Introduction to Machine Learning

Fall 2024

University of Science and Technology of China

Lecturer: Jie Wang | Homework 2

Posted: Oct. 15, 2024 | Due: Oct. 24, 2024

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

Exercise 1: Projection

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

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

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 $\mathbf{V} = (\mathbf{v}_1, \ldots, \mathbf{v}_d)$, and $\mathbf{v}_1, \ldots, \mathbf{v}_d$ are linearly independent.

      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. (Optional) 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. (Optional) 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.

Exercise 2: Projection to a Matrix Space

Let $\mathbb{R}^{n\times n}$ be the linear space of $n \times n$ matrices. The inner product in this space is defined as

$$\langle A, B\rangle = \mathrm{tr}(A^T B).$$

1. Show that the set of diagonal matrices in $\mathbb{R}^{n\times n}$ forms a linear space. Besides, please find the the projection of any matrix onto the space of diagonal matrices.

2. Prove that the set of symmetric matrices, denoted S^n, in $\mathbb{R}^{n\times n}$ forms a linear space. Also, determine the dimension of this linear space.

3. Show that the inner product of any symmetric matrix and skew-symmetric matrix is zero. Moreover, prove that any matrix can be decomposed as the sum of a symmetric matrix and a skew-symmetric matrix.

4. Find the projection of any matrix onto the space of symmetric matrices.

Exercise 3: 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

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

   [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 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.,

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

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 $\lambda_k \approx 0$?

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. (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 6: High-Dimensional Linear Regression for Image Warping (Programming Exercise)

Consider a data set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$ where $\mathbf{x}_i \in \mathbb{R}^n, \mathbf{y}_i \in \mathbb{R}^m,\ i = 1,\ 2,\ \cdots N$, we want to find a map $\phi : \mathbb{R}^n \to \mathbb{R}^m$ such that $\phi(\mathbf{x}_i) = \mathbf{y}_i,\ i = 1,\ 2,\ \cdots N$. Now given a set of basis functions $\phi_i : \mathbb{R}^n \to \mathbb{R}$

$$\phi_i(\mathbf{x}) = \left(\|\mathbf{x} - \mathbf{x}_i\|_2^2 + r^2\right)^{\frac{\mu}{2}}$$

where μ, r are costume constants. We can approximate function ϕ by a $\hat{\phi}(\mathbf{x})$:

$$\hat{\phi}(\mathbf{x}) = \mathbf{A}\mathbf{x} + \mathbf{b} + \mathbf{W}\boldsymbol{\phi}(\mathbf{x})$$

which is a linear combination of basis functions, where $\boldsymbol{\phi}(\mathbf{x}) := \begin{pmatrix}\phi_1(\mathbf{x}) & \phi_2(\mathbf{x}) & \dots & \phi_N(\mathbf{x})\end{pmatrix}^T \in \mathbb{R}^N$, and parameters $\mathbf{W} \in \mathbb{R}^{m\times N}, \mathbf{A} \in \mathbb{R}^{m\times n}, \mathbf{b} \in \mathbb{R}^m$.

1. Find $\mathbf{W}, \mathbf{A}, \mathbf{b}$ such that

$$\min_{\mathbf{W},\mathbf{A},\mathbf{b}} l := \sum_{i=1}^N \left\|\hat{\phi}(\mathbf{x}_i) - \mathbf{y}_i\right\|_2^2 + \lambda_1\|\mathbf{A} - \mathbf{I}\|_f^2 + \lambda_2\|\mathbf{b}\|_2^2 + \lambda_3\|\mathbf{W}\|_f^2 \tag{1}$$

   where $\|\cdot\|_f$ is the Frobenius norm and $\lambda_1, \lambda_2, \lambda_3 \in \mathbb{R}_+$.

   Hint (1) Take the partial derivatives with respect to $\mathbf{W}, \mathbf{A}, \mathbf{b}$, and set them to zero. (2) $\partial_{\mathbf{X}}\|\mathbf{X} - \mathbf{C}\|_f^2 = 2(\mathbf{X} - \mathbf{C})$.

2. Image warping is a technique used to smoothly distort or reshape an image based on specified transformation rules (shown in Figure 1). The user defines these rules by selecting points on the image, with each transformation mapping an initial position $\mathbf{x}_i$ to a target position $\mathbf{y}_i$. These point pairs form the training set $\mathcal{D} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^N$, where $\mathbf{x}_i, \mathbf{y}_i \in \mathbb{R}^2$. A linear function $\hat{\phi}$ is then learned by minimizing the loss l over the training set. This function is applied to every pixel $\hat{\mathbf{x}}_\mathbf{i}$ in the test image, mapping it to an output coordinate $\hat{\mathbf{y}}_\mathbf{i} = \hat{\phi}(\hat{\mathbf{x}}_\mathbf{i})$. Finally, the pixel value at $\hat{\mathbf{y}}_\mathbf{i}$ replaces that at $\hat{\mathbf{x}}_\mathbf{i}$, generating the warped image.

   Now let $r = 0.5, \mu = 1, \lambda_1 = \lambda_2 = \lambda_3 = 1e - 2$. Please implement the image warping method in the provided framework. You can submit any image you like.

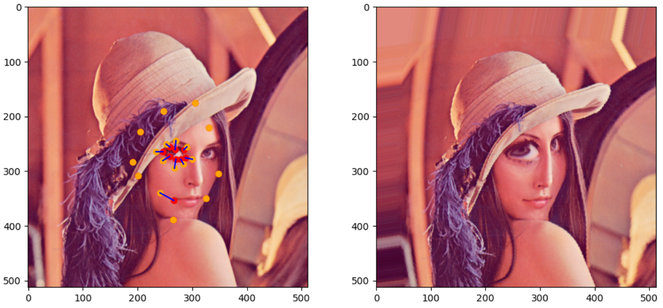

Figure 1: Image warping example

Exercise 7: 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.