Home Assignment 4 Solutions

STAT 154/254: Modern Statistical Prediction & Machine Learning

Ajay Sharma — Fall 2024

Problem 1: Bias and variance for linear ridge regression.

Consider the linear model in which

$$y_i = x_i^\top \beta + \varepsilon_i \quad \text{for } i \in [n],$$

where $\beta \in \mathbb{R}^d$, $x_i \in \mathbb{R}^d$, and $\varepsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. The linear ridge regression estimator gives

$$\hat{\beta} = (X^\top X + \lambda I_d)^{-1} X^\top y,$$

where $y = [y_1, \ldots, y_n]^\top \in \mathbb{R}^n$, and $X = [x_1, \ldots, x_n]^\top \in \mathbb{R}^{n \times d}$. Calculate the expression for $\mathbb{E}\|\mathbb{E}[\hat{\beta}] - \beta\|_2^2$, $\mathbb{E}\|\hat{\beta} - \mathbb{E}[\hat{\beta}]\|_2^2$, and $\mathbb{E}\|\hat{\beta} - \beta\|_2^2$, where the expectation is with respect to the randomness of $\{\varepsilon_i\}_{i \in [n]}$.

Solution. We just need to use the explicit ridge regression solution, the formula $y = X\beta + \varepsilon$, and some linear algebra. To start, we can compute:

$$\begin{aligned}
\widehat{\beta}_\lambda &= (X^\top X + \lambda I)^{-1} X^\top y \\
&= (X^\top X + \lambda I)^{-1} X^\top (X\beta + \varepsilon) \\
&= (X^\top X + \lambda I)^{-1} X^\top X \, \beta + (X^\top X + \lambda I)^{-1} X^\top \varepsilon.
\end{aligned}$$

Taking expectation and using linearity and $\mathbb{E}[\varepsilon] = 0$, we get

$$\mathbb{E}[\widehat{\beta}_\lambda] = (X^\top X + \lambda I)^{-1} X^\top X \, \beta.$$

Thus the bias term is

$$\begin{aligned}
\mathbb{E}[\widehat{\beta}_\lambda] - \beta &= (X^\top X + \lambda I)^{-1} X^\top X \, \beta - \beta \\
&= \left[(X^\top X + \lambda I)^{-1} X^\top X - I\right] \beta \\
&= (X^\top X + \lambda I)^{-1} (X^\top X + \lambda I - \lambda I - (X^\top X)) \, \beta \\
&= (X^\top X + \lambda I)^{-1} \big(\lambda I\big) \beta \\
&= (X^\top X + \lambda I)^{-1} \, \lambda \, \beta.
\end{aligned}$$

Hence

$$\big\|\mathbb{E}[\widehat{\beta}_\lambda] - \beta\big\|_2^2 = \big\|\lambda \, (X^\top X + \lambda I)^{-1} \beta\big\|_2^2. \tag{1}$$

For the variance term, note

$$\widehat{\beta}_\lambda - \mathbb{E}[\widehat{\beta}_\lambda] = (X^\top X + \lambda I)^{-1} X^\top \varepsilon.$$

Hence

$$\begin{aligned}\mathbb{E}\big\|\widehat{\beta}_\lambda - \mathbb{E}[\widehat{\beta}_\lambda]\big\|_2^2 &= \mathbb{E}\big\|\,(X^\top X + \lambda I)^{-1} X^\top \varepsilon\big\|_2^2 \\ &= \mathbb{E}\Big[\varepsilon^\top X\,(X^\top X + \lambda I)^{-1}(X^\top X + \lambda I)^{-1} X^\top \varepsilon\Big].\end{aligned}$$

Using $\mathbb{E}[\varepsilon^\top A^\top A\,\varepsilon] = \sigma^2\,\mathrm{tr}(A^\top A)$ for any A, we get

$$\mathbb{E}\big\|\widehat{\beta}_\lambda - \mathbb{E}[\widehat{\beta}_\lambda]\big\|_2^2 = \sigma^2\;\mathrm{tr}\big((X^\top X + \lambda I)^{-2} X^\top X\big).$$

By the bias–variance decomposition, the total mean-squared error is

$$\mathbb{E}\|\widehat{\beta}_\lambda - \beta\|_2^2 = \mathbb{E}\big\|\widehat{\beta}_\lambda - \mathbb{E}[\widehat{\beta}_\lambda]\big\|_2^2 + \big\|\mathbb{E}[\widehat{\beta}_\lambda] - \beta\big\|_2^2.$$

Problem 2: RKHS inner product and norm.

In this problem, we assume all infinite summations are convergent. In other words, you can treat all summations as finite sums, for index j running from 1 to a finite number m.

Let $\{\varphi_j : \mathbb{R}^d \to \mathbb{R}\}_{j\geq 1}$ be a sequence of linearly independent feature maps (linear independence means $\sum_{j\geq 1} c_j\varphi_j(x) = 0$ for any $x \in \mathbb{R}^d$ implies $c_j = 0$ for any $j \geq 1$). Denote the kernel $k(x, z) = \sum_{j\geq 1} \varphi_j(x)\varphi_j(z)$. For any functions $f(x) = \sum_{j\geq 1} a_j\varphi_j(x)$ and $g(x) = \sum_{j\geq 1} b_j\varphi_j(x)$, we denote its RKHS inner-product by

$$\langle f, g\rangle_{\mathcal{H}} = \sum_{j\geq 1} a_j b_j,$$

and the RKHS norm of f by

$$\|f\|^2_{\mathcal{H}} = \sum_{j\geq 1} a_j^2.$$

By the linear independence of $\{\varphi_j\}$, such inner-product and norm are uniquely defined.

- For any $p \in \mathbb{R}^d$, $k(p, \cdot)$ can be treated as a function on $\mathbb{R}^d$, which maps $x \mapsto k(p, x)$.
- For any $q \in \mathbb{R}^d$, $k(\cdot, q)$ can be treated as a function on $\mathbb{R}^d$, which maps $x \mapsto k(x, q)$.

Consider the following exercises:

i. Show that for any f which can be expressed as a linear combination of $\{\varphi_j\}_{j\geq 1}$, we have $\langle f,\, k(\cdot, q)\rangle_{\mathcal{H}} = f(q)$.

ii. Show that for any $p, q \in \mathbb{R}^d$, we have $\langle k(p, \cdot),\, k(\cdot, q)\rangle_{\mathcal{H}} = k(p, q)$.

iii. Show that suppose $g(x) = \sum_{i=1}^n \alpha_i k(x, x_i)$. For any $x \in \mathbb{R}^d$, we have $\|g\|^2_{\mathcal{H}} = \sum_{i,j=1}^n \alpha_i\alpha_j\, k(x_i, x_j)$.

Solution. First, if

$$f = \sum_{j\geq 1} \alpha_j\varphi_j,$$

then

$$\begin{aligned}\langle f,\, k(\cdot, q)\rangle_{\mathcal{H}} &= \Big\langle \sum_{j\geq 1} \alpha_j\varphi_j,\ \sum_{j\geq 1} \varphi_j(q)\,\varphi_j \Big\rangle_{\mathcal{H}} \\ &= \sum_{j\geq 1} \alpha_j\,\varphi_j(q) \\ &= f(q).\end{aligned}$$

Second, if $p, q \in \mathbb{R}^d$ then by the same reasoning with $f = k(p, \cdot)$,

$$\langle k(p, \cdot),\, k(\cdot, q)\rangle_{\mathcal{H}} = k(p, q).$$

Third, if

$$g = \sum_{i=1}^n \alpha_i\, k(\cdot, x_i),$$

Then by linearity and the above result,

$$\begin{aligned}
\|g\|_{\mathcal{H}}^2 = \langle g, g \rangle_{\mathcal{H}} &= \left\langle \sum_{i=1}^{n} \alpha_i k(\cdot, x_i),\ \sum_{j=1}^{n} \alpha_j k(\cdot, x_j) \right\rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \, \langle k(\cdot, x_i),\ k(\cdot, x_j) \rangle_{\mathcal{H}} \\
&= \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j \, k(x_i, x_j).
\end{aligned}$$

Problem 3: Describe the Bootstrap procedure.

Let $\mathbb{P}_Z$ be a distribution on the real line, with mean $\mu = \mathbb{E}_{Z\sim\mathbb{P}_Z}[Z]$ and variance $\tau = \mathbb{E}_{Z\sim\mathbb{P}_Z}[(Z-\mu)^2]$. Let $z_1, \ldots, z_n \overset{\text{iid}}{\sim} \mathbb{P}_Z$. We define

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} z_i, \quad \hat{\tau} = \frac{1}{n}\sum_{i=1}^{n}(z_i - \hat{\mu})^2$$

as the estimator of μ and τ. Please describe the steps of using the Bootstrap method to estimate the variance of the estimator $\hat{\tau}$.

Requirement of the description (please be as concrete as possible):

- In the first step, describe how the Bootstrap dataset is generated. Please use $(z_i^{(k)})_{i\in[n]}$ to denote the k-th Bootstrap dataset, and denote the number of Bootstrap copies as B.
- In the second step, describe what are the intermediate quantities $\{\hat{\tau}^{(k)}\}_{k\in[B]}$, writing down their mathematical definition using $(z_i^{(k)})_{i\in[n]}$ (you may find it helpful to define intermediate quantities $\hat{\mu}^{(k)}$).
- In the last step, write down the mathematical formula for the Bootstrap estimator $\mathrm{Var}(\hat{\tau})$ using $\{\hat{\tau}^{(k)}\}_{k\in[B]}$.

Solution. First, for $k \in \{1, \ldots, B\}$ we sample with replacement from $\{z_i\}_{i=1}^n$ to get the bootstrap sample $\{z_i^{(k)}\}_{i=1}^n$.

Second, for each $k \in \{1, \ldots, B\}$ we compute

$$\hat{\mu}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} z_i^{(k)}, \qquad \hat{\tau}^{(k)} = \frac{1}{n}\sum_{i=1}^{n}\big(z_i^{(k)} - \hat{\mu}^{(k)}\big)^2.$$

Finally, we compute

$$\bar{\tau} = \frac{1}{B}\sum_{k=1}^{B}\hat{\tau}^{(k)}, \qquad \widehat{\mathrm{Var}}(\hat{\tau}) = \frac{1}{B}\sum_{k=1}^{B}\big(\hat{\tau}^{(k)} - \bar{\tau}\big)^2.$$