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

Lecturer: Jie Wang | Homework 1
Posted: Sep. 17, 2021 | Due: Sep. 29, 2021

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

Exercise 1: Basis and coordinates

Suppose that $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$ is a basis of an n-dimensional vector space V.

1. Show that $\{\lambda_1\mathbf{a}_1, \lambda_2\mathbf{a}_2, \ldots, \lambda_n\mathbf{a}_n\}$ is also a basis of V for nonzero scalars $\lambda_1, \lambda_2, \ldots, \lambda_n$.
2. Let $V = \mathbb{R}^n$ and $(\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n) = (\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n)\mathbf{P}$, where $\mathbf{P} \in \mathbb{R}^{n\times n}$ and $\mathbf{b}_i \in \mathbb{R}^n$, for any $i \in \{1, \ldots, n\}$. Show that $\{\mathbf{b}_1, \mathbf{b}_2, \ldots, \mathbf{b}_n\}$ is also a basis of V for any invertible matrix $\mathbf{P}$.
3. Suppose that the coordinate of a vector $\mathbf{v}$ under the basis $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$ is $\mathbf{x} = (x_1, x_2, \ldots x_n)$.
   (a) What is the coordinate of $\mathbf{v}$ under $\{\lambda_1\mathbf{a}_1, \lambda_2\mathbf{a}_2, \ldots, \lambda_n\mathbf{a}_n\}$?
   (b) What are the coordinates of $\mathbf{w} = \mathbf{a}_1 + \cdots + \mathbf{a}_n$ under $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_n\}$ and $\{\lambda_1\mathbf{a}_1, \lambda_2\mathbf{a}_2, \ldots, \lambda_n\mathbf{a}_n\}$? Note that $\lambda_i \neq 0$ for any $i \in \{1, \ldots, n\}$.

Solution: ■

Exercise 2: Derivatives with matrices

Definition 1 (Differentiability). [?] Let $f : \mathbb{R}^n \to \mathbb{R}^m$ be a function, $\mathbf{x}_0 \in \mathbb{R}^n$ be a point, and let $L : \mathbb{R}^n \to \mathbb{R}^m$ be a linear transformation. We say that f is *differentiable at* $\mathbf{x}_0$ *with derivative* L if we have

$$\lim_{\mathbf{x}\to\mathbf{x}_0;\mathbf{x}\neq\mathbf{x}_0} \frac{\|f(\mathbf{x}) - f(\mathbf{x}_0) - L(\mathbf{x}-\mathbf{x}_0)\|_2}{\|\mathbf{x}-\mathbf{x}_0\|_2} = 0.$$

We denote this derivative by $f'(\mathbf{x}_0)$.

1. Let $\mathbf{x}, \mathbf{a} \in \mathbb{R}^n$ and $\mathbf{y} \in \mathbb{R}^m$. Consider the functions as follows. Please show that they are differentiable and find $f'(\mathbf{x})$.

   (a) $f(\mathbf{x}) = \mathbf{a}^\top \mathbf{x}$.

   (b) $f(\mathbf{x}) = \mathbf{x}^\top \mathbf{x}$.

   (c) $f(\mathbf{x}) = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$, where $\mathbf{A} \in \mathbb{R}^{m\times n}$.

2. Please follow Definition 1 and give the definition of the differentiability of the functions $f : \mathbb{R}^{n\times n} \to \mathbb{R}$.

3. Let $f(\mathbf{X}) = \det(\mathbf{X})$, where $\det(\mathbf{X})$ is the determinant of $\mathbf{X} \in \mathbb{R}^{n\times n}$. Please discuss the differentiability of f rigorously according to your definition in the last part. If f is differentiable, please find $f'(\mathbf{X})$.

4. Let $f(\mathbf{X}) = \mathrm{tr}(\mathbf{A}^\top \mathbf{X})$, where $\mathbf{A}, \mathbf{X} \in \mathbb{R}^{n\times m}$, and $\mathrm{tr}(\cdot)$ denotes the trace of a matrix. Please discuss the differentiability of f and find f' if it is differentiable.

5. Let $\mathbf{S}_{++}^n$ be the space of all positive definite $n \times n$ matrices. Prove the function $f : \mathbf{S}_{++}^n \to \mathbb{R}$ defined by $f(\mathbf{X}) = \mathrm{tr}\,\mathbf{X}^{-1}$ is differentiable on $\mathbf{S}_{++}^n$. (Hint: Expand the expression $(\mathbf{X} + t\mathbf{Y})^{-1}$ as a power series.)

6. Define a function $f : \mathbf{S}_{++}^n \to \mathbb{R}$ by $f(\mathbf{X}) = \log\det\mathbf{X}$. Prove $\nabla f(\mathbf{I}) = \mathbf{I}$. Deduce $\nabla f(\mathbf{X}) = \mathbf{X}^{-1}$ for any $\mathbf{X}$ in $\mathbf{S}_{++}^n$.

Solution: ■

Exercise 3: Rank of matrices

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

1. Please show that
   (a) $\mathbf{rank}(\mathbf{A}) = \mathbf{rank}(\mathbf{A}^\top)$;
   (b) $\mathbf{rank}(\mathbf{AB}) \leq \mathbf{rank}(\mathbf{A})$;
   (c) $\mathbf{rank}(\mathbf{AB}) \leq \mathbf{rank}(\mathbf{B})$;
   (d) $\mathbf{rank}(\mathbf{A}) = \mathbf{rank}(\mathbf{A}^\top\mathbf{A})$.
2. The *column space* of $\mathbf{A}$ is defined by
   $$\mathcal{C}(\mathbf{A}) = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{y} = \mathbf{A}\mathbf{x},\, \mathbf{x} \in \mathbb{R}^n\}.$$
   The *null space* of $\mathbf{A}$ is defined by
   $$\mathcal{N}(\mathbf{A}) = \{\mathbf{x} \in \mathbb{R}^n : \mathbf{A}\mathbf{x} = 0\}.$$
   Notice that, the rank of $\mathbf{A}$ is the dimension of the column space of $\mathbf{A}$.
   Please show that
   (a) $\mathbf{rank}(\mathbf{A}) + \dim(\mathcal{N}(\mathbf{A})) = n$;
   (b) $\mathbf{y} = \mathbf{0}$ if and only if $\mathbf{a}_i^\top\mathbf{y} = 0$ for $i = 1, \ldots, m$, where $\mathbf{y} \in \mathbb{R}^m$ and $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m\}$ is a basis of $\mathbb{R}^m$.
3. Show that
   $$\mathbf{rank}(\mathbf{AB}) = \mathbf{rank}(\mathbf{B}) - \dim(\mathcal{C}(\mathbf{B}) \cap \mathcal{N}(\mathbf{A})). \tag{1}$$
4. Suppose that the first term on the right-hand side (RHS) of Eq. (1) changes to $\mathbf{rank}(\mathbf{A})$. Please find the second term on the RHS of Eq. (1) such that it still holds.
5. Show the results in 1. by Eq. (1) or the one you established in 4.

Solution: ■

Exercise 4: Linear equations

Consider the system of linear equations in $\mathbf{w}$

$$\mathbf{y} = \mathbf{X}\mathbf{w}, \tag{2}$$

where $\mathbf{y} \in \mathbb{R}^n$, $\mathbf{w} \in \mathbb{R}^d$, and $\mathbf{X} \in \mathbb{R}^{n \times d}$.

1. Give an example for "$\mathbf{X}$" and "$\mathbf{y}$" to satisfy the following three situations respectively:
   (a) there exists one unique solution;
   (b) there does not exist any solution;
   (c) there exists more than one solution.
2. Suppose that $\mathbf{X}$ has full column rank and $\mathbf{rank}((\mathbf{X}, \mathbf{y})) = \mathbf{rank}(\mathbf{X})$. Show that the system of linear equations (2) always admits a unique solution.
3. (**Normal equations**) Consider another system of linear equations in $\mathbf{w}$

   $$\mathbf{X}^\top \mathbf{y} = \mathbf{X}^\top \mathbf{X}\mathbf{w}. \tag{3}$$

   Please show that the system (3) always admits a solution. Moreover, does it always admit a unique solution?

Solution: ■

Exercise 5: 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

$$y = w_0 + w_1 x, \tag{4}$$

   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 (4) and the one as follows:

$$y = w_0 + w_1 x + w_2 x^2. \tag{5}$$

   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 (4), or $\hat{w}_0$, $\hat{w}_1$, and $\hat{w}_2$ if you choose the model in (5)].

Solution: ■

Exercise 6: 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 $\mathbf{P}\mathbf{x}$ 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 $\mathbf{P}_{\mathbf{B}}(\mathbf{x})$ the same as $\mathbf{P}_{\mathbf{B}}(\mathbf{P}_{\mathbf{A}}(\mathbf{x}))$? Please show your claim rigorously.

Solution: ■

Exercise 7: Linear regression by maximum likelihood

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 8: 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.,

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

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 9: 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: ■