Introduction to Machine Learning

Fall 2025

University of Science and Technology of China

Lecturer: Jie Wang — Homework 2

Posted: Oct. 20, 2025 — Due: Oct. 27, 2025

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

Exercise 1: Projection to a Function Space

1. Suppose X and Y are both random variables defined in the same sample space Ω with finite second-order moment, i.e. $\mathbb{E}[X^2], \mathbb{E}[Y^2] < \infty$.

   (a) Let $L^2(\Omega) = \{Z : \Omega \to \mathbb{R} \mid \mathbb{E}[Z^2] < \infty\}$ be the set of random variables with finite second-order moment. Please show that $L^2(\Omega)$ is a linear space, and $\langle X, Y \rangle := \mathbb{E}[XY]$ defines an inner product in $L^2(\Omega)$. Then find the projection of Y on the subspace of $L^2(\Omega)$ consisting of all constant variables.

   (b) Please find a real constant $\hat{c}$, such that

   $$\hat{c} = \operatorname*{\mathbf{argmin}}_{c \in \mathbb{R}} \mathbb{E}[(Y - c)^2].$$

   [Hint: you can solve it by completing the square.]

   (c) Please find the necessary and sufficient condition where $\min_{c \in \mathbb{R}} \mathbb{E}[(Y - c)^2] = \mathbb{E}[Y^2]$. Then give it a geometric interpretation using inner product and projection.

2. Suppose X and Y are both random variables defined in the same sample space Ω and all the expectations exist in this problem. Consider the problem

   $$\min_{f:\mathbb{R}\to\mathbb{R}} \mathbb{E}[(f(X) - Y)^2].$$

   (a) Please solve the above problem by completing the square.

   (b) We let $\mathcal{C}(X)$ denote the subspace $\{f(X) \mid f(\cdot) : \mathbb{R} \to \mathbb{R}, \mathbb{E}[f(X)^2] < \infty\}$ of $L^2(\Omega)$. Please show that the solution of the above problem is the projection of Y on $\mathcal{C}(X)$.

   (c) Please show that question 1 is a special case of question 2. Please give a geometric interpretation of conditional expectation.

Exercise 2: Weighted Least Squares

Consider a dataset $\mathcal{D} = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ where $\mathbf{x}_i \in \mathbb{R}^D$ and $y_i \in \mathbb{R}$. Let $\bar{\mathbf{x}}_i = (1, x_{i,1}, \ldots, x_{i,D})^\top$ and define the design matrix

$$\mathbf{X} = \begin{pmatrix} \bar{\mathbf{x}}_1^\top \\ \bar{\mathbf{x}}_2^\top \\ \vdots \\ \bar{\mathbf{x}}_n^\top \end{pmatrix} = \begin{pmatrix} 1 & x_{1,1} & \cdots & x_{1,D} \\ 1 & x_{2,1} & \cdots & x_{2,D} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n,1} & \cdots & x_{n,D} \end{pmatrix} \in \mathbb{R}^{n \times (D+1)}.$$

Let $\mathbf{y} = (y_1, \ldots, y_n)^\top$. Given positive weights $w_i > 0$, define $\mathbf{W} = \mathrm{diag}(w_1, \ldots, w_n)$ and $\mathbf{W}^{1/2} = \mathrm{diag}(\sqrt{w_1}, \ldots, \sqrt{w_n})$.

1. Consider the weighted least squares (WLS) objective

$$L_{\mathrm{WLS}}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n w_i \, (y_i - \bar{\mathbf{x}}_i^\top \mathbf{w})^2 = \frac{1}{n} \|\mathbf{W}^{1/2}(\mathbf{y} - \mathbf{X}\mathbf{w})\|_2^2.$$

   Derive the first–order optimality condition and show that, if $\mathbf{X}^\top \mathbf{W} \mathbf{X}$ is invertible,

$$\hat{\mathbf{w}}_{\mathrm{WLS}} = (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W} \mathbf{y}.$$

2. (Weighted projection) Recall that the Euclidean (orthogonal) projection of $\mathbf{x} \in \mathbb{R}^n$ onto $\mathcal{C}(\mathbf{A})$, which we introduced in Homework 1, is

$$\mathbf{P}_{\mathbf{A}}^{(2)}(\mathbf{x}) := \underset{\mathbf{z} \in \mathcal{C}(\mathbf{A})}{\mathbf{argmin}} \ \|\mathbf{x} - \mathbf{z}\|_2.$$

   (a) Show that, in general, the matrix

$$\mathbf{P}_W := \mathbf{X} \, (\mathbf{X}^\top \mathbf{W} \mathbf{X})^{-1} \mathbf{X}^\top \mathbf{W}$$

   is *not* the Euclidean orthogonal projector onto $\mathcal{C}(\mathbf{X})$ (typically $\mathbf{P}_W^\top \neq \mathbf{P}_W$ unless $\mathbf{W} = \mathbf{I}$), although $\mathbf{P}_W^2 = \mathbf{P}_W$.

   (b) Define the W-norm $\|\mathbf{u}\|_{\mathbf{W}} := \sqrt{\mathbf{u}^\top \mathbf{W} \mathbf{u}}$ and the W-inner product $\langle \mathbf{u}, \mathbf{v} \rangle_{\mathbf{W}} := \mathbf{u}^\top \mathbf{W} \mathbf{v}$. Show that

$$\mathrm{P}_{\mathbf{X}}^{(W)}(\mathbf{y}) := \underset{\mathbf{z} \in \mathcal{C}(\mathbf{X})}{\mathbf{argmin}} \ \|\mathbf{y} - \mathbf{z}\|_{\mathbf{W}} = \mathbf{P}_W \mathbf{y},$$

   and that $\mathbf{P}_W$ is the W-orthogonal projector: $\mathbf{P}_W^2 = \mathbf{P}_W$ and $\mathbf{P}_W^\top \mathbf{W} = \mathbf{W} \mathbf{P}_W$.

Exercise 3: 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 $\lambda_k \approx 0$?

Exercise 4: 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. (Optional) 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.

Exercise 5: Maximum Likelihood with Laplace Noise (L1 Regression)

Consider the linear observation model

$$b_i \;=\; \mathbf{a}_i^\top \mathbf{x} + \varepsilon_i, \qquad i = 1, 2, \ldots, m,$$

where $\mathbf{a}_i \in \mathbb{R}^d$ and $b_i \in \mathbb{R}$ are observed. Assume the noises $\{\varepsilon_i\}$ are i.i.d. and follow the Laplace distribution with density

$$p(z) = \frac{1}{2\lambda} \exp\!\left(-\frac{|z|}{\lambda}\right), \qquad \lambda > 0.$$

Let $\mathbf{A} \in \mathbb{R}^{m \times d}$ be the matrix whose i-th row is $\mathbf{a}_i^\top$, $\mathbf{b} = (b_1, \ldots, b_m)^\top$, and $\mathbf{x} \in \mathbb{R}^d$ is the parameter vector to be estimated.

1. Write down the likelihood $L(\mathbf{x}, \lambda \mid \mathbf{A}, \mathbf{b})$ and the log-likelihood $\log L(\mathbf{x}, \lambda \mid \mathbf{A}, \mathbf{b})$ (you may drop additive constants independent of $(\mathbf{x}, \lambda)$).

2. Show that the maximum likelihood estimator of $\mathbf{x}$ solves the *L1 regression* problem

$$\hat{\mathbf{x}} \;\in\; \operatorname*{\mathbf{argmin}}_{\mathbf{x} \in \mathbb{R}^d} \sum_{i=1}^{m} \big|\, b_i - \mathbf{a}_i^\top \mathbf{x} \,\big|. \tag{1}$$

3. **Joint MLE when λ is unknown.** In practice, the scale parameter λ is unknown. Treat both $\mathbf{x}$ and λ as unknown parameters.

   (a) For a fixed $\mathbf{x}$, maximize the log-likelihood with respect to $\lambda > 0$ and derive the closed form

   $$\lambda^\star(\mathbf{x}) = \frac{1}{m} \sum_{i=1}^{m} \big|\, b_i - \mathbf{a}_i^\top \mathbf{x} \,\big|.$$

   (b) Show that the joint MLE reduces to the same L1 problem as Problem (1), i.e.,

   $$\big(\hat{\mathbf{x}}, \hat{\lambda}\big) \in \operatorname*{\mathbf{argmin}}_{\mathbf{x} \in \mathbb{R}^d,\, \lambda > 0} \left\{ m \log(2\lambda) + \tfrac{1}{\lambda} \sum_{i=1}^{m} \big| b_i - \mathbf{a}_i^\top \mathbf{x} \big| \right\} \quad \Longleftrightarrow \quad \hat{\mathbf{x}} \in \operatorname*{\mathbf{argmin}}_{\mathbf{x}} \sum_{i=1}^{m} \big| b_i - \mathbf{a}_i^\top \mathbf{x} \big|,$$

   with $\hat{\lambda} = \lambda^\star(\hat{\mathbf{x}})$.

4. **Optimization form (LP).** Show that Problem (1) is equivalent to the linear program

$$\min_{\mathbf{x} \in \mathbb{R}^d,\, \mathbf{t} \in \mathbb{R}^m} \sum_{i=1}^{m} t_i \quad \text{s.t.} \quad -t_i \le b_i - \mathbf{a}_i^\top \mathbf{x} \le t_i, \;\; t_i \ge 0, \;\; i = 1, \ldots, m.$$

Exercise 6: Bias-Variance Trade-off (Programming Exercise)

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.)

References

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