Theory of Machine Learning

Exercise sheet 6 — Session 6

Exercise I (checking the maths) 💻. In this exercise, we want to illustrate the decomposition of the ridge excess risk which we obtained in the lecture slide 87. Consider vector-valued inputs and real-valued outputs ($\mathcal{X} = \mathbb{R}^d$ and $\mathcal{Y} = \mathbb{R}$) with $X := (X_1, \ldots, X_n)^\top \in \mathbb{R}^{n\times d}$ the input vector and $Y := (Y_1, \ldots, Y_n)^\top \in \mathbb{R}^n$ the response vector. Let $\phi(x) = (x_1, \ldots, x_d)^\top$ and $\Phi \in \mathbb{R}^{n\times d}$ the matrix of inputs with row i defined as $\Phi_{i,:} := \phi(X_i)^\top$. We work in the fixed design setting where for a fixed input $X \in \mathbb{R}^{n\times d}$, the output is $Y = \Phi\theta^\star + \varepsilon$ (ε i.i.d. $\mathcal{N}(0, \sigma^2)$) and $\theta^\star \in \mathbb{R}^d$.

We set $n = 100$ and $d = 10$, and fix $\theta^\star$ to an arbitrary value. We take i.i.d. $\mathcal{N}(0, \sigma^2)$ noise with small σ.

1. code a function which for any given X, Y and $\lambda > 0$ return the ridge regressor:

$$\hat{\theta}_\lambda = \frac{1}{n}(\widehat{\Sigma} + \lambda\,\mathbf{I}_d)^{-1}\Phi^\top Y\,.$$

2. code a function which estimates the excess risk for a given $\hat{\theta}_\lambda$ defined as:

$$\mathcal{R}(\hat{\theta}_\lambda) - \mathcal{R}^\star = \mathbb{E}_\varepsilon\left[\frac{1}{n}\left\|Y - \Phi\hat{\theta}_\lambda\right\|_2^2\right] - \sigma^2\,.$$

3. Sample uniform training data points X in $[0,1]^d$ and outputs Y according to our assumptions.

4. Make a big loop on λ. For each λ, compute an estimate of excess risk $\mathcal{R}(\hat{\theta}_\lambda) - \mathcal{R}^\star$. What do you observe when you plot the estimated excess risk as a function of lambda? *Hint: the range of λ depends on your problem, beware not to over/undershoot. Bonus: repeat the experiment several times for each lambda to get error bars.*

5. Compute the theoretical bias, variance, and theoretical excess risk as done in the lecture slide 87.

   (a) Compute the theoretical bias b_λ:

$$b_\lambda = \lambda^2(\theta^\star)^\top(\widehat{\Sigma} + \lambda\,\mathbf{I}_d)^{-2}\widehat{\Sigma}\theta^\star\,.$$

   (b) Compute the theoretical variance v_λ:

$$v_\lambda = \frac{\sigma^2}{n}\mathrm{trace}\left(\widehat{\Sigma}^2(\widehat{\Sigma} + \lambda\,\mathbf{I}_d)^{-2}\right)\,.$$

   (c) Compute the theoretical excess risk $\mathbb{E}\left[\mathcal{R}(\hat{\theta}_\lambda)\right] - \mathcal{R}^\star$:

$$\mathbb{E}\left[\mathcal{R}(\hat{\theta}_\lambda)\right] - \mathcal{R}^\star = b_\lambda + v_\lambda\,.$$

   (d) Add them on the previous plot. What do you observe?

6. Add a vertical line corresponding to $\lambda^\star := \frac{\sigma\,\mathrm{trace}(\widehat{\Sigma})^{1/2}}{\|\theta^\star\|\sqrt{n}}$. Is it the best regularization hyper-parameter?

Exercise II (shrinkage) ✎. Assume that $n > d$. Set $\Phi = U\Sigma V^\top$ the singular value decomposition of Φ, and $\sigma_1, \ldots, \sigma_d$ the singular values (which we assume to be positive).

1. Show that, with this notation, the least squares predictions are given by

$$\Phi\hat{\theta} = U J_d U^\top Y \, ,$$

where $J_d = \operatorname{diag}(1, \ldots, 1, 0, \ldots, 0)$ is the $n \times n$ diagonal matrix with d leading 1s on the diagonal.

2. Show that the ridge regression predictions are given by

$$\Phi\hat{\theta}_\lambda = \sum_{j=1}^{d} \frac{\sigma_j^2}{\sigma_j^2 + \lambda} U_{:,j} U_{:,j}^\top Y \, .$$

3. Ridge regression is sometimes classified among the "shrinkage" methods. Explain why.

Exercise III (Expected empirical risk) ✎. Assume that $Y = \Phi\theta^\star + \varepsilon$ where ε is centered and the ε_is are independent, and have common variance σ^2 (assumptions I and II in the lecture).

1. Show that

$$\widehat{R}(\hat{\theta}) = \frac{1}{n} \left\|\Pi\varepsilon\right\|^2 \, ,$$

where $\Pi := \mathbf{I} - \Phi(\Phi^\top\Phi)^{-1}\Phi^\top \in \mathbb{R}^{n\times n}$.

2. Show that

$$\mathbb{E}\left[\widehat{R}(\hat{\theta})\right] = \frac{n-d}{n}\sigma^2 \, .$$

Hint: $\Pi := \mathbf{I} - \Phi(\Phi^\top\Phi)^{-1}\Phi^\top \in \mathbb{R}^{n\times n}$ *is an orthogonal projection matrix.*