EPFL, Autumn 2025
Yoav Zemel
Guillaume Obozinski

Ho Yun

Statistical Machine Learning

Exercise sheet 2

Exercise 2.1 (Continuation of Ex 1.1) Let $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, where $\mathbb{E}(\boldsymbol{\varepsilon}) = \mathbf{0}$, $\mathrm{Cov}(\boldsymbol{\varepsilon}) = \sigma^2\mathbf{I}$ and $\mathbf{X}$ is a non-random full rank matrix of size $n \times p$. This setup contains the Gauss-Markov assumptions of a linear model.

(a) Prove the Gauss-Markov theorem, i.e, $\widehat{\boldsymbol{\beta}}$ is the best **linear unbiased** estimator (BLUE) of $\boldsymbol{\beta}$. "Best" in the sense that for all other linear unbiased estimators $\widetilde{\boldsymbol{\beta}}$ of $\boldsymbol{\beta}$, $\mathrm{Cov}(\widetilde{\boldsymbol{\beta}}) - \mathrm{Cov}(\widehat{\boldsymbol{\beta}})$ is a positive semidefinite matrix.

Hints: Recall that an estimator $\widetilde{\boldsymbol{\beta}}$ is linear if $\widetilde{\boldsymbol{\beta}} = \mathbf{A}\boldsymbol{y}$, for some $\mathbf{A} \in \mathbb{R}^{p\times n}$. Notice that the matrix $\mathbf{A}$ can be decomposed as $\mathbf{A} = \mathbf{B} + (\mathbf{X}^\top\mathbf{X})^{-1}\mathbf{X}^\top$.

(b) Assume now that the errors $\boldsymbol{\varepsilon}$ are normally distributed. Prove that $\widehat{\boldsymbol{\beta}}$ is the best estimator among **all unbiased** estimators. $\widehat{\boldsymbol{\beta}}$ is then a uniformly minimum variance unbiased (UMVU) estimator.

Hint: Remember the Cramér–Rao bound.

Exercise 2.2 (The regression function) Recall that we are interested in the predictive model $f^* : \mathbb{R}^p \to \mathbb{R}$ that minimizes the expected error for the ℓ^2 loss. i.e., we want to find the function f^* such that

$$\mathbb{E}[\ell\{Y, f^*(\boldsymbol{X})\}] = \mathbb{E}[\{Y - f^*(\boldsymbol{X})\}^2] = \min_{f:\mathbb{R}^p\to\mathbb{R}} \mathbb{E}[\{Y - f(\boldsymbol{X})\}^2].$$

(a) Show that $f^*(\boldsymbol{x}) = \mathbb{E}(Y|\boldsymbol{X} = \boldsymbol{x})$.

(b) If we consider the ℓ^1 loss instead, i.e., $\ell(y, \widehat{y}) = |y - \widehat{y}|$, what is f^*? (For simplicity suppose that $\mathbb{P}(Y \mid \boldsymbol{X})$ has a density.)

Exercise 2.3 (Bias-variance tradeoff) In this exercise, we consider the expected ℓ^2 error of a random predictive model $\widehat{f}_n$ (depends on a training set $\mathcal{D}_n$), defined as

$$\mathbb{E}\left[\int_{\mathbb{R}^p} \{\widehat{f}_n(\boldsymbol{x}) - f^*(\boldsymbol{x})\}^2 P_{\boldsymbol{X}}(d\boldsymbol{x})\right]. \tag{1}$$

(a) For any random predictive model $\widehat{f}_n$ and any fixed point $\boldsymbol{x}_0 \in \mathbb{R}^p$, prove that

$$\mathbb{E}\Big[\{\widehat{f}_n(\boldsymbol{x}_0) - f^*(\boldsymbol{x}_0)\}^2\Big] = \Big[\mathrm{bias}\{\widehat{f}_n(\boldsymbol{x}_0)\}\Big]^2 + \mathrm{var}\{\widehat{f}_n(\boldsymbol{x}_0)\}.$$

(b) Find a similar bias-variance decomposition for the expected ℓ^2 error (1).

Exercise 2.4 (Ridge regression)

(a) Consider the linear regression model

$$y_i = \beta_0 + \sum_{j=1}^{p} \beta_j x_{ij} + \varepsilon_i, \quad i = 1, \ldots, n.$$

Define $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_p)^\top$ and the residuals as

$$r_i(\beta_0, \boldsymbol{\beta}) = y_i - \beta_0 - \sum_{j=1}^{p} \beta_j x_{ij}, \quad i = 1, \ldots, n.$$

Show that the OLS estimator $\widehat{\beta}_0 = \overline{y} - \sum_{j=1}^{p} \beta_j x_{\cdot j}$ for any $\boldsymbol{\beta}$, where $x_{\cdot j} = \frac{1}{n} \sum_{i=1}^{n} x_{ij}$. Hence deduce that

$$r_i(\widehat{\beta}_0, \boldsymbol{\beta}) = y_i - \overline{y} - \sum_{j=1}^{p} \beta_j (x_{ij} - x_{\cdot j}), \quad i = 1, \ldots, n.$$

Discuss the implications of this result.

(b) Define the ridge regression estimator as a minimizer of the penalized residual sum of squares,

$$\widehat{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}}{\operatorname{argmin}} \ \frac{1}{n} \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda \boldsymbol{\beta}^\top \boldsymbol{\beta}, \tag{2}$$

where $\lambda \geq 0$ is a parameter that controls the amount of shrinkage. Show that the ridge regression solution always exists, even if $\mathbf{X}$ does not have full rank, and is given by

$$\widehat{\boldsymbol{\beta}}(\lambda) = (\mathbf{X}^\top \mathbf{X} + n\lambda \mathbf{I})^{-1} \mathbf{X}^\top \boldsymbol{y}.$$

Note that the ridge estimator is still linearly depending on the response $\boldsymbol{y}$, as for ordinary least squares.

(c) Show that the ridge regression estimator defined in (2) equals

$$\widehat{\boldsymbol{\beta}}(t) = \underset{\|\boldsymbol{\beta}\|^2 \leq t}{\operatorname{argmin}} \ \|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 \tag{3}$$

for a given $t = t(\lambda)$. *Hint: Use the Karush–Kuhn–Tucker (KKT) method.*

Exercise 2.5 The Gauss-Markov Theorem makes the assumption that the training data is generated as $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$, $\mathbf{X}$ is a non-random full rank matrix of size $n \times p$, where $\mathbb{E}[\boldsymbol{\varepsilon}] = \mathbf{0}$, $\operatorname{Cov}(\boldsymbol{\varepsilon}) = \sigma^2 \mathbf{I}$.

(a) Explain why the Gauss-Markov Theorem still holds for any random design matrix $\mathbf{X}$ (in particular without assuming that the rows of $\mathbf{X}$ are i.i.d.) provided we change the assumptions and assume that $\boldsymbol{y} = \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon}$ with $\mathbb{E}[\varepsilon|\mathbf{X}] = \mathbf{0}$, $\operatorname{Cov}(\varepsilon|\mathbf{X}) = \sigma^2 \mathbf{I}$.

(b) Let $\tilde{\boldsymbol{\beta}}$ be any linear unbiased estimator and let $\widehat{\boldsymbol{\beta}}$ be the linear regression estimator (aka ordinary least squares estimator). Show that as a consequence of the Gauss-Markov theorem:

$$\forall \boldsymbol{x} \in \mathbb{R}^p, \qquad \operatorname{Var}(\boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}) \leq \operatorname{Var}(\boldsymbol{x}^\top \tilde{\boldsymbol{\beta}}).$$

(c) Consider now i.i.d. data (X_i, Y_i) with $Y_i = X_i^\top \boldsymbol{\beta} + \varepsilon_i$, $\mathbb{E}[\varepsilon_i | X_i] = 0$ and $\mathrm{Var}(\varepsilon_i | X_i) = \sigma^2$. For data following this distribution, express the target function for the quadratic risk as a function of $\boldsymbol{\beta}$.

(d) Let $\widehat{f} : \boldsymbol{x} \mapsto \boldsymbol{x}^\top \widehat{\boldsymbol{\beta}}$ and $\tilde{f} : \boldsymbol{x} \mapsto \boldsymbol{x}^\top \tilde{\boldsymbol{\beta}}$ for $\tilde{\boldsymbol{\beta}}$ some unbiased linear estimator based on $\mathbf{X}$ and $\boldsymbol{y}$. Show that for any such $\tilde{f}$, if $\mathcal{R}$ denotes the quadratic risk (i.e. the risk associated with the square loss), then we necessarily have $\mathbb{E}[\mathcal{R}(\widehat{f})] \leq \mathbb{E}[\mathcal{R}(\tilde{f})]$. Show that the same inequality actually holds conditionally on the value of any new $X = \boldsymbol{x}$.