Introduction to Machine Learning
Fall 2021
University of Science and Technology of China

Lecturer: Jie Wang | Homework 2
Posted: Oct. 1, 2021 | Due: Oct. 18, 2021

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

Exercise 1: Conditional Expectations

Recall that, for supervised learning problems, each data instance consists of a D-dimensional input feature vector $X \in \mathbb{R}^D$ and the corresponding output $Y \in \mathbb{R}$. We would like to find a mapping $f(X)$ to estimate the value of Y given a sample of X. Let

$$\ell(y, f(\mathbf{x})) = (f(\mathbf{x}) - y)^2$$

be the square loss. We choose the function $f(X)$ by minimizing the expectation of the square loss:

$$J[f] := \mathbb{E}[\ell(Y, f(X)] = \iint (y - f(\mathbf{x}))^2 p(\mathbf{x}, y)\mathrm{d}\mathbf{x}\mathrm{d}y,$$

where $p(\mathbf{x}, y)$ is the joint PDF.

1. Let h be a function of X and $\epsilon > 0$. Please calculate $J[f + \epsilon h] - J[f]$.
2. Prove that $J[f + \epsilon h] - J[f] \geq 0$ for any $\epsilon > 0$ if and only if
$$\int h(\mathbf{x}) \left\{ \int -2\left(y - f(\mathbf{x})\right) p(\mathbf{x}, y)\mathrm{d}y \right\} \mathrm{d}\mathbf{x} \geq 0.$$
3. Please show that $f^*(\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}]$ is a solution to
$$J[f^*] = \min_f \{J[f]\}.$$
4. Please deduce that
$$\mathbb{E}[\ell(Y, f(X))] = \int \{f(\mathbf{x}) - \mathbb{E}[y|\mathbf{x}]\}^2 p(\mathbf{x})\mathrm{d}\mathbf{x} + \iint \{\mathbb{E}[y|\mathbf{x}] - y\}^2 p(\mathbf{x}, y)\mathrm{d}\mathbf{x}\mathrm{d}y.$$

Solution: ■

Exercise 2: Bias-Variance Trade-off (Programming Exercise. You are required to finish at least one of Exercises 2 and 3.)

We provide you with $L = 100$ data sets, each having $N = 25$ points:

$$\mathcal{D}^{(l)} = \{(x_n, y_n^{(l)})\}_{n=1}^{N}, \quad l = 1, 2, \cdots, L,$$

where x_n are uniformly taken from $[-1, 1]$, and all points $(x_n, y_n^{(l)})$ are independently from the sinusoidal curve $h(x) = \sin(\pi x)$ with an additional disturbance.

1. For each data set $\mathcal{D}^{(l)}$, consider fitting a model with 24 Gaussian basis functions

   $$\phi_j(x) = e^{-(x-\mu_j)^2}, \quad \mu_j = 0.2 \cdot (j - 12.5), \quad j = 1, \cdots 24$$

   by minimizing the regularized error function

   $$L^{(l)}(\mathbf{w}) = \frac{1}{2}\sum_{n=1}^{N}(y_n^{(l)} - \mathbf{w}^\top \boldsymbol{\phi}(x_n))^2 + \frac{\lambda}{2}\mathbf{w}^\top \mathbf{w},$$

   where $\mathbf{w} \in \mathbb{R}^{25}$ is the parameter, $\boldsymbol{\phi}(x) = (1, \phi_1(x), \cdots, \phi_{24}(x))^\top$ and λ is the regular coefficient. What's the closed form of the parameter estimator $\hat{\mathbf{w}}^{(l)}$ for the data set $\mathcal{D}^{(l)}$?

2. For $\log_{10}\lambda = -10, -5, -1, 1$, plot the prediction functions $y^{(l)}(x) = f_{\mathcal{D}^{(l)}}(x)$ on $[-1, 1]$ respectively. For clarity, show only the first 25 fits in the figure for each λ.

3. For $\log_{10}\lambda \in [-3, 1]$, calculate the followings:

   $$\bar{y}(x) = \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(x)] = \frac{1}{L}\sum_{l=1}^{L} y^{(l)}(x)$$

   $$(\text{bias})^2 = \mathbb{E}_X[(\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(X)] - h(X))^2] = \frac{1}{N}\sum_{n=1}^{N}(\bar{y}(x_n) - h(x_n))^2$$

   $$\text{variance} = \mathbb{E}_X[\mathbb{E}_{\mathcal{D}}[(f_{\mathcal{D}}(\mathbf{x}) - \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D}}(\mathbf{x})])^2]] = \frac{1}{N}\sum_{n=1}^{N}\frac{1}{L}\sum_{l=1}^{L}(y^{(l)}(x_n) - \bar{y}(x_n))^2$$

   Plot the three quantities, $(\text{bias})^2$, variance and $(\text{bias})^2 + \text{variance}$ in one figure, as the functions of $\log_{10}\lambda$. (**Hint:** see [1] for an example.)

Solution: ■

Exercise 3: Bayesian Linear Regression (Programming Exercise. You are required to finish at least one of Exercises 2 and 3.)

Consider a single input variable $\mathbf{x}$, a single output variable $\mathbf{y}$ and a linear model of the form $\mathbf{y} = w_0 + w_1\mathbf{x} + \epsilon$, where ϵ is Gaussian distributed with mean of 0 and standard deviation of 0.25.

1. Suppose that, the model parameter $\mathbf{w} = (w_0, w_1)^T \in \mathbb{R}^2$ has a Gaussian prior of the form

$$p(\mathbf{w}) = \mathcal{N}(\mathbf{w}|\mu_0, \Sigma_0) = \frac{1}{2\pi}\frac{1}{|\Sigma_0|^{1/2}} \exp\{-\frac{1}{2}(\mathbf{w} - \mu_0)^T\Sigma_0^{-1}(\mathbf{w} - \mu_0)\}$$

   where $\mu_0 = \mathbf{0}$ and $\Sigma_0 = \frac{1}{2}\mathbf{I}$. Please plot this Gaussian distribution in the form of heat map.

2. Sample six times independently from the prior Gaussian distribution defined above. Please plot the six straight lines $y = w_0 + w_1 x$ using these samples.

3. Now, suppose that we have observed a single data point $(x_1, y_1) = (0.6, 0)$. Please plot the likelihood function $p(y_1|x_1, \mathbf{w})$ for this data point as the function of $\mathbf{w}$, still in the form of heat map.

4. Calculate the posterior distribution of $\mathbf{w}$, denoted by $p(\mathbf{w}|y_1, x_1)$. Please plot the posterior distribution.

5. Sample six times independently from this posterior distribution of $\mathbf{w}$ and plot the six straight lines $y = w_0 + w_1 x$.

6. Then, suppose we observe a new single data point $(x_2, y_2) = (-0.5, 0.6)$. Please plot the corresponding likelihood function $p(y_2|x_2, \mathbf{w})$ of this second point alone, the posterior distribution of $\mathbf{w}$, denote by $p(\mathbf{w}|y_1, y_2, x_1, x_2)$, and six samples drawn from the current posterior function.

7. If we can observe new data points continuously, and then observe the posterior distributions and their sampled linear regression models sequentially. What will you infer from them? Please write down your conclusions.

(**Hint:** see [1] for an example.)

Solution: ■

Exercise 4: Covariance Matrix and Gaussian Distribution

Let $\mathbf{X} = (X_1, X_2, \cdots, X_D)^{\mathrm{T}} \in \mathbb{R}^D$ be a D-dimensional random vector. The covariance matrix of $\mathbf{X}$, denoted by $\Sigma_{\mathbf{X}}$, is defined as

$$\mathrm{Cov}[\mathbf{X}] = \mathbb{E}\big[(\mathbf{X} - \mathbb{E}[\mathbf{X}])(\mathbf{X} - \mathbb{E}[\mathbf{X}])^{\mathrm{T}}\big].$$

1. Please show that $\Sigma_{\mathbf{X}}$ is positive semi-definite.

2. Please show that $\Sigma_{\mathbf{X}}$ doesn't have full rank if and only if $\{X_i - \mathbb{E}[X_i]\}_{i=1}^{D}$ are linearly dependent.

3. Suppose that, the random vector $\mathbf{X}$ has a multivariate Gaussian distribution with the mean vector being $\boldsymbol{\mu}$ and the covariance matrix being $\boldsymbol{\Sigma}$, respectively. The probability density function of $\mathbf{X}$ is

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\top} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\},$$

where $\mathbf{x}$ is a realization of the random vector $\mathbf{X}$. For notational simplicity, let

$$c = (2\pi)^{D/2} |\boldsymbol{\Sigma}|^{1/2}. \tag{1}$$

Clearly, we must have

$$\int_{\mathbb{R}^D} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\top} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\} d\mathbf{x} = c. \tag{2}$$

Now, let us denote the first M components of $\mathbf{X}$ by $\mathbf{X}_a$, and the remaining $D - M$ ones by $\mathbf{X}_b$, so that

$$\mathbf{X} = \begin{pmatrix} \mathbf{X}_a \\ \mathbf{X}_b \end{pmatrix}.$$

We denote the corresponding partitions of the mean vector by

$$\boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix}$$

and the covariance matrix by

$$\boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}.$$

Please show that $\mathbf{X}_a$ has a Gaussian distribution with its mean vector being $\boldsymbol{\mu}_a$ and the covariance matrix being $\boldsymbol{\Sigma}_{aa}$. In other words, please show that

$$p(\mathbf{x}_a) = \int p(\mathbf{x}_a, \mathbf{x}_b) d\mathbf{x}_b = \frac{1}{(2\pi)^{M/2}} \frac{1}{|\boldsymbol{\Sigma}_{aa}|^{1/2}} \exp\left\{ -\frac{1}{2}(\mathbf{x}_a - \boldsymbol{\mu}_a)^\top \boldsymbol{\Sigma}_{aa}^{-1} (\mathbf{x}_a - \boldsymbol{\mu}_a) \right\}.$$

(**Hint:**

(a) you can make use identities similar to (1) and (2) to integrate out $\mathbf{x}_b$.

(b) you may find the following identity useful:

$$|\boldsymbol{\Sigma}| = |\boldsymbol{\Sigma}_{aa}||\boldsymbol{\Sigma}_{bb} - \boldsymbol{\Sigma}_{ba}\boldsymbol{\Sigma}_{aa}^{-1}\boldsymbol{\Sigma}_{ab}| = |\boldsymbol{\Sigma}_{bb}||\boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}|.)$$

Solution: ■

Exercise 5: Limit and Limit Points (Optional)

1. Show that $\{\mathbf{x}_n\}$ in $\mathbb{R}^n$ converges to $\mathbf{x} \in \mathbb{R}^n$ if and only if $\{\mathbf{x}_n\}$ is bounded and has a unique limit point $\mathbf{x}$.

2. (**Limit Points of a Set**). Let C be a subset of $\mathbb{R}^n$. A point $\mathbf{x} \in \mathbb{R}^n$ is called a limit point of C if there is a sequence $\{\mathbf{x}_n\}$ in C such that $\mathbf{x}_n \to \mathbf{x}$ and $\mathbf{x}_n \neq \mathbf{x}$ for all positive integers n. If $\mathbf{x} \in C$ and $\mathbf{x}$ is not a limit point of C, then $\mathbf{x}$ is called an isolated point of C. Let C' be the set of limit points of the set C. Please show the following statements.

   (a) If $C = (0,1) \cup \{2\} \subset \mathbb{R}$, then $C' = [0,1]$ and $x = 2$ is an isolated point of C.

   (b) The set C' is closed.

   (c) The closure of C is the union of C' and C; that is $\mathbf{cl}\, C = C' \cup C$. Moreover, $C' \subset C$ if and only if C is closed.

Solution: ■

Exercise 6: Open and Closed Sets (Optional)

The norm ball $\{\mathbf{y} \in \mathbb{R}^n : \|\mathbf{y} - \mathbf{x}\|_2 < r, \mathbf{x} \in \mathbb{R}^n\}$ is denoted by $B_r(\mathbf{x})$.

1. Given a set $C \subset \mathbb{R}^n$, please show the following are equivalent.
   (a) The set C is closed; that is $\mathbf{cl}\, C = C$.
   (b) The complement of C is open.
   (c) If $B_\epsilon(\mathbf{x}) \cap C \neq \emptyset$ for every $\epsilon > 0$, then $\mathbf{x} \in C$.

2. Given $A \subset \mathbb{R}^n$, a set $C \subset A$ is called open in A if

$$C = \{\mathbf{x} \in C : B_\epsilon(\mathbf{x}) \cap A \subset C \,\text{for some}\, \epsilon > 0\}.$$

   A set C is said to be closed in A if $A \setminus C$ is open in A.
   (a) Let $B = [0, 1] \cup \{2\}$. Please show that $[0, 1]$ is not an open set in $\mathbb{R}$, while it is both open and closed in B.
   (b) Please show that a set $C \subset A$ is open in A if and only if $C = A \cap U$, where U is open in $\mathbb{R}^n$.

Solution: ■

Exercise 7: Bolzano-Weierstrass Theorem

The Least Upper Bound Axiom

Any nonempty set of real numbers with an upper bound has a least upper bound. That is, $\sup C$ *always exists for a nonempty bounded above set* $C \subset \mathbb{R}$.

Please show the following statements from **the least upper bound axiom**.

1. Let C be a nonempty subset of $\mathbb{R}$ that is bounded above. Prove that $u = \sup C$ if and only if u is an upper bound of C and

$$\forall \, \epsilon > 0, \exists \, a \in C \text{ such that } a > u - \epsilon.$$

2. Every bounded sequence in $\mathbb{R}$ has at least one limit point.

3. Every bounded sequence in $\mathbb{R}^n$ has at least one limit point.

Solution: ■

Exercise 8: Extreme Value Theorem

1. Show that a set $C \subset \mathbb{R}^n$ is compact if and only if C is closed and bounded.
2. Let C be a compact subset of $\mathbb{R}^n$ and $f : C \to \mathbb{R}$ be continuous. Please show that there exist $\mathbf{a}, \mathbf{b} \in C$ such that

$$f(\mathbf{a}) \leq f(\mathbf{x}) \leq f(\mathbf{b}), \forall \mathbf{x} \in C.$$

   (**Hint:** first prove that $f(C)$ is compact, in $\mathbb{R}$.)

Solution: ■

References

[1] C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, 2006.