Machine Learning 2 - Homework week 1: PCA

Mai Xuan Bach
ID 11200489

Ngày 27 tháng 12 năm 2022

1 Problem 1. Biến đổi:

We consider an independent and identically distributed (i.i.d.) dataset X = $\{x_1, x_2, ..., x_n\}$ with mean 0, $x_i \in R^D$

We assume there exists a low-dimensional compressed representation:

$$z_n = B^T x_n \in R^M \tag{1}$$

of x_n where projection matrix is

$$B = [b_1, b_2, ..., b_M] \in R^{D \times M} \tag{2}$$

In other words, we want to achieve a matrix of principal components Z, which is: $Z = X.B, Z \in R^{N \times M}$, with M is the number of components we want to reduce to.

So that, we need to find a matrix B that retains as much information as possible when compressing data by projecting it onto the subspace spanned by the columns $b_1, b_2, ..., b_M$ of B. In other words, we want to maximize the variance of data points when projected onto a new space.

We maximize the variance of the low-dimensional code using a sequential approach. We start by seeking a single vector $b_1 \in R^D$ that maximizes the variance of the projected data, i.e., we aim to maximize the variance of the first coordinate z_1. Note that, we can easily prove that $\mu_Z = 0$.

$$\begin{aligned}
Var &= \frac{\sum_{i=1}^N (x_i^T b_1 - \mu_Z)^2}{N} && (3)\\
&= \frac{\sum_{i=1}^N (x_i^T b_1)^2}{N} && (4)\\
&= \frac{\sum_{i=1}^N b_1^T x_i x_i^T b_1}{N} && (5)\\
&= b_1^T (\frac{\sum_{i=1}^N x_i x_i^T}{N}) b_1 && (6)\\
&= b_1^T S b_1 && (7)
\end{aligned}$$

where S is called the data covariance matrix.

Now, the problem turns into optimization problem as follows:

$$\begin{aligned} &\max_{b_1} b_1^T S b_1 \\ &\text{s.t.} ||b_1||^2 = 1 \end{aligned} \tag{8}$$

The Lagrangian function:

$$\mathcal{L}(b_1, \alpha) = b_1^T S b_1 + \alpha(1 - b_1^T b_1) \tag{9}$$

$$\frac{\partial L}{\partial b_1} = 2b_1^T S - 2\alpha b_1^T \tag{10}$$

$$\frac{\partial L}{\partial \alpha} = 1 - b_1^T b_1 \tag{11}$$

Setting these partial derivaties to 0 gives us

$$b_1^T b_1 = 1 \tag{12}$$

$$S b_1 = \alpha b_1 \tag{13}$$

We can see that b_1, α is an eigenvector and an eigenvalue of S respectively. Then,

$$Var = b_1^T S b_1 = b_1^T \alpha b_1 = \alpha \tag{14}$$

The variance of the data projected onto a one-dimensional subspace equals the eigenvalue that is associated with the basis vector b_1 that spans this subspace.

Therefore, to maximize the variance of the low-dimensional code, we choose the basis vector associated with the largest eigenvalue principal component of the data covariance matrix.

This eigenvector is called the first principal component. The second component is the projection of data onto the eigenvector corresponding with the second largest eigenvalue.

In conclusion, we can list out the steps in PCA:

0. Scale data into mean = 0

1. Compute the covariance matrix S

2. Find eigenvalues, eigenvectors of S. Sort the eigenvalues in descending order. Then, reorder the eigenvectors with the sorted eigenvalues.

3. Projection: Project the data onto the eigenvectors.