EPFL, Autumn 2025
Yoav Zemel
Guillaume Obozinski Ho Yun

Statistical Machine Learning

Exercise sheet 3

Exercise 3.1 (Singular Value Decomposition.) Let $\mathbf{X}$ be a $n \times p$ matrix of rank r.

(a) Show that $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$, where $\mathbf{U}$ and $\mathbf{V}$ are $n \times r$ and $p \times r$ semi-orthogonal matrices respectively, in that $\mathbf{U}^\top\mathbf{U} = \mathbf{V}^\top\mathbf{V} = \mathbf{I}_r$ (while $\mathbf{U}\mathbf{U}^\top$ and $\mathbf{V}\mathbf{V}^\top$ are projections but need not be identity matrices unless $\mathbf{U}$ or $\mathbf{V}$ are square matrices, in which case the definition is equivalent to that of an orthogonal matrix.) and $\mathbf{D}$ is a $r \times r$ diagonal matrix with r non-negative entries which we refer to as the singular values of $\mathbf{X}$. This is sometimes called a *compact* SVD.

Hint: For the special case $r = p$, consider the eigendecomposition of $\mathbf{X}^\top\mathbf{X} = \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top$ and let $\mathbf{U} = \mathbf{X}\mathbf{V}\mathbf{\Lambda}^{-1/2}$.

(b) Show that $\mathbf{X} = \tilde{\mathbf{U}}\tilde{\mathbf{D}}\tilde{\mathbf{V}}^\top$ where $\tilde{\mathbf{U}}$ and $\tilde{\mathbf{V}}$ are $n \times n$ and $p \times p$ orthogonal matrices respectively and $\tilde{\mathbf{D}}$ is a $n \times p$ diagonal matrix with r non-negative entries.

(c) To interpret the result from (b) we need to understand the orthonormal basis $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_p)$ of $\mathbb{R}^p$. For each $j = 1, \dots, p$, calculate the norm of $\mathbf{X}\boldsymbol{v}_j$ and discuss the implication of (b); note that $\mathbf{X}\boldsymbol{v}_j$ are the jth coordinates of each row vector of $\mathbf{X}$ in the basis $(\boldsymbol{v}_1, \dots, \boldsymbol{v}_p)$.

Note: The vectors $\boldsymbol{v}_1, \dots, \boldsymbol{v}_p \in \mathbb{R}^p$ are called the *principal components directions* of $\mathbf{X}$. The vectors $\boldsymbol{u}_1, \dots, \boldsymbol{u}_p \in \mathbb{R}^n$ are the *normalized principal component scores*. We will go back to this in this class when discussing unsupervised learning (PCA).

(d) Use the compact singular value decomposition $\mathbf{U}\mathbf{D}\mathbf{V}^\top$ of $\mathbf{X}$ (full rank) to show that applying the hat matrix $\mathbf{H} = \mathbf{X}\left[\mathbf{X}^\top\mathbf{X}\right]^{-1}\mathbf{X}^\top$ to a vector $\boldsymbol{y}$ projects it onto the subspace spanned by the columns $\{\boldsymbol{u}_j\}$ of $\mathbf{U}$ as in (a). In other words, $\mathbf{H}$ is a projection matrix. Using this result, show that the fitted value of $\boldsymbol{y}$ from ordinary least squares can be written in the following way:

$$\widehat{\boldsymbol{y}}^{\text{ols}} = \sum_{j=1}^{r} \boldsymbol{u}_j \boldsymbol{u}_j^\top \boldsymbol{y}.$$

Exercise 3.2 § (Geometric interpretation of linear ridge regression) In this question, we will assume that the design matrix $\mathbf{X}$ is an $n \times p$ full-rank matrix.

(a) Show that

$$\begin{aligned}\widehat{\boldsymbol{\beta}}^{\text{ridge}} &= \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \frac{1}{n}\|\boldsymbol{y} - \mathbf{X}\boldsymbol{\beta}\|^2 + \lambda\|\boldsymbol{\beta}\|_2^2 \\ &= \mathbf{V}(\mathbf{D}^2 + n\lambda\mathbf{I})^{-1}\mathbf{D}\mathbf{U}^\top\boldsymbol{y}.\end{aligned}$$

(b) Show that the fitted values from ridge regression can be written as

$$\widehat{\boldsymbol{y}}^{\text{ridge}}(\lambda) = \sum_{j=1}^{p} \frac{d_j^2}{d_j^2 + n\lambda} \boldsymbol{u}_j \boldsymbol{u}_j^\top \boldsymbol{y},$$

where the d_j are the diagonal elements of the matrix $\mathbf{D}$ from (b). Discuss.

Exercise 3.3 (Artificial feature noising in linear regression) This exercise is inspired by the paper *Dropout Training as Adaptive Regularization*, by S. Wager, S. Wang, and P. Liang, published in Advances in Neural Information Processing Systems (NIPS), in 2013. In this exercise we focus on the case of feature noising applied to linear regression.

Consider linear regression with no intercept (recall from Exercise 2.4(a) that the intercept can always be estimated in a first step, so there is no loss of generality of assuming it is zero). Let $\boldsymbol{y} \in \mathbb{R}^n$ denote the response vector, and $\mathbf{X} \in \mathbb{R}^{n \times p}$ the design matrix with centered columns. Recall the least-squares estimator of the regression parameter

$$\widehat{\boldsymbol{\beta}}^{\text{ols}} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2.$$

Consider the following artificial feature noising scheme: the features $\boldsymbol{x}_i$ are replaced by the artificial features $\tilde{\boldsymbol{x}}_i = \nu(\boldsymbol{x}_i, \boldsymbol{\xi}_i)$ where ν is a noising function and the $\boldsymbol{\xi}_i$ are independent random vectors from some distribution $P_{\boldsymbol{\xi}}$. In practice, artificial features noising is implemented as follows:

1. For $l = 1, \ldots, n$, generate a large number L of independent $\tilde{\boldsymbol{x}}_{il} = \nu(\boldsymbol{x}_i, \boldsymbol{\xi}_{il})$, where $\boldsymbol{\xi}_{il} \sim P_{\boldsymbol{\xi}}$ are independent copies of $\boldsymbol{\xi}_i$.
2. Estimate the regression parameter by

$$\widehat{\boldsymbol{\beta}}^{L} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} \sum_{l=1}^{L} (y_i - \tilde{\boldsymbol{x}}_{il}^\top \boldsymbol{\beta})^2. \tag{1}$$

The goal of this exercise is to study the link between feature noising and regularization. In theory, we study the limiting behavior of $\widehat{\boldsymbol{\beta}}^{L}$ as $L \to \infty$. We will assume that

$$\widehat{\boldsymbol{\beta}}^{\infty} = \operatorname*{argmin}_{\boldsymbol{\beta} \in \mathbb{R}^p} \sum_{i=1}^{n} \mathrm{E}_{\boldsymbol{\xi}_i} \left\{ (y_i - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta})^2 \right\}.$$

(a) Consider the following additive feature noising scheme: $\tilde{\boldsymbol{x}}_i = \nu(\boldsymbol{x}_i, \boldsymbol{\xi}_i) = \boldsymbol{x}_i + \boldsymbol{\xi}_i$ where $\boldsymbol{\xi}_i \sim \mathcal{N}(\mathbf{0}, \sigma^2 \mathbf{I})$ for some $\sigma^2 > 0$. Show that the resulting estimator $\widehat{\boldsymbol{\beta}}^{\infty}$ is a ridge estimator $\widehat{\boldsymbol{\beta}}^{\text{ridge}}(\lambda)$ for some value of λ in terms of σ^2.

(b) Consider the following dropout noising scheme: $\tilde{\boldsymbol{x}}_i = \nu(\boldsymbol{x}_i, \boldsymbol{\xi}_i) = \boldsymbol{x}_i \cdot \boldsymbol{\xi}_i$, where the operator $\cdot$ denotes the element-wise product of two vectors, and the components ξ_{ij} of the vectors $\boldsymbol{\xi}_i$ are independent random variables with

$$\xi_{ij} = \begin{cases} 0, & \text{with probability } \delta, \\ \dfrac{1}{1-\delta}, & \text{with probability } 1 - \delta, \end{cases}$$

i.e., the ξ_{ij} follow independent scaled Bernoulli distributions, for some parameter $\delta \in (0, 1)$. Show that the resulting estimator $\widehat{\boldsymbol{\beta}}^{\infty}$ is a penalized least-squares estimator, and express the regularization term in terms of δ. What happens when the features are normalized so that $\sum_{i=1}^{n} x_{ij}^2 = 1$ for $j = 1, \ldots, p$?

Practical exercise

Exercise 3.4 (Reading for next week's practical) Read §§6.5,6.6 from ISL. This is to familiarize yourself with the packages and the functions in R you will need for the purpose of next week's exercises.