = \frac{\sum a_i \lambda_i^k V_i}{||\sum a_i \lambda_i^k V_i||}$$

מניחין ש $\lambda_1 > \dots > \lambda_n$ נקבל ש:

$$= \frac{\lambda_1^k\left(\sum a_i\left(\frac{\lambda_i}{\lambda_1}\right)^k v_i\right)}{\lambda_1^k\left\|\sum a_i\left(\frac{\lambda_i}{\lambda_1}\right)^k v_i\right\|} = \frac{a_1 v_1}{\|a_1 v_1\|} = v_1$$

עבור $k \to \infty$

v_i מנורמל.

2.1.2 Multivariate Calculus

4. Let $x \in \mathbb{R}^n$ be a fixed vector and $U \in \mathbb{R}^{n \times n}$ a fixed orthogonal matrix. Calculate the Jacobian of the function $f : \mathbb{R}^n \to \mathbb{R}^n$:

$$f(\sigma) = U \cdot \text{diag}(\sigma) U^\top x$$

Where $\text{diag}(\sigma)$ is an $n \times n$ matrix where

$$\text{diag}(\sigma)_{ij} = \begin{cases} \sigma_i & i = j \\ 0 & i \neq j \end{cases}$$

$$f(\sigma) = U\,\text{diag}(\sigma)\,U^T x = U \sum \sigma_i u_i^T x$$

לכן $\frac{\partial f}{\partial \sigma_i} = U$ וסה"כ

$$\boxed{Jac(f) = U\,\text{diag}(U^T x)}$$

5. Use the chain rule to calculate the gradient of $h(\sigma) = \frac{1}{2}||f(\sigma) - y||^2$

$$h(\sigma) = \frac{1}{2}||f(\sigma) - y||^2 =$$

לכן מכלל השרשרת:

$$\nabla h(\sigma) = \left(f(\sigma) - y\right)^T \nabla f(\sigma)$$

6. Calculate the Jacobian of the softmax function $S : \mathbb{R}^d \to [0,1]^k$

$$S(\mathbf{x})_j = \frac{e^{x_j}}{\sum_{l=1}^{k} e^{x_l}}$$

$$\frac{\partial S(x)_j}{\partial x_i} \underset{i \neq j}{=} -\frac{e^{x_j + x_i}}{\left(\sum_l e^{x_l}\right)^2}$$

$$\frac{\partial S(x)_j}{\partial x_i} \underset{i = j}{=} \frac{e^{x_i}}{\sum_l e^{x_l}} - \frac{e^{x_j + x_i}}{\left(\sum_l e^{x_l}\right)^2}$$

$$J(S(x))_{ji} = \begin{cases} & i \neq j \\ & i = j \end{cases}$$

לכן

7. Let $f : \mathbb{R}^d \to \mathbb{R}$ be defined as $f(x,y) = x^3 - 5xy - y^5$. Calculate the Hessian of f.

$$H(f) = \begin{matrix} \frac{\partial f}{\partial x^2} & \frac{\partial f}{\partial x \partial y} \\ \frac{\partial f}{\partial y \partial x} & \frac{\partial f}{\partial y^2} \end{matrix}$$

$$\frac{\partial f}{\partial x} = 3x^2 - 5y \qquad \frac{\partial f}{\partial x \partial y} = \frac{\partial f}{\partial y \partial x} = -5$$

$$\frac{\partial f}{\partial y} = -5x - 5y^4 \qquad \frac{\partial f}{\partial x^2} = 6x$$

$$\frac{\partial f}{\partial y^2} = -20y^3$$

$$\boxed{H(f) = \begin{matrix} 6x & -5 \\ -5 & -20y^3 \end{matrix}}$$

1ɔJ

2.2 Estimation Theory

8. Let $x_1, x_2, \ldots \overset{iid}{\sim} \mathcal{P}$ be a sample of infinity size drawn from some probability distribution function $\mathcal{P}$ with finite expectation and variance. Show that the sample mean estimator $\hat{\mu}_n = \frac{1}{n}\sum x_i$ calculated over the first n samples is a consistent estimator. Hint: for any given fixed value of $n \in \mathbb{N}$ bound from above the probability of deviating more than ε.

$$\hat{\mu}_n = \frac{1}{n}\sum x_i$$

$$p(\hat{\mu}_n - \mu \geq \varepsilon) \leq \frac{1}{\sqrt{4n}}$$

הוכחנו בהרצאה אתד

לכן ברור כי ככל ש n גדל אנו הסיכוי
לתכיג ב ε קטן גם כן.

9. Let $\mathbf{x}_1, \ldots, \mathbf{x}_m \overset{iid}{\sim} \mathcal{N}(\mu, \Sigma)$ be m observations sampled i.i.d from a multivariate Gaussian with expectation of $\mu \in \mathbb{R}^d$ and a covariance matrix $\Sigma \in \mathbb{R}^{d \times d}$. Derive the log-likelihood function of $\mathcal{N}(\mu, \Sigma)$. Hint: follow the approach used to derive the likelihood function for the univariate case.

עבור $\bar{X}_i$ נתיד מתקיים ש:

$$f_{(\mu,\Sigma)}(x_i) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} \exp\left(-\frac{1}{2}(\bar{X}_i - \mu)^T \Sigma^{-1} (\bar{X}_i - \mu)\right)$$

לכן

$$L(x, \mu, \Sigma) = \prod_{i=1}^{m} f_{(\mu,\Sigma)}(\bar{X}_i)$$

נפעיל log ונקבל:

$$L(x,\mu,\Sigma)=\log\prod_{i=1}^{m} f_{(\mu,\Sigma)}(x_i)$$

$$=\sum_{i=1}^{m} -\frac{d}{2}\log(2\pi)-\frac{1}{2}\log(|\Sigma|)-\frac{1}{2}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)$$

$$\boxed{=-\frac{md}{2}\log(2\pi)-\frac{m}{2}\log(|\Sigma|)-\frac{1}{2}\sum_{i=1}^{m}(x_i-\mu)^T\Sigma^{-1}(x_i-\mu)}$$