Regularization

Machine Learning 2022-2023 - UMONS
Souhaib Ben Taieb

Exercise 1

Consider the problem of multiple linear regression, where the aim is to find $\hat{\beta}^{LS} = (\beta_0, ..., \beta_p)^\intercal \in \mathbb{R}^{p+1}$ such that:

$$\hat{\beta}^{LS} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \Big)^2.$$

Assuming that $(\mathbf{X}^\intercal\mathbf{X})^{-1}$ is invertible, we can show that the ordinary least squares estimate is given by $\hat{\beta}^{LS} = (\mathbf{X}^\intercal\mathbf{X})^{-1}\mathbf{X}^\intercal\boldsymbol{y}$ where $\mathbf{X} \in \mathbb{R}^{n\times p}$ and $\boldsymbol{y} \in \mathbb{R}^n$. Consider the problem of ridge regression, where the optimization problem is now formulated as finding $\hat{\beta}^{R} = (\beta_0, ..., \beta_p)^\intercal \in \mathbb{R}^{p+1}$ such that:

$$\hat{\beta}^{R} = \underset{\beta}{\operatorname{argmin}} \sum_{i=1}^{n} \Big(y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij} \Big)^2 + \lambda \sum_{j=1}^{p} \beta_j^2,$$

where $\lambda \geq 0$. The solution is given by $\hat{\beta}^{R} = (\mathbf{X}^\intercal\mathbf{X} + \lambda\mathbf{I}_p)^{-1}\mathbf{X}^\intercal\boldsymbol{y}$, where $\mathbf{I}_p$ is the $p \times p$ identity matrix. Let us consider a simple scenario where $p = n$ and that $\mathbf{X} = \mathbf{I}_p$.

- Prove that $\hat{\beta}^{R} = \frac{\hat{\beta}^{LS}}{\lambda+1}$.
- Given that $\mathrm{Bias}(\hat{\beta}^{LS}) = 0$, compute the bias of the ridge estimator.
- Assuming that the data generative process is $\boldsymbol{y} = \mathbf{X}\beta + \boldsymbol{\varepsilon}$ where $\boldsymbol{\varepsilon} \in \mathbb{R}^n$ is a random noise vector, derive the covariance matrices of $\hat{\beta}^{LS}$ and $\hat{\beta}^{R}$ and show how they relate to one another. You can assume that $\mathbb{E}[\boldsymbol{\varepsilon}\boldsymbol{\varepsilon}^T] = \sigma^2\mathbf{I}_n$ and $\mathbb{E}[\boldsymbol{\varepsilon}] = 0$.

The covariance matrix of a random vector $\boldsymbol{a} \in \mathbb{R}^p$ is given by $\mathrm{Cov}(\boldsymbol{a}) = \mathbb{E}\Big[(\boldsymbol{a} - \mathbb{E}[\boldsymbol{a}])(\boldsymbol{a} - \mathbb{E}[\boldsymbol{a}])^\intercal\Big] \in \mathbb{R}^{p\times p}$.

Exercise 2

Suppose that the columns of $\mathbf{X}_1$ are orthonormal, and that $\mathbf{X}_2 = 10\mathbf{X}_1$. Show that the ordinary least squares estimates are equivariant, meaning that multiplying $\mathbf{X}$ by a constant c scales the coefficients estimates by a factor $\frac{1}{c}$. Is it also the case for the ridge estimates?