PC7. Ecole Polytechnique. MAP 569. Machine Learning II.

Exercise 1 (Refresher on matrices)

1. Let $\mathbf{A}$ be a $n \times d$ matrix with real entries. Show that $\mathrm{range}(\mathbf{A}) = \mathrm{range}(\mathbf{A}\mathbf{A}^\top)$.
2. Let $\{U_k\}_{1 \le k \le r}$ be a family of r orthonormal vectors of $\mathbb{R}^n$. Show that $\sum_{k=1}^r U_k U_k^\top$ is the matrix associated with the orthogonal projection onto $\mathbf{H} = \{\sum_{k=1}^r \alpha_k U_k\ ;\ \alpha_1, \ldots, \alpha_r \in \mathbb{R}\}$. Deduce that if $\mathbf{A}$ is a $n \times d$ matrix with real entries such that each column of $\mathbf{A}$ is in $\mathbf{H}$, then,

$$\left(\sum_{k=1}^r U_k U_k^\top\right) \mathbf{A} = \mathbf{A}\,.$$

3. Let $p < d$ and $\mathbf{B} \in \mathbb{R}^{d \times p}$ such that $\mathbf{B}^\top \mathbf{B} = I_p$. Let us denote $\mathbf{B} = (b_{ij})_{\substack{1 \le i \le d \\ 1 \le j \le p}}$ the components of $\mathbf{B}$ and for all $i \in [\![1, d]\!]$, $\alpha_i = \sum_{j=1}^p b_{ij}^2$. Show that $\sum_{i=1}^d \alpha_i = p$ and $\alpha_i \le 1$.

Exercise 2 (Principal Component Analysis) Principal component analysis is a multivariate technique which aims at analyzing the statistical structure of high dimensional dependent observations by representing data using orthogonal variables called *principal components*. Reducing the dimensionality of the data is motivated by several practical reasons such as improving computational complexity. Let $(X_i)_{1 \leqslant i \leqslant n}$ be i.i.d. random variables in $\mathbb{R}^d$ and consider the matrix $\mathbf{X} \in \mathbb{R}^{n \times d}$ such that the i-th row of $\mathbf{X}$ is the observation $X_i^\top$. In this exercise, it is assumed that data are preprocessed so that the columns of $\mathbf{X}$ are centered. This means that for all $1 \leqslant k \leqslant d$, $\sum_{i=1}^n X_{i,k} = 0$. Let $\mathbf{\Sigma}_n$ be the empirical covariance matrix:

$$\mathbf{\Sigma}_n = n^{-1} \sum_{i=1}^n X_i X_i^\top\,.$$

Principal Component Analysis aims at reducing the dimensionality of the observations $(X_i)_{1 \leqslant i \leqslant n}$ using a *compression* matrix $\mathbf{U} \in \mathbb{R}^{d \times p}$ with orthonormal columns with $p \leqslant d$ so that for each $1 \leqslant i \leqslant n$, $\mathbf{U}^\top X_i$ ia a low dimensional representation of X_i. The original observation may then be partially recovered using $\mathbf{U}$. Principal Component Analysis computes $\mathbf{U}$ using the least squares approach:

$$\mathbf{U}_\star \in \underset{\substack{U \in \mathbb{R}^{d \times p} \\ U^\top U = I_p}}{\mathrm{argmin}} \sum_{i=1}^n \|X_i - \mathbf{U}\mathbf{U}^\top X_i\|^2\,,$$

1. Prove that for all matrix $\mathbf{A} \in \mathbb{R}^{n \times d}$ with rank r, there exist $\sigma_1 \geqslant \ldots \geqslant \sigma_r > 0$ such that

$$\mathbf{A} = \sum_{k=1}^r \sigma_k u_k v_k^\top\,,$$

where $\{u_1, \ldots, u_r\} \subset \mathbb{R}^n$ and $\{v_1, \ldots, v_r\} \subset \mathbb{R}^d$ are two families of orthonormal vectors. The vectors $\{u_1, \ldots, u_r\}$ (resp. $\{v_1, \ldots, v_r\}$) are the left-singular (resp. right-singular) vectors associated with $\{\sigma_1, \ldots, \sigma_r\}$, the singular values of $\mathbf{A}$. *If $\mathbf{U}$ denotes the $\mathbb{R}^{n \times r}$ matrix with columns given by $\{u_1, \ldots, u_r\}$ and $\mathbf{V}$ denotes the $\mathbb{R}^{d \times r}$ matrix with columns given by $\{v_1, \ldots, v_r\}$, then the singular value decomposition of $\mathbf{A}$ may also be written as*

$$\mathbf{A} = \mathbf{U}\mathbf{D}_r\mathbf{V}^\top\,,$$

where $\mathbf{D}_r = \mathrm{diag}(\sigma_1, \ldots, \sigma_r)$. Then, $\mathbf{A}^\top\mathbf{A}$ and $\mathbf{A}\mathbf{A}^\top$ are positive semidefinite such that

$$\mathbf{A}^\top\mathbf{A} = \mathbf{V}\mathbf{D}_r^2\mathbf{V}^\top \quad \text{and} \quad \mathbf{A}\mathbf{A}^\top = \mathbf{U}\mathbf{D}_r^2\mathbf{U}^\top\,.$$

In the framwework of this exercise, $n\mathbf{\Sigma}_n = \mathbf{X}^\top\mathbf{X}$ so that diagonalizing $n\mathbf{\Sigma}_n$ is equivalent to computing the singular value decomposition of $\mathbf{X}$.

2. Prove that solving the PCA least squares optimization problem boils down to computing

$$\mathbf{U}_\star \in \underset{\mathbf{U} \in \mathbb{R}^{d \times p}\,,\, \mathbf{U}^\top\mathbf{U} = \mathbf{I}_p}{\mathrm{argmax}} \{\mathrm{Trace}(\mathbf{U}^\top\mathbf{\Sigma}_n\mathbf{U})\}\,.$$

3. Let $\lambda_1 \geqslant \ldots \geqslant \lambda_d$ be real numbers and denote $f : \alpha \in \mathbb{R}^d \mapsto \sum_{i=1}^d \alpha_i \lambda_i$. Show that

$$\sup\left\{f(\alpha) : \alpha \in [0,1]^d, \sum_{i=1}^d \alpha_i = p\right\}$$

is attained for $\alpha^\star = (\mathbb{1}_{i \leq p})_{1 \leq i \leq d}$.

4. Let $\{\vartheta_1, \ldots, \vartheta_d\}$ be orthonormal eigenvectors associated with the eigenvalues $\lambda_1 \geqslant \ldots \geqslant \lambda_d$ of $\mathbf{\Sigma}_n$. Prove that a solution to the PCA least squares optimization problem is given by the matrix $\mathbf{U}_\star$ with columns $\{\vartheta_1, \ldots, \vartheta_p\}$.

5. For any dimension $1 \leqslant p \leqslant d$, let $\mathcal{F}_d^p$ be the set of all vector subpaces of $\mathbb{R}^d$ with dimension p. Consider the linear span V_p defined as

$$V_p \in \underset{V \in \mathcal{F}_d^p}{\operatorname{argmin}} \sum_{i=1}^n \|X_i - \pi_V(X_i)\|^2 ,$$

where π_V is the orthogonal projection onto the linear span V. Prove that $V_1 = \operatorname{span}\{v_1\}$ where

$$v_1 \in \underset{v \in \mathbb{R}^d \,;\, \|v\|=1}{\operatorname{argmax}} \sum_{i=1}^n \langle X_i, v\rangle^2 .$$

6. For all $2 \leqslant p \leqslant d$, following the same steps, prove that a solution to the optimization problem is given by $V_p = \operatorname{span}\{v_1, \ldots, v_p\}$ where

$$v_1 \in \underset{v \in \mathbb{R}^d \,;\, \|v\|=1}{\operatorname{argmax}} \sum_{i=1}^n \langle X_i, v\rangle^2 \quad \text{and for all } 2 \leqslant k \leqslant p\,, \quad v_k \in \underset{\substack{v \in \mathbb{R}^d \,;\, \|v\|=1\,; \\ v \perp v_1, \ldots, v \perp v_{k-1}}}{\operatorname{argmax}} \sum_{i=1}^n \langle X_i, v\rangle^2 . \qquad (1)$$

7. Prove that the vectors $\{v_1, \ldots, v_k\}$ defined by (1) can be chosen as the orthonormal eigenvectors associated with the k largest eigenvalues of the empirical covariance matrix $\mathbf{\Sigma}_n$.

8. The orthonormal eigenvectors associated with the eigenvalues of Σ_n allow to define the principal components as follows. Then, as $V_d = \operatorname{span}\{\vartheta_1, \ldots, \vartheta_d\}$, for all $1 \leqslant i \leqslant n$,

$$\pi_{V_d}(X_i) = \sum_{k=1}^d \langle X_i, \vartheta_k\rangle \vartheta_k = \sum_{k=1}^d (X_i^\top \vartheta_k)\vartheta_k = \sum_{k=1}^d c_k(i)\vartheta_k ,$$

where for all $1 \leqslant k \leqslant d$, the k-th principal component is defined as $c_k = \mathbf{X}\vartheta_k$. Prove that $(c_1, \ldots, c_d)$ are orthogonal vectors.

Exercise 3 (Kernel Principal Component Analysis) Let $(X_i)_{1 \leq i \leq n}$ be n observations in a general space $\mathcal{X}$, $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ a positive kernel and $\mathbf{K} = (k(X_i, X_j))_{1 \leqslant i,j \leqslant n}$. $\mathcal{W}$ denotes the Reproducing Kernel Hilbert Space associated with k and for all $x \in \mathcal{X}$, $\phi(x)$ denotes the function $\phi(x) : y \to k(x,y)$. The aim is now to perform a PCA on $(\phi(X_1), \ldots, \phi(X_n))$. It is assumed that $\sum_{i=1}^n \phi(X_i) = 0$.

1. Prove that

$$f_1 = \underset{f \in \mathcal{W} \,;\, \|f\|_{\mathcal{W}}=1}{\operatorname{argmax}} \sum_{i=1}^n \langle \phi(X_i), f\rangle_{\mathcal{W}}^2$$

may be written

$$f_1 = \sum_{i=1}^n \alpha_1(i)\phi(X_i) \,, \quad \text{where} \quad \alpha_1 = \underset{\alpha \in \mathbb{R}^n \,;\, \alpha^\top \mathbf{K}\alpha = 1}{\operatorname{argmax}} \alpha^\top \mathbf{K}^2 \alpha .$$

2. Prove that $\alpha_1 = \lambda_1^{-1/2} b_1$ where b_1 is the unit eigenvector associated with the largest eigenvalue λ_1 of $\mathbf{K}$. What about $(f_2, \ldots, f_p)$ defined iteratively as in (1)?

3. Write $H_p = \operatorname{span}\{f_1, \ldots, f_p\}$. Prove that, for all $1 \leqslant i \leqslant n$,

$$\pi_{H_p}(\phi(X_i)) = \sum_{j=1}^p \lambda_j \alpha_j(i) f_j .$$