Introduction to Machine Learning

Spring 2020

University of Science and Technology of China

Lecturer: Jie Wang | Homework 6

Posted: May. 6, 2020 | Due: May. 14, 2020

Name: San Zhang | ID: PBXXXXXXXX

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

Exercise 1: Decision Tree 5pts

Please build a decision tree based on the information gain to classify the following dataset (you need to show the calculation steps in detail).

Sample	A_1	A_2	A_3	Response
x_1	0	0	0	0
x_2	1	0	1	0
x_3	0	1	0	0
x_4	0	1	1	1
x_5	1	1	0	1

Table 1: Dataset

The dataset consists of five samples x_1, x_2, x_3, x_4, x_5. For each sample, we can observe the features A_1, A_2, A_3 and the corresponding response.

Solution: ■

Exercise 2: 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 3: 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 2.

   (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.

conv3-64	max pool	conv3-256	conv1-512	max pool	FC-1000	FC-18

Table 2: The architecture of convolutional neural network

Solution: ■

Exercise 4: 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 $\mathbf{A}$.

Solution: ■

Exercise 5: Tail Probabilities 30pts

Let X be a random variable on $\mathbb{R}$. You can assume that X is a continuous random variable.

1. **(Markov's inequality)** For all $t > 0$, show that
$$\mathbf{P}(|X| \geq t) \leq \frac{\mathbf{E}[|X|]}{t}.$$

2. **(Chebyshev's inequality)** For all $t > 0$, show that
$$\mathbf{P}(|X - \mathbb{E}[X]| \geq t) \leq \frac{\mathbb{V}[X]}{t^2}.$$

3. A random variable X on $\mathbb{R}$ is σ-subgaussian if for all $\lambda \in \mathbb{R}$, it holds that $\mathbb{E}[\exp(\lambda X)] \leq \exp(\lambda^2\sigma^2/2)$.

   Assume that $\mathbb{E}[X] = 0$ and $X \in [a, b]$.

   (a) Show that X is $\frac{b-a}{2}$-subgaussian.

   (b) **(Hoeffding's inequality)** Show that
$$\mathbf{P}(X \geq t) \leq \exp\left(-\frac{2t^2}{(b-a)^2}\right)$$

Solution: ■

Exercise 6: Learning Intervals 20pts

1. Show that $(1-\epsilon)^m \leq e^{-m\epsilon}$, where $m \in \mathbb{N}^+$ and $0 \leq \epsilon < 1$.

2. Let the target concept class be $C = \{[a,b] : a < b, a, b \in \mathbb{R}\}$ and the hypotheses class $H = C$, and the version space be $VS_{H,D}$. Each $c \in C$ labels the points inside the interval positive and the others negative. A consistent learner will pick a consistent hypothesis—if any—$h \in H$ according to a set of i.i.d. samples $\{(x_i, c(x_i))\}_{i=1}^m$ that obey an unknown absolute continuous distribution $\mathcal{D}$. $\mathcal{D}$'s p.d.f. is $p(x)$. Please find an upper bound of
$$\mathbf{P}[\exists\, h \in VS_{H,D} \text{ and } error_{\mathcal{D}}(h) > \epsilon],$$
and the corresponding sample complexity.

Solution: ■

Exercise 7: VC-dimension 10pts

Given the instance space $X = \mathbb{R}^2$, the hypothesis space H is the set of all linear threshold functions defined on $\mathbb{R}^2$. Find $VC(H)$ and prove it.

Solution: ■