Introduction to Machine Learning

Fall 2019

University of Science and Technology of China

Lecturer: Jie Wang	Homework 1
Posted: Seq. 20, 2019	Due: Seq. 30, 2019
Name: San Zhang	ID: PBXXXXXXXX

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

Exercise 1: Linear regression 20pts

Given 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 (1)].

Solution:

1. The average fitting error of the linear model over the whole data set is

$$L(w_0, w_1) = \frac{1}{n}\sum_{i=1}^{n}(y_i - (w_1 x_i + w_0))^2.$$

   As L is a quadratic function, it can attain its minimum. Let

$$\begin{cases} \frac{\partial L}{\partial w_0} = 0, \\ \frac{\partial L}{\partial w_1} = 0, \end{cases}$$

which is equivalent to

$$\begin{cases} \frac{2}{n}\sum_{i=1}^{n}((w_1x_i+w_0)-y_i) & =0, \\ \frac{2}{n}\sum_{i=1}^{n}x_i((w_1x_i+w_0)-y_i) & =0. \end{cases} \tag{3}$$

Solve the above equation. We know that

$$\begin{aligned} w_1 &= \frac{n\sum_{i=1}^{n}x_iy_i-\sum_{i=1}^{n}x_i\sum_{i=1}^{n}y_i}{n\sum_{i=1}^{n}x_i^2-(\sum_{i=1}^{n}x_i)^2}, \\ w_0 &= \frac{\sum_{i=1}^{n}y_i-w_1\sum_{i=1}^{n}x_i}{n}. \end{aligned}$$

2. The true model is $y=0.5x+1+\epsilon$, $\epsilon\sim N(0,0.1)$. We can see that the model in (2) has lower fitting error, which indicates the existence of overfitting.

■

Exercise 2: Rank of matrices 20pts

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) let $\mathbf{y} \in \mathbb{R}^m$, show that $\mathbf{y} = 0$ if and only if $\mathbf{a}_i^\top \mathbf{y} = 0$ for $i = 1, \ldots, m$, where $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_m\}$ is a basis of $\mathbb{R}^m$.

Solution:

1. Let $\mathcal{C}(\mathbf{A}) = \{\mathbf{y} \in \mathbb{R}^m : \mathbf{y} = \mathbf{A}\mathbf{x},\, \mathbf{x} \in \mathbb{R}^n\}$ denote the *column space* of $\mathbf{A}$. Then, $\mathbf{rank}(\mathbf{A}) = \dim(\mathcal{C}(\mathbf{A}))$.

   (a) Let $\mathbf{rank}(\mathbf{A}) = r$. Therefore, the dimension of the column space of $\mathbf{A}$ is r. Let $\mathbf{A}_r = (\mathbf{a}_1, \ldots, \mathbf{a}_r)$. WLOG, suppose that $\{\mathbf{a}_1, \ldots, \mathbf{a}_r\}$ are linearly independent. Consider the linear combination of $\{\mathbf{A}^\top\mathbf{a}_1, \ldots, \mathbf{A}^\top\mathbf{a}_r\}$:

$$\lambda_1\mathbf{A}^\top\mathbf{a}_1 + \cdots + \lambda_r\mathbf{A}^\top\mathbf{a}_r = 0, \tag{4}$$

   where $\lambda_1, \ldots, \lambda_r \in \mathbb{R}$. Let $\boldsymbol{\lambda} = (\lambda_1, \ldots, \lambda_r)^\top$, and then the equation (4) can be reformulated as

$$\mathbf{A}^\top\mathbf{A}_r\boldsymbol{\lambda} = 0,$$

   which implies that

$$\begin{aligned}
&\mathbf{A}_r^\top\mathbf{A}_r\boldsymbol{\lambda} = 0\\
\Rightarrow&\boldsymbol{\lambda}^\top\mathbf{A}_r^\top\mathbf{A}_r\boldsymbol{\lambda} = 0\\
\Rightarrow&\mathbf{A}_r\boldsymbol{\lambda} = 0\\
\Rightarrow&\boldsymbol{\lambda} = 0. \quad \text{(linearly independent)}
\end{aligned}$$

Therefore, $\{\mathbf{A}^\top \mathbf{a}_1, \ldots, \mathbf{A}^\top \mathbf{a}_r\}$ are linearly independent. As $\mathbf{A}^\top \mathbf{a}_1, \ldots, \mathbf{A}^\top \mathbf{a}_r \in \mathcal{C}(\mathbf{A}^\top)$, we know that

$$\mathbf{rank}(\mathbf{A}^\top) = \dim(\mathcal{C}(\mathbf{A}^\top) \geq r = \mathbf{rank}(\mathbf{A}).$$

In the same manner, we can show that

$$\mathbf{rank}(\mathbf{A}) \geq \mathbf{rank}(\mathbf{A}^\top).$$

Therefore,

$$\mathbf{rank}(\mathbf{A}) = \mathbf{rank}(\mathbf{A}^\top).$$

(b) Let $\mathbf{B} = (\mathbf{b}_1, \ldots, \mathbf{b}_p)$, where $\mathbf{b}_i$ denotes the ith column of $\mathbf{B}$. Since $\mathbf{A}\mathbf{b_i} \in \mathcal{C}(\mathbf{A})$, we have $\mathcal{C}(\mathbf{AB}) \subset \mathcal{C}(\mathbf{A})$. And thus $\mathbf{rank}(\mathbf{AB}) \leq \mathbf{rank}(\mathbf{A})$.

(c)

$$\mathbf{rank}(\mathbf{AB}) = \mathbf{rank}(\mathbf{B}^\top \mathbf{A}^\top) \leq \mathbf{rank}(\mathbf{B}^\top) = \mathbf{rank}(\mathbf{B}).$$

(d) Suppose that $\mathbf{e}_i = (0, \ldots, 0, 1, 0, \ldots)$, where the ith entry is one and the other entries are zero. Similar to the Exercise (a), we assume that $\mathbf{a}_j$ is the ith column of $\mathbf{A}$ and $\{\mathbf{a}_1, \ldots, \mathbf{a}_r\}$ are linearly independent. Then, we have

$$\mathbf{A}^\top \mathbf{a}_j = \mathbf{A}^\top \mathbf{A} \mathbf{e}_i,$$

which implies that

$$\mathbf{A}^\top \mathbf{a}_j \in \mathcal{C}(\mathbf{A}^\top \mathbf{A}).$$

From the Exercise (a), we know that $\{\mathbf{A}^\top \mathbf{a}_1, \ldots, \mathbf{A}^\top \mathbf{a}_r\}$ are linearly independent. Thus, we have

$$\mathbf{rank}(\mathbf{A}^\top \mathbf{A}) \geq \mathbf{rank}(\mathbf{A}).$$

By the result of the Exercise (c), we know that $\mathbf{rank}(\mathbf{A}^\top \mathbf{A}) \leq \mathbf{rank}(\mathbf{A})$. Therefore, $\mathbf{rank}(\mathbf{A}^\top \mathbf{A}) = \mathbf{rank}(\mathbf{A})$.

2. (a) Let $\mathbf{rank}(\mathbf{A}) = r$ and $\mathbf{a}_i$ be the ith column of $\mathbf{A}$. WLOG, suppose that $\{\mathbf{a}_1, \ldots, \mathbf{a}_r\}$ are linearly independent, $\mathbf{A}_r = (\mathbf{a}_1, \ldots, \mathbf{a}_r)$ and $\mathbf{A} = (\mathbf{A}_r \quad \mathbf{a}_{r+1} \ldots \mathbf{a}_n)$. As $\{\mathbf{a}_1, \ldots, \mathbf{a}_r\}$ is a basis of $\mathcal{C}(\mathbf{A})$, we know that $\mathbf{a}_i$, $i = r+1, \ldots, n$ can be written as the linear combination of $\{\mathbf{a}_1, \ldots, \mathbf{a}_r\}$. That is, there exists $\mathbf{M} \in \mathbb{R}^{r \times (n-r)}$, such that $(\mathbf{a}_{r+1} \ldots \mathbf{a}_n) = \mathbf{A}_r \mathbf{M}$. Thus,

$$\mathbf{A} = (\mathbf{A}_r \quad \mathbf{A}_r \mathbf{M}) = \mathbf{A}_r (\mathbf{I} \quad \mathbf{M}).$$

Next, we show that $\dim(\mathcal{N}(\mathbf{A})) = n-r$. Let $\mathbf{x} = (x_1, \dots, x_n)^\top$, $\mathbf{x}_r = (x_1, \dots, x_r)^\top$ and $\mathbf{x}_{n-r} = (x_{r+1}, \dots, x_n)^\top$. Considering the linear equation $\mathbf{Ax} = 0$, we have

$$\begin{aligned} \mathbf{Ax} &= 0 \\ \mathbf{A}_r(\mathbf{I}_r \quad \mathbf{M})\mathbf{x} &= 0 \\ (\mathbf{I}_r \quad \mathbf{M}) \begin{pmatrix} \mathbf{x}_r \\ \mathbf{x}_{n-r} \end{pmatrix} &= 0 \qquad (\mathbf{rank}(\mathbf{A}_r) = r) \\ \mathbf{x}_r &= -\mathbf{M}\mathbf{x}_{n-r} \end{aligned}$$

Thus the solution of $\mathbf{Ax} = 0$ is

$$\mathbf{x} = \begin{pmatrix} -\mathbf{M} \\ \mathbf{I}_{n-r} \end{pmatrix} \mathbf{x}_{n-r}.$$

Therefore, the dimension of the solution space of $\mathbf{Ax} = 0$ is $n - r$, which implies that $\dim(\mathcal{N}(\mathbf{A})) = n - r$.

(b) $\Rightarrow$ Trivial.

$\Leftarrow$ Let $\mathbf{C} = (\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_m)$. Since $\{\mathbf{a}_1, \mathbf{a}_2, \dots, \mathbf{a}_m\}$ is a basis of $\mathbb{R}^m$, we have $\mathbf{rank}(\mathbf{C}^\top) = m$. Thus, $\mathbf{C}^\top \mathbf{y} = 0 \Rightarrow \mathbf{y} = 0$.

■

Exercise 3: Projection 30pts

Let $C \subset \mathbb{R}^n$ be a closed convex set and $\mathbf{x} \in \mathbb{R}^n$. Define

$$\mathbf{P}_C(\mathbf{x}) = \arg\min_{\mathbf{y}\in C} \|\mathbf{y}-\mathbf{x}\|_2.$$

We call $\mathbf{P}_C(\mathbf{x})$ the projection of the point $\mathbf{x}$ onto the convex set C.

1. Show that any finite dimensional space is convex.

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 $\mathbf{u}, \mathbf{w} \in \mathbb{R}^n$.

    (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 $\mathbf{P}_{\mathbf{V}}(\mathbf{w})$, which is the projection of $\mathbf{w}$ onto $\mathcal{C}(\mathbf{V})$, 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 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 are the eigenspaces corresponding to $\lambda = 1$ and $\lambda = 0$, respectively.*)

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

Solution:

1. Suppose that M is a n dimensional space, and $\{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_n\}$ is a basis of M. Then for all $\mathbf{x}, \mathbf{y} \in M$, there exists α_i and β_i such that $\mathbf{x} = \sum_{i=1}^n \alpha_i\mathbf{x}_i$ and $\mathbf{y} = \sum_{i=1}^n \beta_i\mathbf{x}_i$.

    For all $\lambda \in (0, 1)$, we have

    $$\begin{aligned}\lambda\mathbf{x} + (1-\lambda)\mathbf{y} &= \lambda\sum_{i=1}^n \alpha_i\mathbf{x}_i + (1-\lambda)\sum_{i=1}^n \beta_i\mathbf{x}_i\\ &= \sum_{i=1}^n(\lambda\alpha_i + (1-\lambda)\beta_i)\mathbf{x}_i \in M.\end{aligned}$$

Therefore, M is convex.

2. (a) Let $\mathbf{span}\{\mathbf{v}_1\}$ denote the subspace spanned by $\mathbf{v}_1$. For every $\mathbf{y} \in \mathbf{span}\{\mathbf{v}_1\}$, there exists $\lambda \in \mathbb{R}$ such that $\mathbf{y} = \lambda\mathbf{v}_1$. Then we have

$$\begin{aligned}\min_{\mathbf{y}\in C}\|\mathbf{y}-\mathbf{x}\|_2^2 &= \min_{\lambda\in\mathbb{R}}\|\lambda\mathbf{v}_1-\mathbf{x}\|_2^2\\ &= \min_{\lambda\in\mathbb{R}}(\lambda^2\|\mathbf{v}_1\|_2^2 - 2\lambda\langle\mathbf{v}_1,\mathbf{x}\rangle + \|\mathbf{x}\|_2^2)\\ &= \min_{\lambda\in\mathbb{R}}\left(\|\mathbf{v}_1\|_2^2\left(\lambda-\frac{\langle\mathbf{v}_1,\mathbf{x}\rangle}{\|\mathbf{v}_1\|_2^2}\right)^2 - \frac{|\langle\mathbf{v}_1,\mathbf{x}\rangle|^2}{\|\mathbf{v}_1\|_2^2}+\|\mathbf{x}\|_2^2\right).\end{aligned}$$

Notice that $\|\mathbf{v}_1\|_2^2 > 0$. We have

$$\begin{aligned}\mathbf{P}_{\mathbf{v}_1}(\mathbf{x}) &= \arg\min_{\mathbf{y}=\lambda\mathbf{v}_1}\|\mathbf{y}-\mathbf{x}\|_2\\ &= \left(\arg\min_{\lambda}\|\lambda\mathbf{v}_1-\mathbf{x}\|_2\right)\mathbf{v}_1\\ &= \frac{\langle\mathbf{v}_1,\mathbf{x}\rangle}{\|\mathbf{v}_1\|_2^2}\mathbf{v}_1.\end{aligned}$$

(b) Let $\alpha, \beta \in \mathbb{R}$ and $\mathbf{u}, \mathbf{w} \in \mathbb{R}^n$. We have

$$\begin{aligned}\mathbf{P}_{\mathbf{v}_1}(\alpha\mathbf{u}+\beta\mathbf{w}) &= \frac{\langle\mathbf{v}_1,\alpha\mathbf{u}+\beta\mathbf{w}\rangle}{\|\mathbf{v}_1\|_2^2}\mathbf{v}_1\\ &= \alpha\frac{\langle\mathbf{v}_1,\mathbf{u}\rangle}{\|\mathbf{v}_1\|_2^2}\mathbf{v}_1 + \beta\frac{\langle\mathbf{v}_1,\mathbf{v}\rangle}{\|\mathbf{v}_1\|_2^2}\mathbf{v}_1\\ &= \alpha\mathbf{P}_{\mathbf{v}_1}(\mathbf{u}) + \beta\mathbf{P}_{\mathbf{v}_1}(\mathbf{w}).\end{aligned}$$

(c)

$$\begin{aligned}\mathbf{P}_{\mathbf{v}_1}(\mathbf{x}) &= \frac{\langle\mathbf{v}_1,\mathbf{x}\rangle}{\|\mathbf{v}_1\|_2^2}\mathbf{v}_1\\ &= \mathbf{v}_1\frac{\mathbf{v}_1^\top\mathbf{x}}{\mathbf{v}_1^\top\mathbf{v}_1}\\ &= \left(\frac{\mathbf{v}_1\mathbf{v}_1^\top}{\mathbf{v}_1^\top\mathbf{v}_1}\right)\mathbf{x}\\ \Rightarrow \mathbf{H}_1 &= \frac{\mathbf{v}_1\mathbf{v}_1^\top}{\mathbf{v}_1^\top\mathbf{v}_1}.\end{aligned}$$

(d) i. Assume that $\{\mathbf{v}_1, \ldots, \mathbf{v}_d, \mathbf{v}_{d+1}, \ldots, \mathbf{v}_n\}$ is a basis of $\mathbb{R}^n$, where $\langle\mathbf{v}_i, \mathbf{v}_j\rangle = 0$ for all $1 \leq i \leq d$ and $d+1 \leq j \leq n$. For all $\mathbf{y} \in \mathcal{C}(\mathbf{A})$, there exist $\lambda_y^i \in \mathbb{R}$, such that $\mathbf{y} = \sum_{i=1}^d \lambda_y^i\mathbf{v}_i$. For all $\mathbf{z} \in \mathbb{R}^n$, there exist $\lambda_z^i \in \mathbb{R}$, such that $\mathbf{z} = \sum_{i=1}^n \lambda_z^i\mathbf{v}_i$. Let $\boldsymbol{\lambda}_z = (\lambda_z^1, \ldots, \lambda_z^d)^\top$. Then, $\mathbf{z} = \mathbf{V}\boldsymbol{\lambda}_z$ and $\boldsymbol{\lambda}_z = (\mathbf{V}^\top\mathbf{V})^{-1}\mathbf{V}^\top\mathbf{z}$.

Suppose that $\mathbf{u} = \mathbf{P}_{\mathcal{C}(\mathbf{A})}(\mathbf{x})$, then

$$\begin{aligned}
\|\mathbf{u}-\mathbf{x}\|_2^2 &= \left\| \sum_{i=1}^{d} \lambda_z^i \mathbf{v}_i - \mathbf{x} \right\|_2^2 \\
&= \arg\min_{\lambda_y^i \in \mathbb{R}} \left\| \sum_{i=1}^{d} \lambda_y^i \mathbf{v}_i - \mathbf{x} \right\|_2^2 \\
&= \arg\min_{\lambda_y^i \in \mathbb{R}} \left\| \sum_{i=1}^{d} \lambda_y^i \mathbf{v}_i - \sum_{i=1}^{n} \lambda_x^i \mathbf{v}_i \right\|_2^2 \\
&= \arg\min_{\lambda_y^i \in \mathbb{R}} \left\| \sum_{i=1}^{d} (\lambda_y^i - \lambda_x^i) \mathbf{v}_i - \sum_{i=d+1}^{n} \lambda_x^i \mathbf{v}_i \right\|_2^2 \\
&= \arg\min_{\lambda_y^i \in \mathbb{R}} \left\| \sum_{i=1}^{d} (\lambda_y^i - \lambda_x^i) \mathbf{v}_i \right\|^2 + \left\| \sum_{i=d+1}^{n} \lambda_x^i \mathbf{v}_i \right\|_2^2 .
\end{aligned}$$

It is easy to see that $\lambda_z^i = \lambda_x^i$, $i = 1, \ldots, d$. That is,

$$\begin{aligned}
\mathbf{z} &= \mathbf{V}\boldsymbol{\lambda}_z \\
&= \mathbf{V}\boldsymbol{\lambda}_x \\
&= \mathbf{V}(\mathbf{V}^\top \mathbf{V})^{-1}\mathbf{V}^\top \mathbf{x}.
\end{aligned}$$

Therefore, we know that $\mathbf{H} = \mathbf{V}(\mathbf{V}^\top \mathbf{V})^{-1}\mathbf{V}^\top$.

ii. If we further assume that $\mathbf{v}_i^\top \mathbf{v}_j = 0$, $\forall\, i \neq j$, then

$$\mathbf{V}^\top \mathbf{V} = \mathbf{diag}\,\{\|\mathbf{v}_1\|_2^2, \ldots, \|\mathbf{v}_d\|_2^2\}.$$

Thus,

$$\begin{aligned}
\mathbf{H} &= \mathbf{V}(\mathbf{V}^\top \mathbf{V})^{-1}\mathbf{V}^\top \\
&= \sum_{i=1}^{d} \mathbf{v}_i \mathbf{v}_i^\top / \|\mathbf{v}_i\|_2^2.
\end{aligned}$$

3. (a) Suppose that $\mathbf{x}$ is the eigenvector corresponding to the eigenvalue λ. Then, we have

$$\mathbf{P}\mathbf{x} = \lambda \mathbf{x},$$

which implies that

$$\mathbf{P}^2\mathbf{x} = \mathbf{P}(\mathbf{P}\mathbf{x}) = \mathbf{P}(\lambda\mathbf{x}) = \lambda\mathbf{P}\mathbf{x} = \lambda^2\mathbf{x}.$$

On the other hand, it is easy to see that $\mathbf{P}(\mathbf{P}\mathbf{x}) = \mathbf{P}\mathbf{x} = \lambda\mathbf{x}$. Thus, we have

$$\lambda\mathbf{x} = \lambda^2\mathbf{x},$$

which implies that

$$\lambda = 0 \text{ or } 1.$$

When $\lambda = 0$, $\mathbf{Px} = \mathbf{0}$. That is, $\mathbf{x} \in \mathcal{N}(\mathbf{P})$. For all $\mathbf{y} \in \mathcal{N}(\mathbf{P})$, $\mathbf{Py} = 0$. Therefore, a vector $\mathbf{y}$ is the eigenvector corresponding to 0 iff $\mathbf{y} \in \mathcal{N}(\mathbf{P})$.

When $\lambda = 1$, $\mathbf{Px} = \mathbf{x}$. This implies that $\mathbf{x} \in \mathcal{C}(\mathbf{P})$. For all $\mathbf{y} \in \mathcal{C}(\mathbf{P})$, $\mathbf{Py} = \mathbf{y}$. Therefore, a vector $\mathbf{y}$ is the eigenvector corresponding to 1 iff $\mathbf{y} \in \mathcal{C}(\mathbf{P})$.

(b) "$\Rightarrow$:" Suppose that $\mathbf{P}$ is a projection matrix, and $\mathbf{Px} = \mathbf{v}$. By the definition of the projection operator, there exist subspaces $U, V \subset \mathbb{R}^n$, such that $\mathbb{R}^n = U \oplus V$. For all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x}$ can be uniquely written as $\mathbf{x} = \mathbf{u} + \mathbf{v}$, where $\mathbf{u} \in U$ and $\mathbf{v} \in V$. We have $\mathbf{Px} = \mathbf{v} \in V$.

It is easy to see that

$$\mathbf{P}^2\mathbf{x} = \mathbf{P}(\mathbf{Px}) = \mathbf{Pv} = \mathbf{0} + \mathbf{v},$$

where $\mathbf{0} \in U$ and $\mathbf{v} \in V$. Thus, $\mathbf{P}^2\mathbf{x} = \mathbf{v}$ for all $\mathbf{x} \in \mathbb{R}^n$, i.e., $\mathbf{P}^2 = \mathbf{P}$.

"$\Leftarrow$:" First, we show that $\mathbb{R}^n = \mathcal{C}(\mathbf{P}) \oplus \mathcal{N}(\mathbf{P})$ if $\mathbf{P}^2 = \mathbf{P}$.

Suppose that $\mathbf{x} \in \mathcal{N}(\mathbf{P})$. Then, $\mathbf{P}(\mathbf{x} - \mathbf{Px}) = \mathbf{Px} - \mathbf{P}^2\mathbf{x} = 0$. Thus, $\mathcal{N}(\mathbf{P}) \subset \{\mathbf{x} - \mathbf{Px}\}$.

Suppose that $\mathbf{y} \in \{\mathbf{x} - \mathbf{Px}\}$. Then, there exists $\mathbf{x}_0$, such that $\mathbf{y} = \mathbf{x}_0 - \mathbf{P}x_0$. As $\mathbf{Py} = \mathbf{P}(\mathbf{x}_0 - \mathbf{Px}_0) = 0$, we know that $\{\mathbf{x} - \mathbf{Px}\} \subset \mathcal{N}(\mathbf{P})$.

Therefore, we have $\{\mathbf{x} - \mathbf{Px}\} = \mathcal{N}(\mathbf{P})$.

For all $\mathbf{x} \in \mathbb{R}^n$, $\mathbf{x} = (\mathbf{x} - \mathbf{Px}) + \mathbf{Px}$, where $(\mathbf{x} - \mathbf{Px}) \in \mathcal{N}(\mathbf{P})$ and $\mathbf{Px} \in \mathcal{C}(\mathbf{P})$. Thus, $\mathbb{R}^n \subset \mathcal{C}(\mathbf{P}) + \mathcal{N}(\mathbf{P})$. Further, we know that $\dim(\mathcal{C}(\mathbf{P})) + \dim(\mathcal{N}(\mathbf{P})) = n$, which implies that

$$\mathbb{R}^n = \mathcal{C}(\mathbf{P}) \oplus \mathcal{N}(\mathbf{P}).$$

By the definition of the projection, we know that $\mathbf{P}$ is a projection matrix.

Remark: Note that we use the following definition of the projection operator.

Definition: Let V be a vector space and let U and W be subspaces of V such that $V = U \oplus W$. Then v can be written uniquely as $v = u + w$ where $u \in U$ and $w \in W$. The Projection Operator Onto U is the linear operator $P_{U,W}$ defined by $P_{U,W}(v) = u$ for all $v \in V$.

If we use the definition that $\mathbf{P}_C(\mathbf{x}) = \arg\min_{\mathbf{y} \in C} \|\mathbf{y} - \mathbf{x}\|_2$, then we require an additional condition that $\mathbf{P}^\top = \mathbf{P}$, which implies that $\mathcal{C}(\mathbf{P}) \perp \mathcal{N}(\mathbf{P})$.

■

Exercise 4: 5pts

Let $\mathbf{x} \in \mathbf{R}^n$. Find the gradients of the following functions.

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

Solution:

1. $\nabla f(\mathbf{x}) = \mathbf{a}$.
2. $\nabla f(\mathbf{x}) = 2\mathbf{x}$.
3. $\nabla f(\mathbf{x}) = 2\mathbf{A}^\top (\mathbf{A}\mathbf{x} - \mathbf{y})$.

■

Exercise 5: Second-order sufficient optimality conditions 10pts

Suppose that $f : \mathbb{R}^n \to \mathbb{R}$ is twice differentiable at $\mathbf{x}$. If $\nabla f(\mathbf{x}) = 0$ and the Hessian matrix $\mathbf{H}(\mathbf{x})$ is positive definite, then $\mathbf{x}$ is a strict local minimum.

1. Show the above result by contradiction.
2. Show the result by NOT using contradiction. [*Hint: you may need eigen-decomposition.*]

Solution:

1. Suppose that $\mathbf{x}$ is not a strict local minimum. Then, there is a sequence $\{\mathbf{x}_k\}$ with $\mathbf{x}_k \to \mathbf{x}$ and $f(\mathbf{x}_k) \leq f(\mathbf{x}), \forall k = 1, 2, \ldots$. Let

$$\mathbf{d}_k = \frac{\mathbf{x}_k - \mathbf{x}}{\|\mathbf{x}_k - \mathbf{x}\|_2}.$$

As $\|\mathbf{d}_k\| = 1$ for all k, we can find a subsequence $\{\mathbf{d}_{k_j}\}$ that converges to a vector $\mathbf{d}$ (Bolzano–Weierstrass Theorem), i.e., $\lim_{j\to\infty} \mathbf{d}_{k_j}$. Moreover, for each j

$$\begin{aligned} f(\mathbf{x}_{k_j}) =& f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{x}_{k_j} - \mathbf{x}\rangle + \frac{1}{2}\langle \mathbf{H}(\mathbf{x})(\mathbf{x}_{k_j} - \mathbf{x}), \mathbf{x}_{k_j} - \mathbf{x}\rangle \\ &+ \|\mathbf{x}_{k_j} - \mathbf{x}\|_2^2 \phi_{\mathbf{x}}(\mathbf{x}_{k_j} - \mathbf{x}), \end{aligned}$$

where $\lim_{\mathbf{x}_{k_j}\to\mathbf{x}} \phi_{\mathbf{x}}(\mathbf{x}_{k_j} - \mathbf{x}) = 0$. Thus,

$$\frac{f(\mathbf{x}_{k_j}) - f(\mathbf{x})}{\|\mathbf{x}_{k_j} - \mathbf{x}\|_2^2} = \frac{1}{2}\langle \mathbf{H}(\mathbf{x})\mathbf{d}_{k_j}, \mathbf{d}_{k_j}\rangle + \phi_{\mathbf{x}}(\mathbf{x}_{k_j} - \mathbf{x}),$$

which implies that $\langle \mathbf{H}(\mathbf{x})\mathbf{d}, \mathbf{d}\rangle \leq 0$ contradicting the fact that $\mathbf{H}(\mathbf{x}) \succ 0$.

2. As f is twice differentiable, the Taylor's theorem leads to:

$$f(\mathbf{x} + t\mathbf{d}) = f(\mathbf{x}) + t\langle \nabla f(\mathbf{x}), \mathbf{d}\rangle + \frac{1}{2}t^2\langle \mathbf{H}(\mathbf{x})\mathbf{d}, \mathbf{d}\rangle + \|t\mathbf{d}\|_2^2 \phi_{\mathbf{x}}(t\mathbf{d}), \tag{5}$$

where $\mathbf{d}$ is a vector with unit length, i.e., $\|\mathbf{d}\|_2 = 1$, and $\lim_{t\to 0} \phi_{\mathbf{x}}(t\mathbf{d}) = 0$.

By the assumption, the Hessian matrix $\mathbf{H}(\mathbf{x})$ is positive definite. Thus,

$$\mathbf{H}(\mathbf{x}) = \mathbf{U}\mathbf{\Sigma}\mathbf{U}^\top, \quad \text{(eigen-decomposition)}$$

where $\mathbf{\Sigma} = \mathbf{diag}\,(\lambda_1, \lambda_2, ..., \lambda_n), \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n > 0$ and $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$. Let $\mathbf{U} = (\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_n)$. Clearly, the set of vectors $\{\mathbf{u}_1, \mathbf{u}_2, ..., \mathbf{u}_n\}$ is an orthonormal basis of $\mathbb{R}^n$. Thus, we can write the vector $\mathbf{d}$ in (5) as

$$\mathbf{d} = \sum_{i=1}^n \alpha_i \mathbf{u}_i,$$

where $\alpha_i = \langle \mathbf{d}, \mathbf{u}_i\rangle$ and $\sum_{i=1}^n \alpha_i^2 = 1$.

By the assumption, we also have $\nabla f(\mathbf{x}) = 0$. Therefore, (5) becomes

$$\begin{aligned} f(\mathbf{x}+t\mathbf{d}) &= f(\mathbf{x}) + \frac{1}{2}t^2\langle \mathbf{H}(\mathbf{x})\mathbf{d}, \mathbf{d}\rangle + \|t\mathbf{d}\|_2^2\phi_{\mathbf{x}}(t\mathbf{d}) \\ &= f(\mathbf{x}) + \frac{1}{2}t^2\langle \mathbf{U}\boldsymbol{\Sigma}\mathbf{U}^\top\mathbf{d}, \mathbf{d}\rangle + \|t\mathbf{d}\|_2^2\phi_{\mathbf{x}}(t\mathbf{d}) \\ &= f(\mathbf{x}) + \frac{1}{2}t^2\sum_{i=1}^n \lambda_i\alpha_i^2 + \|t\mathbf{d}\|_2^2\phi_x(t\mathbf{d}) \\ &\geq f(\mathbf{x}) + \frac{1}{2}t^2\sum_{i=1}^n \lambda_n\alpha_i^2 + \|t\mathbf{d}\|_2^2\phi_{\mathbf{x}}(t\mathbf{d}) \\ &= f(\mathbf{x}) + t^2(\lambda_n/2 + \phi_{\mathbf{x}}(t\mathbf{d})). \end{aligned}$$

Notice that $\lim_{t\to 0}\phi_{\mathbf{x}}(t\mathbf{d}) = 0$. Thus, for $\lambda_n/4$, there exists a $\delta > 0$ such that

$$|\phi_{\mathbf{x}}(t\mathbf{d})| < \lambda_n/4, \, \forall\, |t| = \|t\mathbf{d}\| < \delta,$$

Consequently, we have

$$f(\mathbf{x}+t\mathbf{d}) > f(\mathbf{x}) + t^2\lambda_n/4, \, \forall\, |t| = \|t\mathbf{d}\| < \delta,$$

which implies that $\mathbf{x}$ is a strict local minimum. This completes the proof.

■

Exercise 6: Identically independently distributed 10pts

Suppose that the training samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are i.i.d.. show that

$$p(\mathbf{x}_1, \ldots, \mathbf{x}_n) = \prod_{i=1}^n p(\mathbf{x}_i).$$

Solution: As the training samples $\{(\mathbf{x}_i, y_i)\}_{i=1}^n$ are i.i.d., we have

$$\begin{aligned} p((\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)) &= p(\mathbf{x}_1, \ldots, \mathbf{x}_n, y_1, \ldots, y_n) && (6) \\ &= \prod_{i=1}^n p(\mathbf{x}_i, y_i). && (7) \end{aligned}$$

Therefore, we know that

$$\begin{aligned} p(\mathbf{x}_1, \ldots, \mathbf{x}_n) &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} p(\mathbf{x}_1, \ldots, \mathbf{x}_n, y_1, \ldots, y_n) \mathrm{d}y_1 \cdots \mathrm{d}y_n \\ &= \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{i=1}^n p(\mathbf{x}_i, y_i) \mathrm{d}y_1 \cdots \mathrm{d}y_n \\ &= \prod_{i=1}^n \int_{-\infty}^{\infty} p(\mathbf{x}_i, y_i) \mathrm{d}y_i \\ &= \prod_{i=1}^n p(\mathbf{x}_i). \end{aligned}$$

■

Exercise 7: First-order condition II 5pts

Suppose that f is continuously differentiable. Prove that f is convex if and only if $\mathbf{dom} f$ is convex and

$$\langle \nabla f(\mathbf{x}) - \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle \geq 0.$$

Solution: $\Rightarrow$ The convexity of f implies that

$$\begin{aligned} f(\mathbf{y}) &\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle, \\ f(\mathbf{x}) &\geq f(\mathbf{y}) + \langle \nabla f(\mathbf{y}), \mathbf{x} - \mathbf{y} \rangle. \end{aligned}$$

Adding them together leads to desired result.

$\Leftarrow$ Let $\mathbf{x}_t = \mathbf{x} + t(\mathbf{y} - \mathbf{x})$. Then,

$$\begin{aligned} f(\mathbf{y}) &= f(\mathbf{x}) + \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle dt \\ &= f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \frac{1}{t} \langle \nabla f(\mathbf{x}_t) - \nabla f(\mathbf{x}), \mathbf{x}_t - \mathbf{x} \rangle dt \\ &\geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle. \end{aligned}$$

■