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 $\mathbf{U} = \mathbf{X}\mathbf{V}\mathbf{\Lambda}^{-1/2}$.

Solution: We first assume that $p = r$. In that case $\mathbf{X}^\top\mathbf{X}$ and $\mathbf{\Lambda}$ are positive semi-definite matrices and we can let $\mathbf{U} := \mathbf{X}\mathbf{V}\mathbf{D}^{-1}$ with $\mathbf{D} := \mathbf{\Lambda}^{1/2}$. We have $\mathbf{U}^\top\mathbf{U} = \mathbf{D}^{-1}\mathbf{V}^\top\mathbf{X}^\top\mathbf{X}\mathbf{V}\mathbf{D}^{-1} = \mathbf{D}^{-1}\mathbf{V}^\top\mathbf{V}\mathbf{D}^2\mathbf{V}^\top\mathbf{V}\mathbf{D}^{-1} = \mathbf{I}$ so that $\mathbf{U}$ is a semi-orthogonal matrix. Since $r = p$, $\mathbf{V}$ is a square semi-orthogonal matrix which entails that it is an orthogonal matrix and we have $\mathbf{U}\mathbf{D}\mathbf{V}^\top := \mathbf{X}\mathbf{V}\mathbf{D}^{-1}\mathbf{D}\mathbf{V}^\top = \mathbf{X}$.

If $r < p$, then we consider again the eigenvalue decomposition of $\mathbf{X}^\top\mathbf{X}$ which we write this time $\mathbf{X}^\top\mathbf{X} = \tilde{\mathbf{V}}^\top\tilde{\mathbf{\Lambda}}\tilde{\mathbf{V}}$. Since $\mathbf{X}$ is of rank r, we have $\tilde{\mathbf{\Lambda}} = \mathrm{Diag}(\lambda_1, \dots, \lambda_r, 0, \dots, 0)$, assuming that the diagonal of $\tilde{\mathbf{\Lambda}}$ is sorted in decreasing order. Still under that assumption, if we let $\mathbf{\Lambda} = \mathrm{Diag}(\lambda_1, \dots, \lambda_r)$ and $\mathbf{V} \in \mathbb{R}^{p \times r}$ be the semi-orthogonal matrix obtained by extracting the r first columns of $\tilde{\mathbf{V}}$, we get that $\mathbf{X}^\top\mathbf{X} = \mathbf{V}^\top\mathbf{\Lambda}\mathbf{V}$. By the previous reasoning, we have that $\mathbf{U} := \mathbf{X}\mathbf{V}\mathbf{D}^{-1}$ is again a semi-orthogonal matrix. We also have that $\mathbf{U}\mathbf{D}\mathbf{V}^\top := \mathbf{X}\mathbf{V}\mathbf{V}^\top = \mathbf{X}$ and so we need to argue that $\mathbf{X}\mathbf{V}\mathbf{V}^\top = \mathbf{X}$ in spite of the fact that we don't have $\mathbf{V}\mathbf{V}^\top = \mathbf{I}_p$ here. It is easy to see that this is true using $\mathbf{V}^\top\mathbf{V} = \mathbf{I}_r$ as follows

$$\begin{aligned}(\mathbf{X}\mathbf{V}\mathbf{V}^\top - \mathbf{X})^\top(\mathbf{X}\mathbf{V}\mathbf{V}^\top - \mathbf{X}) &= \mathbf{V}\mathbf{V}^\top\mathbf{X}^\top\mathbf{X}\mathbf{V}\mathbf{V}^\top - \mathbf{X}^\top\mathbf{X}\mathbf{V}\mathbf{V}^\top - \mathbf{V}\mathbf{V}^\top\mathbf{X}^\top\mathbf{X} + \mathbf{X}^\top\mathbf{X} \\ &= \mathbf{V}(\mathbf{V}^\top\mathbf{V})\mathbf{\Lambda}(\mathbf{V}^\top\mathbf{V})\mathbf{V}^\top - \mathbf{V}\mathbf{\Lambda}(\mathbf{V}^\top\mathbf{V})\mathbf{V}^\top - \mathbf{V}(\mathbf{V}^\top\mathbf{V})\mathbf{\Lambda}\mathbf{V}^\top + \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top \\ &= \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top - \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top - \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top + \mathbf{V}\mathbf{\Lambda}\mathbf{V}^\top = \mathbf{0}_{p\times p}\end{aligned}$$

where $\mathbf{0}_{m\times n}$ denotes the $m \times n$ matrix with all entries equal to zero. Thus, $\mathbf{X}\mathbf{V}\mathbf{V}^\top = \mathbf{X}$ and the conclusion follows.

(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.

Solution: Using the result in (a), we can write $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$. Let $\{\boldsymbol{u}_j\}_{j=1}^r$ and $\{\boldsymbol{v}_j\}_{j=1}^r$ be the columns of $\mathbf{U}$ and $\mathbf{V}$ respectively and $\mathbf{D} = \mathrm{diag}[d_1, \dots, d_r]$ where d_j are corresponding singular values of $\mathbf{X}$. Then we can complete $\{\boldsymbol{u}_j\}_{j=1}^r$ and $\{\boldsymbol{v}_j\}_{j=1}^r$

to orthonormal bases $\{\boldsymbol{u}_j\}_{j=1}^n$ and $\{\boldsymbol{v}_j\}_{j=1}^p$ of the spaces $\mathbb{R}^n$ and $\mathbb{R}^p$ respectively. Let $\mathbf{U}_1 = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_p]$, $\mathbf{D}_1 = \mathrm{diag}[d_1, \ldots, d_p]$ where $d_j = 0$ for $j > r$ and $\tilde{\mathbf{V}} = [\boldsymbol{v}_1 \cdots \boldsymbol{v}_p]$. Then,

$$\mathbf{U}\mathbf{D}\mathbf{V}^\top = \sum_{j=1}^{r} d_j \boldsymbol{u}_j \boldsymbol{v}_j^\top = \sum_{j=1}^{p} d_j \boldsymbol{u}_j \boldsymbol{v}_j^\top = \mathbf{U}_1 \mathbf{D}_1 \tilde{\mathbf{V}}^\top$$

Let $\tilde{\mathbf{U}} = [\boldsymbol{u}_1 \cdots \boldsymbol{u}_n]$ and $\tilde{\mathbf{D}} = [\mathbf{D}_1^\top \ \mathbf{0}_{p \times n-p}]^\top$. It follows that $\tilde{\mathbf{U}}\tilde{\mathbf{D}} = \mathbf{U}_1\mathbf{D}_1$. Therefore, $\mathbf{X} = \tilde{\mathbf{U}}\tilde{\mathbf{D}}\tilde{\mathbf{V}}^\top$. Hence proved.

(c) To interpret the result from (b) we need to understand the orthonormal basis $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$ of $\mathbb{R}^p$. For each $j = 1, \ldots, p$, 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 $(\boldsymbol{v}_1, \ldots, \boldsymbol{v}_p)$.

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

Solution: By noting that $\mathbf{X}\mathbf{V} = \mathbf{U}\mathbf{D}$, we have that $\mathbf{X}\boldsymbol{v}_j = \boldsymbol{u}_j d_j$. Thus, the sample variance of $\mathbf{X}\boldsymbol{v}_j$ is simply given by

$$\frac{1}{n}\|\mathbf{X}\boldsymbol{v}_j\|^2 = \frac{1}{n}\|\boldsymbol{u}_j d_j\|^2 = \frac{1}{n} d_j^2 \boldsymbol{u}_j^\top \boldsymbol{u}_j = \frac{d_j^2}{n}.$$

(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:

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

Solution: By substituting in $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$ into the least squares estimate, it is easy to see that $\widehat{\boldsymbol{y}}^{\mathrm{ols}} = \mathbf{X}\widehat{\boldsymbol{\beta}}^{\mathrm{ols}} = \mathbf{U}\mathbf{U}^\top\boldsymbol{y}$.

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

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

Solution: We minimize the error as in Exercise 1.2 (a), to get

$$\boldsymbol{\beta} = \left[n\lambda\mathbf{I} + \mathbf{X}^\top\mathbf{X}\right]^{-1}\mathbf{X}^\top\boldsymbol{y}.$$

Note that the inversion of $n\lambda\mathbf{I} + \mathbf{X}^\top\mathbf{X}$ is trivial for $\lambda > 0$, so the ridge solution always exists. Then we apply SVD by substituting $\mathbf{X} = \mathbf{U}\mathbf{D}\mathbf{V}^\top$.

(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.

Solution: We see that $\widehat{\boldsymbol{y}}^{\text{ridge}} = \mathbf{X}\widehat{\boldsymbol{\beta}}^{\text{ridge}} = \mathbf{U}\boldsymbol{\Delta}\mathbf{U}^\top \boldsymbol{y}$, where Δ is diagonal with $\Delta_{jj} = \frac{d_j^2}{d_j^2 + n\lambda}$.

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.

Solution: Write

$$\begin{aligned} \mathrm{E}_{\boldsymbol{\xi}_i} \left\{ (y_i - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta})^2 \right\} &= \left\{ \mathrm{E}_{\boldsymbol{\xi}_i} (y_i - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}) \right\}^2 + \mathrm{var}_{\boldsymbol{\xi}_i} (y_i - \tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}) \\ &= (y_i - \boldsymbol{x}_i^\top \boldsymbol{\beta})^2 + \mathrm{var}_{\boldsymbol{\xi}_i} (\tilde{\boldsymbol{x}}_i^\top \boldsymbol{\beta}) \end{aligned}$$

where we used the fact that $\mathrm{E}_{\boldsymbol{\xi}_i}\tilde{\boldsymbol{x}}_i = \boldsymbol{x}_i$, i.e., the noising scheme is unbiased. It follows that

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

We now explicit the regularization term in the case of the additive noising scheme. We have $\mathrm{var}_{\boldsymbol{\xi}_i}(\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta}) = \boldsymbol{\beta}^\top\mathrm{var}_{\boldsymbol{\xi}_i}(\tilde{\boldsymbol{x}}_i)\boldsymbol{\beta} = \sigma^2\boldsymbol{\beta}^\top\boldsymbol{\beta} = \sigma^2\|\boldsymbol{\beta}\|_2^2$, thus the regularization term in (2) is $n\sigma^2\|\boldsymbol{\beta}\|_2^2$, which shows that $\widehat{\boldsymbol{\beta}}^{\infty} = \widehat{\boldsymbol{\beta}}^{\text{ridge}}(\sigma^2)$.

(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 $j = 1, \ldots, p$?

Solution: Note that the variables ξ_{ij} have expectation 1, hence the dropout noising scheme is unbiased and the same argument as in (b) leads to (2). Now, $\mathrm{var}_{\xi_{ij}}(\tilde{x}_{ij}) = \mathrm{var}_{\xi_{ij}}(x_{ij}\xi_{ij}) = x_{ij}^2\frac{\delta}{1-\delta}$, thus $\mathrm{var}_{\boldsymbol{\xi}_i}(\tilde{\boldsymbol{x}}_i^\top\boldsymbol{\beta}) = \frac{\delta}{1-\delta}\sum_{j=1}^p x_{ij}^2\beta_j^2$ and the regularization term in (2) equals

$$\sum_{i=1}^{n}\tfrac{\delta}{1-\delta}\sum_{j=1}^{p}x_{ij}^2\beta_j^2 = \tfrac{\delta}{1-\delta}\sum_{j=1}^{p}\beta_j^2\sum_{i=1}^{n}x_{ij}^2,$$

which is also a ridge regularization when $\sum_{i=1}^n x_{ij}^2 = 1$ (but not otherwise). Note: Dropout regularization is very popular in machine learning (it is widely used in fitting deep neural network models). Feature noising methods are equivalent to regularization but in general (for non-linear models) the regularization term has no closed-form (see the paper of Wager et al. if you want to know more).

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.