Introduction to Machine Learning

Fall 2022

University of Science and Technology of China

Lecturer: Jie Wang — Homework 2

Posted: Oct. 11, 2022 — Due: Oct. 25, 2022

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

Exercise 1: Linear regression

Consider a data set $\{(x_i, y_i)\}_{i=1}^n$, where $x_i, y_i \in \mathbb{R}$.

1. If we want to fit the data by a linear model

$$y = w_0 + w_1 x, \tag{1}$$

please find $\hat{w}_0$ and $\hat{w}_1$ by the least squares approach (you need to find expressions of $\hat{w}_0$ and $\hat{w}_1$ by $\{(x_i, y_i)\}_{i=1}^n$, respectively).

2. **Programming Exercise** We provide you a data set $\{(x_i, y_i)\}_{i=1}^{30}$. Consider the model in (1) and the one as follows:

$$y = w_0 + w_1 x + w_2 x^2. \tag{2}$$

Which model do you think fits better the data? Please detail your approach first and then implement it by your favorite programming language. The required output includes

   (a) your detailed approach step by step;

   (b) your code with detailed comments according to your planned approach;

   (c) a plot showing the data and the fitting models;

   (d) the model you finally choose [$\hat{w}_0$ and $\hat{w}_1$ if you choose the model in (1), or $\hat{w}_0$, $\hat{w}_1$, and $\hat{w}_2$ if you choose the model in (2)].

Solution: ■

Exercise 2: Projection

Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{x} \in \mathbb{R}^m$. Define

$$\mathbf{P}_{\mathbf{A}}(\mathbf{x}) = \underset{\mathbf{z} \in \mathbb{R}^m}{\mathbf{argmin}} \{\|\mathbf{x} - \mathbf{z}\|_2 : \mathbf{z} \in \mathcal{C}(\mathbf{A})\}.$$

We call $\mathbf{P}_{\mathbf{A}}(\mathbf{x})$ the projection of the point $\mathbf{x}$ onto the column space of $\mathbf{A}$.

1. Please prove that $\mathbf{P}_{\mathbf{A}}(\mathbf{x})$ is unique for any $\mathbf{x} \in \mathbb{R}^m$.
2. Let $\mathbf{v}_i \in \mathbb{R}^n$, $i = 1, \ldots, d$ with $d \leq n$, which are linearly independent.
   (a) For any $\mathbf{w} \in \mathbb{R}^n$, please find $\mathbf{P}_{\mathbf{v}_1}(\mathbf{w})$, which is the projection of $\mathbf{w}$ onto the subspace spanned by $\mathbf{v}_1$.
   (b) Please show $\mathbf{P}_{\mathbf{v}_1}(\cdot)$ is a linear map, i.e.,
   $$\mathbf{P}_{\mathbf{v}_1}(\alpha\mathbf{u} + \beta\mathbf{w}) = \alpha\mathbf{P}_{\mathbf{v}_1}(\mathbf{u}) + \beta\mathbf{P}_{\mathbf{v}_1}(\mathbf{w}),$$
   where $\alpha, \beta \in \mathbb{R}$ and $\mathbf{w} \in \mathbb{R}^n$.
   (c) Please find the projection matrix corresponding to the linear map $\mathbf{P}_{\mathbf{v}_1}(\cdot)$, i.e., find the matrix $\mathbf{H}_1 \in \mathbb{R}^{n \times n}$ such that
   $$\mathbf{P}_{\mathbf{v}_1}(\mathbf{w}) = \mathbf{H}_1\mathbf{w}.$$
   (d) Let $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_d)$.
      i. For any $\mathbf{w} \in \mathbb{R}^n$, please find $\mathbf{P}_{\mathbf{V}}(\mathbf{w})$ and the corresponding projection matrix $\mathbf{H}$.
      ii. Please find $\mathbf{H}$ if we further assume that $\mathbf{v}_i^\top \mathbf{v}_j = 0$, $\forall\, i \neq j$.
3. (a) Suppose that
   $$\mathbf{A} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}.$$
   What are the coordinates of $\mathbf{P}_{\mathbf{A}}(\mathbf{x})$ with respect to the column vectors in $\mathbf{A}$ for any $\mathbf{x} \in \mathbb{R}^2$? Are the coordinates unique?
   (b) Suppose that
   $$\mathbf{A} = \begin{bmatrix} 1 & 2 \\ 1 & 2 \end{bmatrix}.$$
   What are the coordinates of $\mathbf{P}_{\mathbf{A}}(\mathbf{x})$ with respect to the column vectors in $\mathbf{A}$ for any $\mathbf{x} \in \mathbb{R}^2$? Are the coordinates unique?
4. A matrix $\mathbf{P}$ is called a projection matrix if $\mathbf{Px}$ is the projection of $\mathbf{x}$ onto $\mathcal{C}(\mathbf{P})$ for any $\mathbf{x}$.

(a) Let λ be the eigenvalue of $\mathbf{P}$. Show that λ is either 1 or 0. (*Hint: you may want to figure out what the eigenspaces corresponding to* $\lambda = 1$ *and* $\lambda = 0$ *are, respectively.*)

(b) Show that $\mathbf{P}$ is a projection matrix if and only if $\mathbf{P}^2 = \mathbf{P}$ and $\mathbf{P}$ is symmetric.

5. Let $\mathbf{B} \in \mathbb{R}^{m \times s}$ and $\mathcal{C}(\mathbf{B})$ be its column space. Suppose that $\mathcal{C}(\mathbf{B})$ is a proper subspace of $\mathcal{C}(\mathbf{A})$. Is $\mathbf{P_B}(\mathbf{x})$ the same as $\mathbf{P_B}(\mathbf{P_A}(\mathbf{x}))$? Please show your claim rigorously.

Solution: ■

Exercise 3: Linear regression by maximum likelihood (optional)

Suppose that the samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are i.i.d., where $\mathbf{x}_i = (x_{i,1}, \ldots, x_{i,d})^\top \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. For any $i \in \{1, \ldots, n\}$, we assume that

$$y_i = w_0 + w_1 x_{i,1} + \cdots + w_d x_{i,d} + \epsilon_i,$$

where $\mathbf{w} = (w_0, w_1, \ldots, w_d)^\top \in \mathbb{R}^{d+1}$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. For simplicity, we define $\bar{\mathbf{x}}_i = (1, x_{i,1}, \ldots, x_{i,d})^\top$, $\mathbf{X} = (\bar{\mathbf{x}}_1, \ldots, \bar{\mathbf{x}}_n)^\top$, and $\mathbf{y} = (y_1, \ldots, y_n)^\top$, where $\mathbf{X}$ has full rank.

1. Please find the maximum likelihood estimation (MLE) $\hat{\mathbf{w}}$ of the weights $\mathbf{w}$. Specifically, please give the expression of $\hat{w}_0$.

2. Please find the MLE of σ.

Solution: ■

Exercise 4: Multicollinearity

Consider the linear regression problem formulated as below:

$$\mathbf{y} = \mathbf{X}\mathbf{w} + \mathbf{e}, \mathbb{E}(\mathbf{e}) = \mathbf{0}, \mathrm{Cov}(\mathbf{e}) = \sigma^2\mathbf{I_n},$$

where $\mathbf{y} = (y_1, \ldots, y_n)^\top$ and $\mathbf{X} \in \mathbb{R}^{n\times p}$. Suppose that $\mathbf{X}^\top\mathbf{X}$ is invertible, then $\hat{\mathbf{w}} = \left(\mathbf{X}^\top\mathbf{X}\right)^{-1}\mathbf{X}^\top\mathbf{y}$ is the least squares estimator of $\mathbf{w}$.

1. Recall that the covariance matrix of p-dimensional random vectors is defined as

$$\mathrm{Cov}(\hat{\mathbf{w}}) = \mathbb{E}[(\hat{\mathbf{w}} - \mathbb{E}(\hat{\mathbf{w}}))(\hat{\mathbf{w}} - \mathbb{E}(\hat{\mathbf{w}}))^\top].$$

   Please show that

   (a) $\mathbb{E}(\hat{\mathbf{w}}) = \mathbf{w}$;

   (b) $\mathrm{Cov}(\hat{\mathbf{w}}) = \sigma^2\left(\mathbf{X}^\top\mathbf{X}\right)^{-1}$.

2. We usually measure the quality of an estimator by mean squared error (MSE). The mean squared error (MSE) of estimator $\hat{\mathbf{w}}$ is defined as

$$\mathrm{MSE}(\hat{\mathbf{w}}) = \mathbb{E}[\|\hat{\mathbf{w}} - \mathbf{w}\|^2].$$

   Please derive that MSE can be decomposed into the variance of the estimator and the squared bias of the estimator, i.e.,

$$\begin{aligned}\mathrm{MSE}(\hat{\mathbf{w}}) &= \mathrm{trCov}(\hat{\mathbf{w}}) + \|\mathbb{E}\hat{\mathbf{w}} - \mathbf{w}\|^2 \\ &= \sum_{i=1}^{p}\mathrm{Var}(\hat{w_i}) + \sum_{i=1}^{p}(\mathbb{E}\hat{w_i} - w_i)^2.\end{aligned}$$

3. Please show that

$$\mathrm{MSE}(\hat{\mathbf{w}}) = \sigma^2\sum_{i=1}^{p}\frac{1}{\lambda_i},$$

   where $\lambda_1, \lambda_2, \ldots, \lambda_p$ are the eigenvalues of $\mathbf{X}^\top\mathbf{X}$.

4. What would happen if there exists an eigenvalue $\lambda_k \approx 0$?

Solution: ■

Exercise 5: Regularized least squares

Suppose that $\mathbf{X} \in \mathbb{R}^{n \times d}$.

1. Please show that $\mathbf{X}^\top \mathbf{X}$ is always positive semi-definite. Moreover, $\mathbf{X}^\top \mathbf{X}$ is positive definite if and only if $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_d$ are linearly independent.

2. Please show that $\mathbf{X}^\top \mathbf{X} + \lambda \mathbf{I}$ is always invertible, where $\lambda > 0$ and $\mathbf{I} \in \mathbb{R}^{d \times d}$ is an identity matrix.

3. Consider the regularized least squares linear regression and denote

$$\mathbf{w}^*(\lambda) = \operatorname*{\mathbf{argmin}}_{\mathbf{w}} L(\mathbf{w}) + \lambda \Omega(\mathbf{w}),$$

   where $L(\mathbf{w}) = \frac{1}{n} \|\mathbf{y} - \mathbf{X}\mathbf{w}\|_2^2$ and $\Omega(\mathbf{w}) = \|\mathbf{w}\|_2^2$. For regular parameters $0 < \lambda_1 < \lambda_2$, please show that $L(\mathbf{w}^*(\lambda_1)) < L(\mathbf{w}^*(\lambda_2))$ and $\Omega(\mathbf{w}^*(\lambda_1)) > \Omega(\mathbf{w}^*(\lambda_2))$. Explain intuitively why this holds.

Solution: ■

Exercise 6: Conditional Expectations (optional)

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 7: Bias-Variance Trade-off (Programming Exercise. You are required to finish at least one of Exercises 7 and 8.)

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 8: Bayesian Linear Regression (Programming Exercise. You are required to finish at least one of Exercises 7 and 8.)

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 9: 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

$$c = (2\pi)^{D/2} |\boldsymbol{\Sigma}|^{1/2}. \tag{3}$$

   Clearly, we must have

$$\int_{\mathbb{R}^D} \exp\left\{-\frac{1}{2}(\mathbf{x} - \boldsymbol{\mu})^{\top} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})\right\} d\mathbf{x} = c. \tag{4}$$

   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 (3) and (4) 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 10: Determinant and geometric (optional)

The determinant of a square matrix can be view as the signed volume spanned by the columns or rows of the matrix.

1. Consider three vectors $\mathbf{a} = (1, 2, 3, 4)^\top$, $\mathbf{b} = (5, 6, 7, 8)^\top$ and $\mathbf{c} = (7, -11, 1, 3)^\top$ in $\mathbb{R}^4$. Please find the volume of the parallelepipedon spanned by $\mathbf{a}$, $\mathbf{b}$ and $\mathbf{c}$. You may first find a unit vector $\mathbf{n}$ such that $\mathbf{n} \perp \text{Span}(\{\mathbf{a}, \mathbf{b}, \mathbf{c}\})$ and then calculate the volume of the parallelogram spanned by $\mathbf{a}$, $\mathbf{b}$, $\mathbf{c}$ and $\mathbf{n}$. Explain why you can do so.

2. Consider two vectors $\mathbf{a} = (1, 2, 3, 4)^\top$ and $\mathbf{b} = (5, 6, 7, 8)^\top$ in $\mathbb{R}^4$. Please find the area of the parallelogram spanned by $\mathbf{a}$ and $\mathbf{b}$.

3. Now we want to calculate the n-dimension volume of an n-dimension parallelepiped in $\mathbb{R}^m$ $(n \leq m)$. The parallelepiped P has the form

$$P = \left\{ \mathbf{a} + \sum_{i=1}^{n} \lambda_i \mathbf{b}_i; 0 \leq \lambda_i \leq 1, 1 \leq i \leq n \right\},$$

   where $\mathbf{a} \in \mathbb{R}^m$ and $\mathbf{b}_i \in \mathbb{R}^m$, $1 \leq i \leq n$. Show that the volume is given by

$$\text{V}(P) = \sqrt{\det(\mathbf{B}^\top \mathbf{B})},$$

   where $\mathbf{B}$ denotes the $m \times n$ matrix whose i^{th} column is the vector $\mathbf{b}_i$.

   [Hint: you may follow the idea in question 1.]

4. Suppose $\mathbf{A} \in \mathbb{R}^{n \times n}$, please show that

$$|\det(\mathbf{A})| \leq \prod_{i=1}^{n} \|\alpha_i\|_2,$$

   where α_i is the i^{th} row of $\mathbf{A}$. Then explain the geometrical meaning of this inequality.

Solution: ■

Exercise 11: Calculus in Bayesian linear regression

1. In Bayesian linear regression lecture, we suppose that the model parameter $\mathbf{w}$ has a Gaussian prior of the form $p(\mathbf{w}) = \mathcal{N}(\mathbf{w} \mid \boldsymbol{\mu}_0, \boldsymbol{\Sigma}_0)$. Then, given a set of input data instances $\{\mathbf{x}_i\}_{i=1}^n$, the joint distribution of the corresponding target variables is a Gaussian $p(\mathbf{y} \mid \mathbf{w}, \mathbf{X}) = \mathcal{N}\left(\mathbf{y} \mid \mathbf{X}\mathbf{w}, \sigma^2\mathbf{I}\right)$. We first find the joint distribution over $\mathbf{w}$ and $\mathbf{y}$. Let

$$\mathbf{z} = \begin{pmatrix} \mathbf{w} \\ \mathbf{y} \end{pmatrix}.$$

The log of the joint distribution is

$$\ln p(\mathbf{z}) = \ln p(\mathbf{w}) + \ln p(\mathbf{y} \mid \mathbf{w}).$$

The calculation of Gaussian density $p(\mathbf{z})$ is a tedious work. We are going to find a simpler way to calculate $p(\mathbf{z})$.

   (a) Let $\boldsymbol{\xi}$ is an n-dimension Gaussian random vector, satisfying $\mathbb{E}(\boldsymbol{\xi}) = \boldsymbol{\mu}, \mathrm{Cov}(\boldsymbol{\xi}) = \boldsymbol{\Sigma}$. Show that for any $n \times n$ matrix $\mathbf{A}$, $\mathbf{A}\boldsymbol{\xi} \sim \mathcal{N}\left(\mathbf{A}\boldsymbol{\mu}, \mathbf{A}\boldsymbol{\Sigma}\mathbf{A}^\top\right)$.

   (b) We can rewrite $\mathbf{y}$ as $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\epsilon}$, where $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ and is independent of $\mathbf{w}$. Please find the covariance matrix of $\mathbf{y}$ and the matrix $\mathbb{E}[(\mathbf{w} - \mathbb{E}[\mathbf{w}])(\mathbf{y} - \mathbb{E}[\mathbf{y}])^\top]$. Then write down the mean and covariance matrix of $\mathbf{z}$.

2. (Optional) We know that a multivariate Gaussian random vector $\mathbf{X}$ with uncorrelated components has mean vector $\boldsymbol{\mu}$ and the invertible covariance matrix $\boldsymbol{\Sigma}$. The probability density function of $\mathbf{X}$ is

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

But now we concentrate on the multivariate Gaussian random vector $\mathbf{X}$ with correlated components, which means that $|\boldsymbol{\Sigma}| = 0$ and $\boldsymbol{\Sigma}$ is not invertible.

Specifically, Suppose X is a Gaussian random variable with mean μ and variance σ. Another random variable $Y = aX + b$, where a, b are non-zero real numbers. Please show that X and Y are correlated, then find the joint density function of X and Y.

Hint: you may use the Dirac Delta function. This function, typically written $\delta(x)$, is defined as:

$$\delta(x) = \begin{cases} \infty & x = 0 \\ 0 & \text{otherwise} \end{cases}$$

with the properties that

   (a) $\int_{-\infty}^{\infty} \delta(x)dx = 1$;

   (b) $\int_{-\infty}^{\infty} f(x)\delta(x - x_0)\,dx = f(x_0)$ for any function $f(x)$ that is continuous around $x = x_0$.

Solution: ■

Exercise 12: Inverse of block matrix

Please prove Lemma 1 in Bayesian linear regression lecture.

Lemma 1.

Suppose that the involved matrices are invertible. Then,

$$\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{C} & \mathbf{D} \end{pmatrix}^{-1} = \begin{pmatrix} \mathbf{M} & -\mathbf{MBD}^{-1} \\ -\mathbf{D}^{-1}\mathbf{CM} & \mathbf{D}^{-1} + \mathbf{D}^{-1}\mathbf{CMBD}^{-1} \end{pmatrix}$$

where

$$\mathbf{M} = \left(\mathbf{A} - \mathbf{BD}^{-1}\mathbf{C}\right)^{-1}.$$

Please imitate the process of finding an inverse of a matrix $\mathbf{X}$, i.e., we first write $(\mathbf{X}, \mathbf{I})$ and then executing elementary row operations to get $(\mathbf{I}, \mathbf{X}^{-1})$.

Solution: ■

References

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