Theory of Machine Learning

Exercise sheet 5 — Session 5

Exercise I (degrees of freedom) ✎. As in the lecture, let us set $\delta := \text{trace}\left(\widehat{\Sigma}^2(\widehat{\Sigma}+\lambda\,\mathbf{I})^{-2}\right)$. Let $\mu_1,\dots,\mu_d$ be the eigenvalues of $\widehat{\Sigma}$.

1. Show that, for any $\mu,\lambda>0$, $(\mu+\lambda)^{-2}\mu\lambda \leq 1/2$.
2. Deduce from the previous question the following factoid used in the lecture: all eigenvalues of $\lambda(\widehat{\Sigma}+\lambda\,\mathbf{I})^{-2}\widehat{\Sigma}$ are smaller than $1/2$.
3. Show that δ can be written
$$\sum_{j=1}^{d}\frac{\mu_j^2}{(\mu_j+\lambda)^2}\,.$$
4. What do you think of the following statement: "the degrees of freedom provides a soft count of the number of eigenvalues that are larger than λ."

Exercise II (shrinkage) ✎. Assume that $n>d$. Set $\Phi = U\Sigma V^\top$ the singular value decomposition of Φ, and $\sigma_1,\dots,\sigma_d$ 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 $J_d = \text{diag}\,(1,\dots,1,0,\dots,0)$ 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.*

Exercise IV (checking the maths) 💻. In this exercise, we want to illustrate the decomposition of the ridge excess risk which we obtained in the lecture. We set $n=100$ and $d=10$, $\varphi=$ identity, and fix $\theta^\star$ to an arbitrary value. We take i.i.d. $\mathcal{N}\left(0,\sigma^2\right)$ noise with small σ.

1. code functions which for any given x_is, y_is and $\lambda>0$ return the ridge regressor $\hat{\theta}_\lambda$.
2. code a function which estimates the excess risk for a given $\hat{\theta}_\lambda$.

3. Make a big loop on λ. For each λ, generate uniform training data in $[0, 1]$ and outputs according to our assumptions. Use the previous functions to plot the estimated excess risk as a function of λ. What do you observe? *Hint: the range of λ depends on your problem, beware not to over/undershoot. Bonus: generate several sets of test points to get error bars.*

4. Compute the theoretical bias, variance, and theoretical excess risk. Add them on the previous plot. What do you observe?

5. Add a vertical line corresponding to $\lambda^\star$. Is it the best regularization hyperparameter?