# Quiz 1: Introduction to Machine Learning

University of Science and Technology of China

Oct. 12, 2019

---

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

**Problem 1: The ETA Problem** 10pts

1. (3pts) Please write down the pipeline of a typical machine learning solution.
2. (7pts) Given a data set $\{(x_i, y_i)\}_{i=1}^n$, where $x_i, y_i \in \mathbb{R}$. 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).

**Solution:**

1. (a) Data preparation.
   (b) Select a hypothesis set.
   (c) Select a hypothesis.
2. The average fitting error of the linear model over the whole data set is

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

Let

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

which is equivalent to

$$\begin{cases} \frac{2}{n}\sum_{i=1}^{n}((w_1 x_i + w_0) - y_i) & = 0, \\ \frac{2}{n}\sum_{i=1}^{n} x_i((w_1 x_i + w_0) - y_i) & = 0. \end{cases} \tag{2}$$

Solving the equation (2), we know that

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

■

**Problem 2: Linear Algebra** 30pts

1. (10pts) Let $V$ be a finite dimensional linear space. Suppose that $A = (\mathbf{u}_1, \mathbf{u}_2, \ldots, \mathbf{u}_m)$ and $B = (\mathbf{v}_1, \mathbf{v}_2, \ldots, \mathbf{v}_n)$ are two bases of $V$. Show that $m = n$.

2. (10pts) Suppose that $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ is a list of vectors with $\mathbf{x}_p \neq \mathbf{0}$. Let

$$\mathbf{y}_i = \mathbf{x}_i + a_i \mathbf{x}_p,\ i = 1, 2, \ldots, p-1,$$

   where $a_1, a_2, \ldots, a_{p-1} \in \mathbb{R}$.

   Show that $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{p-1})$ is linearly independent if $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ is linearly independent.

3. (10pts) Let $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m \times n}$. Show that

$$\mathbf{rank}(\mathbf{A} + \mathbf{B}) \leq \mathbf{rank}([\mathbf{A}\ \mathbf{B}]) \leq \mathbf{rank}(\mathbf{A}) + \mathbf{rank}(\mathbf{B}).$$

**Solution:**

1. As $A$ is a basis of $V$, each vector in $B$ can be written as a linear combination of $A$. That is, for all integers $j \in [1, n]$, we have

$$\mathbf{v}_j = a_{1j}\mathbf{u}_1 + \cdots + a_{mj}\mathbf{u}_m.$$

   Suppose that there exist $\lambda_1, \cdots, \lambda_n \in \mathbb{R}$, such that

$$\lambda \mathbf{v}_1 + \cdots + \lambda_n \mathbf{v}_n = 0.$$

   Then we have

$$\begin{aligned} &\lambda_1(a_{11}\mathbf{u}_1 + \cdots + a_{m1}\mathbf{u}_m) + \cdots + \lambda_n(a_{1n}\mathbf{u}_1 + \cdots + a_{mn}\mathbf{u}_m) = 0, \\ \Rightarrow &(a_{11}\lambda_1 + \cdots + a_{1n}\lambda_n)\mathbf{u}_1 + \cdots + (a_{m1}\lambda_1 + \cdots + a_{mn}\lambda_n)\mathbf{u}_m = 0. \end{aligned}$$

   Consider the linear system

$$\begin{cases} a_{11}\lambda_1 & + \cdots + a_{1n}\lambda_n = 0, \\ & \cdots\cdots \\ a_{m1}\lambda_1 & + \cdots + a_{mn}\lambda_n = 0. \end{cases} \tag{3}$$

   There are $n$ variables and $m$ equations in (3). If $n > m$, then (3) has a nonzero solution, which implies that $B$ is linear dependent. This leads to a contradiction. Therefore, we have $n \leq m$.

   In the same manner, we can show that $m \geq n$.

   Thus, $m = n$.

2. Suppose that there exist $\lambda_1, \ldots, \lambda_{p-1} \in \mathbb{R}$ such that

$$\sum_{i=1}^{p-1} \lambda_i \mathbf{y}_i = 0.$$

As $\mathbf{y}_i = \mathbf{x}_i + a_i \mathbf{x}_p$ for all integers $i \in [1, p-1]$, we have

$$\begin{aligned} &\sum_{i=1}^{p-1} \lambda_i (\mathbf{x}_i + a_i \mathbf{x}_p) \\ =&\sum_{i=1}^{p-1} \lambda_i \mathbf{x}_i + \sum_{i=1}^{p-1} \lambda_i a_i \mathbf{x}_p. \end{aligned}$$

As $(\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_p)$ is linearly independent, we know that $\lambda_i = 0$ for $i = 1, 2, \ldots, p-1$ and $\sum_{i=1}^{p-1} \lambda_i a_i = 0$. Therefore, $(\mathbf{y}_1, \mathbf{y}_2, \ldots, \mathbf{y}_{p-1})$ is linearly independent.

3. As the elementary transformations do not change the rank of a matrix, we have

$$\begin{aligned} \mathbf{rank}([\mathbf{A}\ \mathbf{B}]) = \mathbf{rank}([\mathbf{A}+\mathbf{B}\ \mathbf{B}]) &\geq \mathbf{rank}(\mathbf{A}+\mathbf{B}), \\ \mathbf{rank}(\mathbf{A}) + \mathbf{rank}(\mathbf{B}) = \mathbf{rank}\begin{pmatrix} \mathbf{A} & \mathbf{O} \\ \mathbf{O} & \mathbf{B} \end{pmatrix} = \mathbf{rank}\begin{pmatrix} \mathbf{A} & \mathbf{B} \\ \mathbf{O} & \mathbf{B} \end{pmatrix} &\geq \mathbf{rank}([\mathbf{A}\ \mathbf{B}]). \end{aligned}$$

■

**Problem 3: Convex Optimization** 30pts

A continuously differentiable function $f : \mathbb{R}^n \to \mathbb{R}$ is called strongly convex if there exists a constant $\mu > 0$ such that

$$f(\mathbf{y}) \geq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|_2^2, \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n.$$

The constant $\mu$ is called the convexity parameter of $f$.

1. (20pts) Suppose that the function $g$ is twice continuously differentiable. Show that $g$ is strongly convex if and only if $\nabla^2 g(\mathbf{x}) - m\mathbf{I}$ is positive semi-definite for some $m > 0$ and all $\mathbf{x} \in \mathbb{R}^n$, where $\mathbf{I}$ is the identity matrix.

2. (10pts) Consdier the problem

$$\min_{\mathbf{x}} \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2, \tag{4}$$

   where $\mathbf{x} \in \mathbb{R}^d$, $\mathbf{A} \in \mathbb{R}^{n \times d}$, and $\mathbf{y} \in \mathbb{R}^n$. Suppose $\mathbf{rank}(\mathbf{A}) = d$. Show that it admits a unique solution.

3. (**Bonus** 10pts) The vector $\mathbf{y} - \mathbf{A}\mathbf{x}$ is the so-called residual. The informal idea of regression is to minimize the norm of the residual such that the model can well explain the data. However, there are special cases where the norm of the residual will never be less than the norm of $\mathbf{y}$ no matter which $\mathbf{x}$ you choose for nontrivial data ($\mathbf{y} \neq \mathbf{0}, \mathbf{A} \neq \mathbf{0}$). Can you figure out one of these cases?

**Solution:**

1. Let the convexity parameter of $g$ is $\mu$. Let

$$\phi(\mathbf{x}) = g(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|_2^2.$$

   The strong convexity of $g$ implies that

$$\begin{aligned}
& g(\mathbf{y}) \geq g(\mathbf{x}) + \langle \nabla g(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{\mu}{2}\|\mathbf{y} - \mathbf{x}\|_2^2, \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \\
\Leftrightarrow\ & g(\mathbf{y}) - \frac{\mu}{2}\|\mathbf{y}\|_2^2 \geq g(\mathbf{x}) - \frac{\mu}{2}\|\mathbf{x}\|_2^2 + \langle \nabla g(\mathbf{x}) - \mu\mathbf{x}, \mathbf{y} - \mathbf{x} \rangle, \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \\
\Leftrightarrow\ & \phi(\mathbf{y}) \geq \phi(\mathbf{x}) + \langle \nabla \phi(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle, \forall\, \mathbf{x}, \mathbf{y} \in \mathbb{R}^n, \\
\Leftrightarrow\ & \phi(\mathbf{x}) \text{ is convex}, \\
\Leftrightarrow\ & \nabla^2 \phi(\mathbf{x}) \text{ is positive semidefinite}, \\
\Leftrightarrow\ & \nabla^2 g(\mathbf{x}) - \mu\mathbf{I} \text{ is positive semidefinite}.
\end{aligned}$$

2. Let $\phi(\mathbf{x}) = \|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2$, then $\nabla^2 \phi(\mathbf{x}) = 2\mathbf{A}^\top \mathbf{A}$.

First, we show that $\phi(\mathbf{x})$ is strongly convex. $\mathbf{A}^\top\mathbf{A}$ is positive definite as

$$\begin{aligned} &\mathbf{x} \neq 0 \\ \Rightarrow &\mathbf{A}\mathbf{x} \neq 0 \quad (\mathbf{rank}(\mathbf{A}) = d) \\ \Rightarrow &\mathbf{x}^\top\mathbf{A}^\top\mathbf{A}\mathbf{x} = \|\mathbf{A}\mathbf{x}\|_2^2 > 0. \end{aligned}$$

Thus,

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

where $\mathbf{\Sigma} = (\lambda_1, \lambda_2, ..., \lambda_n), \lambda_1 \geq \lambda_2 \geq ... \geq \lambda_n > 0$ and $\mathbf{U}^\top\mathbf{U} = \mathbf{I}$. Therefore, $2\mathbf{A}^\top\mathbf{A} - 2\lambda_n\mathbf{I} = 2\mathbf{U}(\lambda_1 - \lambda_n, \lambda_2 - \lambda_n, ..., 0)\mathbf{U}^\top$ is postive semidefinite and $\phi(\mathbf{x})$ is strongly convex.

Next, we show that the strongly convex function $\phi(\mathbf{x})$ admits a unique solution.

Existence: We have

$$\lim_{\|\mathbf{x}\| \to +\infty} \phi(\mathbf{x}) = +\infty,$$

as

$$\phi(\mathbf{x}) \geq \phi(\mathbf{0}) + \langle \nabla\phi(\mathbf{0}), x \rangle + \lambda_n \|x\|_2^2.$$

Thus, there exists $R > 0$ such that $\phi(\mathbf{x}) \geq \phi(\mathbf{0})$ for all $\mathbf{x} \notin B_R(\mathbf{0})$, where $B_R(\mathbf{x}) = \{\mathbf{y} : \|\mathbf{y} - \mathbf{x}\| \leq R\}$. As $B_R(0)$ is compact, the continuous function $\phi(\mathbf{x})$ attains the minimum in $B_R(\mathbf{0})$, i.e.,

$$\exists \mathbf{x}^* \in B_R(\mathbf{0}), \text{ s.t. } \phi(\mathbf{x}^*) = \min_{\mathbf{x} \in B_R(\mathbf{0})} \phi(\mathbf{x}). \quad \text{(Weierstrass's Theorem)}$$

For $\mathbf{x} \notin B_R(\mathbf{0})$. we have

$$\phi(\mathbf{x}) \geq \phi(\mathbf{0}) \geq \phi(\mathbf{x}^*)$$

Therefore, $\mathbf{x}^*$ is a global minimum point.

Uniqueness: Suppose that there exist $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$ such that $\phi(\mathbf{a}) = \phi(\mathbf{b}) = \min_{\mathbf{x}} \phi(\mathbf{x})$ and $\mathbf{a} \neq \mathbf{b}$. Thus $\nabla\phi(\mathbf{a}) = \mathbf{0}$ and

$$\phi(\mathbf{b}) \geq \phi(\mathbf{a}) + \lambda_n \|\mathbf{b} - \mathbf{a}\|_2^2 > \phi(\mathbf{a}),$$

which leads to a contradiction.

3. To find these cases, let

$$\begin{aligned} &\|\mathbf{y} - \mathbf{A}\mathbf{x}\|_2^2 \geq \|\mathbf{y}\|_2^2 \\ \Leftrightarrow &\|\mathbf{y}\|_2^2 + \|\mathbf{A}\mathbf{x}\|_2^2 - 2\langle \mathbf{y}, \mathbf{A}\mathbf{x} \rangle = \|\mathbf{y}\|_2^2 \\ \Leftrightarrow &\|\mathbf{A}\mathbf{x}\|_2^2 \geq 2\langle \mathbf{y}, \mathbf{A}\mathbf{x} \rangle = 2\langle \mathbf{A}^\top\mathbf{y}, \mathbf{x} \rangle. \end{aligned} \tag{5}$$

If all columns of $\mathbf{A}$ are orthogonal to $\mathbf{y}$, i.e.,

$$\mathbf{A}^\top \mathbf{y} = \mathbf{0},$$

then the inequality (5) holds for all $\mathbf{x} \in \mathbb{R}^d$.

■

**Problem 4: Gradient Descent** 30pts

Consider the following problem

$$\min_{\mathbf{x}} f(\mathbf{x}) \tag{6}$$

where the function $f$ is convex and its gradient is Lipschitz continuous with constant $L > 0$. Suppose that $f$ can attain its minimum.

Consider the sequence $\{\mathbf{x}_k\}_{k=0}^{\infty}$ generated by

$$\mathbf{x}_k = \begin{cases} \mathbf{x}_{k-1} - \alpha \nabla f(\mathbf{x}_{k-1}), & \text{if } k \geq 1, \\ \mathbf{0}, & \text{if } k = 0, \end{cases} \tag{7}$$

where $\alpha \in \left(0, \frac{1}{L}\right]$. Suppose that $\mathbf{x}^*$ is an optimal solution to the problem (6). Show that

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{1}{2\alpha k} \|\mathbf{x}^*\|_2^2.$$

**Solution:** The convexity of $f$ implies

$$\begin{aligned} f(\mathbf{x}_k) - f(\mathbf{x}^*) &\leq \langle \nabla f(\mathbf{x}_k), \mathbf{x}_k - \mathbf{x}^* \rangle \\ &= \frac{1}{2\alpha} \left( \alpha^2 \|\nabla f(\mathbf{x}_k)\|_2^2 + \|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_k - \mathbf{x}^* - \alpha \nabla f(\mathbf{x}_k)\|_2^2 \right) \qquad (8) \\ &= \frac{\alpha}{2} \|\nabla f(\mathbf{x}_k)\|_2^2 + \frac{1}{2\alpha} \left( \|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2 \right) \qquad (9) \end{aligned}$$

where the equation (8) comes from $2\langle \mathbf{a}, \mathbf{b} \rangle = \|\mathbf{a}\|_2^2 + \|\mathbf{b}\|_2^2 - \|\mathbf{a} - \mathbf{b}\|_2^2$ and the equation (9) comes from the update rule (7).

As its gradient is Lipschitz continuous, we have

$$\begin{aligned} f(\mathbf{y}) &= f(\mathbf{x}) + \int_{\mathbf{x}}^{\mathbf{y}} \langle \nabla f(\mathbf{z}), d\mathbf{z} \rangle \\ &\overset{\mathbf{z}=\mathbf{x}+t(\mathbf{y}-\mathbf{x})}{=} f(\mathbf{x}) + \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})), \mathbf{y} - \mathbf{x} \rangle dt \\ &= f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \langle \nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle dt \\ &\leq f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 \|\nabla f(\mathbf{x} + t(\mathbf{y} - \mathbf{x})) - \nabla f(\mathbf{x})\|_2 \|\mathbf{y} - \mathbf{x}\|_2 dt \\ &= f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \int_0^1 Lt \|\mathbf{y} - \mathbf{x}\|_2^2 dt \\ &= f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y} - \mathbf{x} \rangle + \frac{L}{2} \|\mathbf{y} - \mathbf{x}\|_2^2, \forall \mathbf{x}, \mathbf{y}. \end{aligned}$$

In view of the update rule in (7), we have

$$\begin{aligned} f(\mathbf{x}_{k+1}) &\leq f(\mathbf{x}_k) + \langle \nabla f(\mathbf{x}_k), \mathbf{x}_{k+1} - \mathbf{x}_k \rangle + \frac{L}{2}\|\mathbf{x}_{k+1} - \mathbf{x}_k\|_2^2 \\ &= f(\mathbf{x}_k) - \alpha \|\nabla f(\mathbf{x}_k)\|_2^2 + \frac{L}{2}\alpha^2 \|\nabla f(\mathbf{x}_k)\|_2^2 \\ &= f(\mathbf{x}_k) - \alpha \left(1 - \frac{L}{2}\alpha\right) \|\nabla f(\mathbf{x}_k)\|_2^2. \end{aligned} \tag{10}$$

Combining the inequality (9) and (10) leads to

$$\begin{aligned} f(\mathbf{x}_{k+1}) - f(\mathbf{x}^*) &\leq -\alpha \left(\frac{1}{2} - \frac{\alpha L}{2}\right) \|\nabla f(\mathbf{x}_k)\|_2^2 + \frac{1}{2\alpha} \left(\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2\right) \\ &\leq \frac{1}{2\alpha} \left(\|\mathbf{x}_k - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_{k+1} - \mathbf{x}^*\|_2^2\right), \end{aligned}$$

where the last step comes from $0 < \alpha \leq \frac{1}{L}$. By summing up the above inequality for $i = 0, 1, \ldots, k-1$, we have

$$\begin{aligned} k\left(f(\mathbf{x}_k) - f(\mathbf{x}^*)\right) &\leq \sum_{i=0}^{k-1} f(\mathbf{x}_{i+1}) - f(\mathbf{x}^*) \\ &\leq \frac{1}{2\alpha} \left(\|\mathbf{x}_0 - \mathbf{x}^*\|_2^2 - \|\mathbf{x}_k - \mathbf{x}^*\|_2^2\right) \\ &\leq \frac{1}{2\alpha} \left(\|\mathbf{x}^*\|_2^2\right), \end{aligned}$$

where the first step comes from the inequality (8). Therefore,

$$f(\mathbf{x}_k) - f(\mathbf{x}^*) \leq \frac{1}{2\alpha k}\|\mathbf{x}^*\|_2^2.$$

■