Introduction to Machine Learning

Spring 2020

University of Science and Technology of China

Lecturer: Jie Wang — Homework 8

Posted: May. 24, 2020 — Due: May. 31, 2020

Name: San Zhang — ID: PBXXXXXXXX

Notice, to get the full credits, please present your solutions step by step.

Exercise 1: Basic Matrix Manipulations 10pts

For an arbitrary matrix M, we denote its i^{th} row, j^{th} column, and $(i,j)^{th}$ entry by $\mathbf{m}_{i,*}$, $\mathbf{m}_{*,j}$, and $m_{i,j}$, respectively.

1. Suppose that $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{m\times d}$, $C \in \mathbb{R}^{d\times n}$, and $A = BC$. Show that

$$A = \sum_{\ell=1}^{d} \mathbf{b}_{*,\ell}\mathbf{c}_{\ell,*}.$$

2. Suppose that $A \in \mathbb{R}^{m\times n}$, $B \in \mathbb{R}^{m\times p}$, $C \in \mathbb{R}^{p\times q}$, $D \in \mathbb{R}^{q\times n}$, and $A = BCD$. Show that

$$A = \sum_{i=1}^{p}\sum_{j=1}^{q} c_{i,j}\mathbf{b}_{*,i}\mathbf{d}_{j,*}.$$

Solution: ■

Exercise 2: Trace 30pts

For any matrixes $X \in \mathbb{R}^{n\times n}$, let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$ be n eigenvalues of the matrix X. Define $\lambda(X) = (\lambda_1, \lambda_2, \ldots, \lambda_n)^\top$ and

$$\operatorname{diag}(\lambda(X)) = \begin{bmatrix} \lambda_1 & 0 & \cdots & 0 \\ 0 & \lambda_2 & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \lambda_n \end{bmatrix} \in \mathbb{R}^{n\times n}.$$

Suppose that $A \in \mathbb{R}^{n\times n}$, $B \in \mathbb{R}^{n\times n}$, and $C \in \mathbb{R}^{n\times n}$.

1. Show that the trace of a matrix is a linear mapping, i.e.,

$$\operatorname{tr}(A+B) = \operatorname{tr}(A) + \operatorname{tr}(B)$$
$$\operatorname{tr}(cA) = c\operatorname{tr}(A),\ c \text{ is a constant.}$$

2. [**Cyclic property**] Show that the trace is invariant under cyclic permutations, i.e.,

$$\operatorname{tr}(ABC) = \operatorname{tr}(CAB) = \operatorname{tr}(BCA).$$

3. [**Derivatives of Traces**] The derivative of a scalar function $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ of a matrix $X \in \mathbb{R}^{m\times n}$ is defined by

$$\nabla_X f = \begin{bmatrix} \frac{\partial f}{\partial X_{11}} & \cdots & \frac{\partial f}{\partial X_{1j}} & \cdots & \frac{\partial f}{\partial X_{1n}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f}{\partial X_{i1}} & \cdots & \frac{\partial f}{\partial X_{ij}} & \cdots & \frac{\partial f}{\partial X_{in}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{\partial f}{\partial X_{m1}} & \cdots & \frac{\partial f}{\partial X_{mj}} & \cdots & \frac{\partial f}{\partial X_{mn}} \end{bmatrix},$$

where $X_{i,j}$ is the entry in the i^{th} row and j^{th} column of the matrix X. Please find the gradients of the following functions.

   (a) $f(X) = \operatorname{tr}(X)$, $X \in \mathbb{R}^{n\times n}$;

   (b) $f(X) = \operatorname{tr}(MX)$, $M \in \mathbb{R}^{n\times m}$, $X \in \mathbb{R}^{m\times n}$;

   (c) $f(X) = \operatorname{tr}(MXN)$, $M \in \mathbb{R}^{n\times m}$, $X \in \mathbb{R}^{m\times m}$, $N \in \mathbb{R}^{m\times n}$;

   (d) $f(X) = \operatorname{tr}(X^2)$, $X \in \mathbb{R}^{n\times n}$;

   (e) $f(X) = \operatorname{tr}(XMX^\top)$, $X \in \mathbb{R}^{n\times n}$, $M \in \mathbb{R}^{n\times n}$.

4. [**Fan's inequality**] Show the inequality

$$\operatorname{tr}(AB) \leq \lambda(A)^\top \lambda(B).$$

Moreover, please show that the equality holds if and only if there is an orthogonal matrix $U \in \mathbb{R}^{n\times n}$ such that $A = U\operatorname{diag}(\lambda(A))U^\top$ and $B = U\operatorname{diag}(\lambda(B))U^\top$. [Hint: Eigenvalue decomposition.]

Solution: ■

Exercise 3: SVD 30pts

Let $A \in \mathbb{R}^{m\times n}$, $\text{rank}(A) = r$. The SVD of A is $A = U\Sigma V^\top$, where we sort the diagonal entries of Σ in the descending order $\sigma_1 \geq \sigma_2 \geq \ldots \geq \sigma_r > 0$, and

$$U_1 = (\mathbf{u}_{*,1}, \mathbf{u}_{*,2}, \ldots, \mathbf{u}_{*,r}), U_2 = (\mathbf{u}_{*,r+1}, \ldots, \mathbf{u}_{*,m}),$$
$$V_1 = (\mathbf{v}_{*,1}, \mathbf{v}_{*,2}, \ldots, \mathbf{v}_{*,r}), V_2 = (\mathbf{v}_{*,r+1}, \ldots, \mathbf{v}_{*,n}).$$

The column space of A is the set

$$\mathcal{C}(A) = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{y} = A\mathbf{x}, \mathbf{x} \in \mathbb{R}^n\}.$$

The null space of A is the set

$$\mathcal{N}(A) = \{\mathbf{y} \in \mathbb{R}^n : A\mathbf{y} = 0\}.$$

1. Show that
   (a) $P_{\mathcal{C}(A)}(\mathbf{x}) = U_1 U_1^\top \mathbf{x}$;
   (b) $P_{\mathcal{N}(A)}(\mathbf{x}) = V_2 V_2^\top \mathbf{x}$;
   (c) $P_{\mathcal{C}(A^\top)}(\mathbf{x}) = V_1 V_1^\top \mathbf{x}$;
   (d) $P_{\mathcal{N}(A^\top)}(\mathbf{x}) = U_2 U_2^\top \mathbf{x}$.
2. The Frobenius norm of A is

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{i,j}^2}.$$

   (a) Show that $\|A\|_F^2 = \text{tr}(A^\top A)$.
   (b) Let $B \in \mathbb{R}^{m\times n}$. Suppose that $\mathcal{C}(A) \perp \mathcal{C}(B)$, that is,

$$\langle \mathbf{a}, \mathbf{b} \rangle = 0, \forall\, \mathbf{a} \in \mathcal{C}(A),\ \mathbf{b} \in \mathcal{C}(B).$$

   Show that

$$\|A + B\|_F^2 = \|A\|_F^2 + \|B\|_F^2.$$

3. Given $K < r$, $K \in \mathbb{N}$, please solve the problem as follows.

$$\min_{X \in \mathbb{R}^{m\times n}} \{\|A - X\|_F : \text{rank}(X) \leq K\}.$$

   For simplicity, you can assume that all singular values of A are different.
4. **Programming Exercise** We provide you a grayscale image ("Alan_Turing.jpg"). Suppose that A is the data matrix of the image. We have $A \in \mathbb{R}^{512\times 512}$ and $r = \text{rank}(A) = 512$. In this exercise, you are expected to implement an image compression algorithm following the steps below. You can use your favorite programming language.

(a) Compute the SVD $A = U\Sigma V^\top = \sum_{i=1}^r \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$, where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_r > 0$ are the diagonal entries of Σ, $\mathbf{u}_i$ is the ith column of U, and $\mathbf{v}_i$ is the ith column of V.

(b) Use the first k ($k < r$) terms of SVD to approximate the original image A. Then, we get the compressed images, of which the data matrices are $A_k = \sum_{i=1}^k \sigma_i \mathbf{u}_i \mathbf{v}_i^\top$. Compute A_k for $k = 2, 4, 8, 16, 32, 64, 128, 256$.

(c) Plot A_k as images for all k.

Please put the compressed images and their corresponding k in this file.

Solution: ■

Exercise 4: PCA 30pts

Suppose that we have a set of data instances $\{\mathbf{x}_i\}_{i=1}^n \subset \mathbb{R}^d$. Let $\widetilde{X} \in \mathbb{R}^{d\times n}$ be the matrix whose i^{th} column is $\mathbf{x}_i - \bar{\mathbf{x}}$, where $\bar{\mathbf{x}}$ is the sample mean, and S be the sample variance matrix.

1. For $G \in \mathbb{R}^{d\times K}$, let us define

$$f(G) = \mathrm{tr}(G^\top S G). \tag{1}$$

   Show that $f(GQ) = f(G)$ for any orthogonal matrix $Q \in \mathbb{R}^{K\times K}$.

2. Please find $\mathbf{g}_1$ defined as follows by the Lagrange multiplier method.

$$\mathbf{g}_1 := \underset{\mathbf{g}\in\mathbb{R}^d}{\mathbf{argmax}}\{f(\mathbf{g}) : \|\mathbf{g}\|_2 = 1\}, \tag{2}$$

   where f is defined by (1). Notice that, the vector $\mathbf{g}_1$ is the first principal component vector of the data.

3. Please find $\mathbf{g}_2$ defined as follows by the Lagrange multiplier method.

$$\mathbf{g}_2 := \underset{\mathbf{g}\in\mathbb{R}^d}{\mathbf{argmax}}\{f(\mathbf{g}) : \|\mathbf{g}\|_2 = 1, \langle \mathbf{g}, \mathbf{g}_1\rangle = 0\},$$

   where $\mathbf{g}_1$ is given by (2). Similar to $\mathbf{g}_1$, the vector $\mathbf{g}_2$ is the second principal component vector of the data.

4. Please derive the first K principal component vectors by repeating the above process.

5. What is $f(\mathbf{g}_k)$, $k = 1, \ldots, K$? What about their meaning?

6. When are the first K principal component vectors unique?

Solution: ■