Université Paris Dauphine-PSL

Fundamentals of Machine Learning

Unsupervised learning – Project Ex 3.1

Auteurs :
Elise Chin
Mathide Da Cruz

February 10, 2022

Derivation of theoretical results

Exercise 3.1 *Demonstrate that the matrix of scalar products can be derived from the distance matrix via double-centering, in the case of Euclidean distance.*

Given $\mathbf{x}_1, ..., \mathbf{x}_n \in \mathbb{R}^m$, we define the squared distance matrix $D = [d_{ij}^2]_{ij} \in \mathbb{R}^{n \times n}$ where $d_{ij}^2 = ||\mathbf{x}_i - \mathbf{x}_j||_2^2$ and the Gram matrix $G = [g_{ij}]_{ij} \in \mathbb{R}^{n \times n}$ where $g_{ij} = x_i^T x_j$. We suppose, without loss of generality, that $\mathbf{x}_i$ are mean-centered, i.e. $\sum_i \mathbf{x}_i = 0$. We want to prove that the Gram matrix G can be computed from the squared distance matrix D via double-centering. It is a way of transforming a rectangular matrix to the one having mean values of rows equal to zero and mean values of columns equal to zero, by multiplying the centering matrix $C = I_n - \frac{1}{n}\mathbf{1}^T\mathbf{1}$[1] to both sides of D.

In other words, we want to find a constant α such that $G = \alpha CDC$.

First, we develop the term d_{ij}^2 for any i, $j \in [\![1, n]\!]$:

$$\begin{aligned} d_{ij}^2 &= ||\mathbf{x}_i - \mathbf{x}_j||_2^2 \\ &= \mathbf{x}_i^T\mathbf{x}_i - 2\mathbf{x}_i^T\mathbf{x}_j + \mathbf{x}_j^T\mathbf{x}_j \\ &= g_{ii} - 2g_{ij} + g_{jj} \end{aligned} \tag{1}$$

We denote that:

$$\begin{aligned} \sum_{j=1}^n d_{ij}^2 = \sum_{j=1}^n d_{ji}^2 = \sum_{j=1}^n g_{ii} - 2g_{ij} + g_{jj} \\ &= ng_{ii} - 2\sum_{j=1}^n g_{ij} + Tr(G) \\ &= ng_{ii} + Tr(G) - 2\mathbf{x}_i^T(\sum_{j=1}^n \mathbf{x}_j) \\ &= ng_{ii} + Tr(G) \qquad \text{because } \mathbf{x}_i \text{ are mean-centered, so } \sum_{j=1}^n \mathbf{x}_j = 0 \end{aligned}$$

Let $CD = [b_{ij}]_{ij}$ and $CDC = [b'_{ij}]_{ij}$. For any i, $j \in [\![1, n]\!]$:

$$\begin{aligned} b_{ij} &= d_{ij} - \frac{1}{n}\sum_{k=1}^n d_{kj}^2 \\ &= (g_{ii} - 2g_{ij} + g_{jj}) - \frac{1}{n}\sum_{k=1}^n d_{kj}^2 \\ &= (g_{ii} - 2g_{ij} + g_{jj}) - \frac{1}{n}(ng_{jj} + Tr(G)) \\ &= g_{ii} - 2g_{ij} - \frac{1}{n}Tr(G) \end{aligned}$$

Thus, we obtain:

[1] $\mathbf{1}^T\mathbf{1} \in \mathbb{R}^{n \times n}$ is a square matrix full of 1s.

$$
\begin{aligned}
b'_{ij} &= b_{ij} - \frac{1}{n}\sum_{k=1}^{n} b_{ik} \\
&= g_{ii} - 2g_{ij} - \frac{1}{n}Tr(G) - \frac{1}{n}\sum_{k=1}^{n}(g_{ii} - 2g_{ik} - \frac{1}{n}Tr(G)) \\
&= g_{ii} - 2g_{ij} - \frac{1}{n}Tr(G) - \frac{1}{n}(ng_{ii} - Tr(G) - 2\sum_{k=1}^{n} g_{ik}) \\
&= g_{ii} - 2g_{ij} - \frac{1}{n}Tr(G) - g_{ii} + \frac{1}{n}Tr(G) \\
&= -2g_{ij}
\end{aligned}
$$

We deduce that $CDC = -2G \iff G = -\frac{1}{2}CDC$.