Introduction to Machine Learning

Spring 2021

University of Science and Technology of China

Lecturer: Jie Wang | Homework 6

Posted: May 29, 2021 | Due: June 11, 2021

Name: San Zhang | ID: PBXXXXXXXX

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

Exercise 1: Softmax and Cross Entropy 20pts

The softmax function $\mathbf{f}: \mathbb{R}^n \to \mathbb{R}^n$ is defined by:

$$f_i(\mathbf{x}) = \frac{\exp(x_i)}{\sum_{k=1}^{n} \exp(x_k)}, i = 1, \ldots, n,$$

where x_i is the i^{th} component of $\mathbf{x} \in \mathbb{R}^n$. The function $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), f_2(\mathbf{x}), \ldots, f_n(\mathbf{x}))^\top$ converts each input $\mathbf{x}$ into a probability (stochastic) vector in which all entries are nonnegative and add up to one.

1. Please find the gradient and Jacobian matrix of $\mathbf{f}(\mathbf{x})$, i.e., $\nabla \mathbf{f}(\mathbf{x})$ and $D\mathbf{f}(\mathbf{x})$.
2. Show that $\mathbf{f}(\mathbf{x}) = \mathbf{f}(\mathbf{x} - c\mathbf{1})$, where $c = \max\{x_1, x_2, ..., x_n\}$ and $\mathbf{1}$ is a vector all of whose components are one. When do we need this transformation?
3. Please find the gradient of cross entropy function:

$$g(\mathbf{x}) = -\sum_{i=1}^{n} H_i \log(f_i(\mathbf{x})),$$

where $\mathbf{H} = (H_1, H_2, \ldots, H_n)^\top \in \mathbb{R}^n$ is a one-hot vector.

Solution: ■

Exercise 2: Convolutional Neural Network 25pts

1. The average pooling in convolutional neural network can be formulated as

$$f_1(\mathbf{x}) = \frac{\sum_{i=1}^{n} x_i}{n},$$

where x_i is the i^{th} component of $\mathbf{x} \in \mathbb{R}^n$. Please derive the gradient of $f_1(\mathbf{x})$.

2. The max pooling in convolutional neural network can be formulated as

$$f_2(\mathbf{x}) = \max\{x_1, \ldots, x_n\},$$

where x_i is the i^{th} component of $\mathbf{x} \in \mathbb{R}^n$.

   (a) Find the set containing all differentiable points of f_2.

   (b) We call $\mathbf{d}(\mathbf{x})$ is a subgradient at $\mathbf{x}$ of f_2 if

$$f_2(\mathbf{y}) \geq f_2(\mathbf{x}) + \langle \mathbf{d}(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle, \forall \mathbf{x}, \mathbf{y}.$$

   Find a subgradient $\mathbf{d}(\mathbf{x})$ of f_2 at $\mathbf{x}$.

3. Suppose that we have a convolutional neural network as shown in Table 1.

   (a) The convolutional layer parameters are denoted as "conv⟨filter size⟩-⟨number of filters⟩".

   (b) The fully connected layer parameters are denoted as "FC⟨number of neurons⟩".

   (c) The window size of pooling layers is 2.

   (d) The stride of convolutinal layers is 1.

   (e) The stride of pooling layers is 2.

   (f) You may want to use padding in both convolutional and pooling layers if necessary.

   (g) For convenience, we assume that there is no activation function and bias.

   Suppose that the input is a **210 × 160 RGB** image. Please derive the size of all feature maps and the number of parameters.

conv3-32	conv5-32	max pool	conv3-64	conv5-64	max pool	FC-128	FC-10

Table 1: The architecture of convolutional neural network

Solution: ■

Exercise 3: Matrix Calculus 20pts

Let $L = f(\mathbf{h}(\mathbf{A}\mathbf{x}+\mathbf{b}))$, where $\mathbf{A} \in \mathbb{R}^{m\times n}$, $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{b} \in \mathbb{R}^m$, and $f : \mathbb{R}^m \to \mathbb{R}$. Define $\mathbf{z} = \mathbf{A}\mathbf{x}+\mathbf{b} \in \mathbb{R}^m$ and $\mathbf{w} = \mathbf{h}(\mathbf{z}) = (\sigma(z_1), \ldots, \sigma(z_m))^\top$, where z_i is the i^{th} component of $\mathbf{z}$ and

$$\sigma(z_i) = \frac{1}{1+\exp(-z_i)}.$$

Assume $\nabla_{\mathbf{w}} f$ is known.

1. Please derive $\nabla_{\mathbf{x}} L$.
2. Please derive

$$\nabla_{\mathbf{A}} L = \begin{bmatrix} \frac{\partial L}{\partial A_{11}} & \cdots & \frac{\partial L}{\partial A_{1j}} & \cdots & \frac{\partial L}{\partial A_{1n}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{\partial L}{\partial A_{i1}} & \cdots & \frac{\partial L}{\partial A_{ij}} & \cdots & \frac{\partial L}{\partial A_{in}} \\ \cdots & \cdots & \cdots & \cdots & \cdots \\ \frac{\partial L}{\partial A_{m1}} & \cdots & \frac{\partial L}{\partial A_{mj}} & \cdots & \frac{\partial L}{\partial A_{mn}} \end{bmatrix},$$

where $A_{i,j}$ is the entry in the i^{th} row, j^{th} column of the matrix $\mathbf{A}$.

Exercise 4: 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 5: Trace 30pts

For any matrices $X \in \mathbb{R}^{n\times n}$, 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

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

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.,

$$\begin{aligned} \mathrm{tr}(A+B) &= \mathrm{tr}(A) + \mathrm{tr}(B) \\ \mathrm{tr}(cA) &= c\,\mathrm{tr}(A),\ c \text{ is a constant.} \end{aligned}$$

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

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

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) $f(X) = \mathrm{tr}(X)$, $X \in \mathbb{R}^{n\times n}$;
   (b) $f(X) = \mathrm{tr}(MX)$, $M \in \mathbb{R}^{n\times m}$, $X \in \mathbb{R}^{m\times n}$;
   (c) $f(X) = \mathrm{tr}(MXN)$, $M \in \mathbb{R}^{n\times m}$, $X \in \mathbb{R}^{m\times m}$, $N \in \mathbb{R}^{m\times n}$;
   (d) $f(X) = \mathrm{tr}(X^2)$, $X \in \mathbb{R}^{n\times n}$;
   (e) $f(X) = \mathrm{tr}(XMX^\top)$, $X \in \mathbb{R}^{n\times n}$, $M \in \mathbb{R}^{n\times n}$.

Solution: ■

Exercise 6: SVD 30pts

Let $A \in \mathbb{R}^{m\times n}$, $\mathbf{rank}(A) = r$. 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) $P_{\mathcal{C}(A)}(\mathbf{x}) = U_1U_1^\top\mathbf{x}$;
   (b) $P_{\mathcal{N}(A)}(\mathbf{x}) = V_2V_2^\top\mathbf{x}$;
   (c) $P_{\mathcal{C}(A^\top)}(\mathbf{x}) = V_1V_1^\top\mathbf{x}$;
   (d) $P_{\mathcal{N}(A^\top)}(\mathbf{x}) = U_2U_2^\top\mathbf{x}$.
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 $\|A\|_F^2 = \mathrm{tr}(A^\top A)$.
   (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.
$$\min_{X\in\mathbb{R}^{m\times n}} \{\|A - X\|_F : \mathbf{rank}(()X) \leq K\}.$$
   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 $r = \mathbf{rank}(A) = 512$. 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 7: 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

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

   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: ■