EPFL, Autumn 2024
Yoav Zemel
Guillaume Obozinski

Ho Yun
Shivang Sachar

# Statistical Machine Learning

## Exercise sheet 10

We will use several times in this exercise sheet *Von Neumann's inequality.* We will prove it in a last exercise. *Von Neuman's inequality* says that if we let $A, B \in \mathbb{R}^{d\times K}$, with $K \le d$, two matrices whose singular values are respectively $\sigma_1(A) \ge \ldots \ge \sigma_K(A)$ and $\sigma_1(B) \ge \ldots \ge \sigma_K(B)$, then

$$|\mathrm{tr}(A^\top B)| \le \sum_{k=1}^{K} \sigma_k(A)\sigma_k(B).$$

**Exercise 10.1** Proving part of Eckart-Young's theorem...

(a) Use Von Neumann's inequality to show that if $A$ and $B$ are as above then

$$\|A - B\|_F^2 \ge \sum_{k=1}^{K} (\sigma_k(A) - \sigma_k(B))^2.$$

(b) Solve the problem

$$\min_B \sum_{k=1}^{K} (\sigma_k(A) - \sigma_k(B))^2 \quad \text{s.t.} \quad \mathrm{rank}(B) \le r.$$

(c) For a fixed matrix $A$, show that there exists a matrix $B$ of rank $r$ such that

$$\|A - B\|_F^2 = \sum_{k=r+1}^{K} \sigma_k(A)^2$$

(d) If $A = USV^\top$, let $U_{[r]} \in \mathbb{R}^{d\times r}$, $S_{[r]} \in \mathbb{R}^{r\times r}$, and $V_{[r]} \in \mathbb{R}^{d\times r}$) denote respectively the matrices formed of the $r$ first column of $U$, $S$ and $V$. Use the previous result to show that $B^* = U_{[r]}S_{[r]}V_{[r]}^\top$ minimizes:

$$\min_B \|A - B\|_F^2 \quad \text{s.t.} \quad \mathrm{rank}(B) \le r,$$

**Exercise 10.2** Probabilistic version of PCA. Let $\mathbf{D} \in \mathbb{R}^{p\times K}$ be a fixed full column rank matrix (thus with $K \le p$). We consider the following generative model:

$$\mathbf{z} \sim \mathcal{N}(0, \mathbf{I}_K), \quad \mathbf{x} \mid \mathbf{z} \sim \mathcal{N}(\mathbf{D}\mathbf{z}, \sigma^2\mathbf{I}_p)$$

(a) Use that with previous model we equivalently have that $\mathbf{x} = \mathbf{D}\mathbf{z} + \varepsilon$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2\mathbf{I}_p)$ where $\varepsilon$ and $\mathbf{z}$ are independent, to obtain the marginal distribution of $\mathbf{x}$; in particular compute its mean and it covariance.

(b) Assume that $\mathbf{x}_1, \ldots, \mathbf{x}_n$ is an i.i.d sample from the model above. Express its log-likelihood as a function of $\widehat{\Sigma} = \frac{1}{n}\sum_{i=1}^n \mathbf{x}_i\mathbf{x}_i^\top$.

(c) Verify that $[\mathbf{D}\mathbf{D}^\top + \sigma^2\mathbf{I}_p]^{-1} = \sigma^{-2}\mathbf{I}_p - \sigma^{-2}\mathbf{D}[\sigma^2\mathbf{I}_K + \mathbf{D}^\top\mathbf{D}]^{-1}\mathbf{D}^\top$.

(d) Show that when $\sigma^2 \to 0$, then $\sigma^2\ell(\mathbf{D}, \sigma^2)$ converges to $-\frac{n}{2}\mathrm{tr}(\widehat{\Sigma}(\mathbf{I} - \mathbf{H}))$ with $\mathbf{H} = \mathbf{D}[\mathbf{D}^\top\mathbf{D}]^{-1}\mathbf{D}^\top$.

(e) Using Von Neumann's inequality, prove that the projector on the subspace spanned by the $K$ top eigenvectors of $\widehat{\Sigma}$ maximizes $\mathrm{tr}(\widehat{\Sigma}\mathbf{H})$.

(f) Explain why when $\sigma^2$ is small, the maximum likelihood estimator for $D$ can be expected to be a matrix whose columns span the $k$th right principal subspace of $\mathbf{X}$ and whose singular values are the top singular values of $\mathbf{X}$ where $\mathbf{X}$ is the design matrix of the data.

(g) In which sense is the probabilistic model introduced at the beginning of this exercise a probabilistic counterpart of PCA?

---

## Practical Exercise

---

**Exercise 10.3** (PCA and Dimensionality Reduction) Import the file `data.csv` using the `read.csv` function in `R`. It contains a list of 10 dimensional vectors with their class.

(a) Using the `svd` function, compute the principal components of the given data set (exclude the class).

(b) How many principal components do you need to explain more than 99% of the variance?

(c) Plot the first two principal components and describe the shape of the data. Would it be a good idea to use linear classifiers to classify this data set? If not, why not?

(d) Construct a function $f : \mathbb{R}^2 \to \mathbb{R}^3$ to transform the data by mapping the first two principal components so as to render the data easily classifiable using linear classifiers. *Note: You may choose the function by inspection or trial and error.*

---

## Bonus Exercise

---

**Exercise 10.4** (von Neumann's Inequality). The goal of this problem is to establish *Von Neumann's inequality.* Let $A, B \in \mathbb{R}^{d\times d}$ two matrices whose singular values are respectively $\sigma_1(A) \geq \ldots \geq \sigma_d(A)$ and $\sigma_1(B) \geq \ldots \geq \sigma_d(B)$. *Von Neuman's inequality* says that

$$|\mathrm{tr}(A^\top B)| \leq \sum_{k=1}^d \sigma_k(A)\sigma_k(B).$$

(a) Why can we assume without loss of generality that $A$ is a diagonal matrix?[*Hint: inject the SVD of A.*]

(b) If $A = D$ is diagonal, prove that Von Neumann's inequality is equivalent to the inequality $|\mathrm{tr}(DUSV^\top)| \le \mathrm{tr}(DS)$, where $USV^\top$ is the SVD of $B$.

(c) Let $P_k = \mathrm{Diag}(\underbrace{1,\ldots,1}_{k\text{ ones}}, \underbrace{0,\ldots,0}_{d-k\text{ zeros}})$. Let $\sigma_{d+1}(A) = \sigma_{d+1}(B) = 0$ by convention, and let $a_k = \sigma_k(A) - \sigma_{k+1}(A)$ and $b_k = \sigma_k(B) - \sigma_{k+1}(B)$, so that we have

$$D = \sum_{k=1}^{d} a_k P_k \quad \text{and} \quad S = \sum_{l=1}^{d} b_l P_l.$$

Show that Von Neumann's inequality can equivalently written as

$$|\sum_{k=1}^{d}\sum_{l=1}^{d} a_k\, b_l\, \mathrm{tr}(P_k U P_l V^\top)| \le \sum_{l=1}^{d} a_k\, b_l\, \mathrm{tr}(P_k P_l).$$

(d) Deduce from the previous question that it is sufficient to prove

$$|\,\mathrm{tr}(P_k U P_l V^\top)| \le \mathrm{tr}(P_k P_l),$$

which is actually exactly Von Neumann's inequality but for a particular kind of matrix.

(e) Let $\mathbf{u}_k$ denote the $k$th column of $U$ and $\mathbf{v}_l$ denote the $l$th column of $V$. Show that $\mathrm{tr}(P_k U P_l V^\top) = \sum_{i=1}^{l} \langle P_k \mathbf{u}_i, \mathbf{v}_i \rangle \le l$ and deduce that in fact $\mathrm{tr}(P_k U P_l V^\top) \le \min(k, l)$.

(f) Use this last result to prove Von Neumann's inequality.

(g) Assume now that $A, B \in \mathbb{R}^{d\times K}$ with $K < d$, why is the inequality still true?