DDA3020 Programming Report 4

Yu Kangqi

May 10, 2023

1 PCA and K-means from scratch

1.1 PCA

Denote the ith data item as $\boldsymbol{x_i}$, the subspace that we would like to project $\boldsymbol{x_i}$ to is $\mathcal{S}$. The projection of $\boldsymbol{x_i}$ to $\mathcal{S}$ is formed by $\{\boldsymbol{u_k}\}_{k=1}^{K}$.
The projection length from x_i to u_k is defined as:

$$z_{ik} = \boldsymbol{u_k}^\top(\boldsymbol{x_i} - \boldsymbol{\mu})$$

where $\boldsymbol{\mu} = \frac{1}{N}\sum_{i}^{N}\boldsymbol{x_i}$.
The representation of $\boldsymbol{x_i}$ will be $\tilde{\boldsymbol{x}}_{\boldsymbol{i}}$:

$$\boldsymbol{z_i} = \boldsymbol{U}^\top(\boldsymbol{x_i} - \boldsymbol{\mu})$$

Then the reconstruction of $\boldsymbol{x_i}$ will be:

$$\tilde{\boldsymbol{x}}_{\boldsymbol{i}} = \boldsymbol{U}\boldsymbol{z_i} + \boldsymbol{\mu}$$

The object of PCA is to maixmize the variance of the representation in the subspace $\mathcal{S}$. The variance is defined as:

$$\sum_{i}^{N}||\tilde{\boldsymbol{x}}_{\boldsymbol{i}} - \tilde{\boldsymbol{\mu}}||^2$$

where $\tilde{\boldsymbol{\mu}} = \frac{1}{N}\sum_{i}^{N}\tilde{\boldsymbol{x}}_{\boldsymbol{i}}$.

So, to reduce the dimension, we can solve the following optimization problem:

$$\max_{\boldsymbol{U}} \sum_{i}^{N} ||\tilde{\boldsymbol{x}}_{\boldsymbol{i}} - \tilde{\boldsymbol{\mu}}||^2$$

$$s.t. \quad \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$$

Claim: $\tilde{\boldsymbol{\mu}} = \boldsymbol{\mu}$

Proof: $\tilde{\boldsymbol{\mu}} = \sum_{i}^{N}\tilde{\boldsymbol{x}}_{\boldsymbol{i}} = \sum_{i}^{N}\boldsymbol{U}\boldsymbol{z}_{\boldsymbol{i}} + \boldsymbol{\mu} = \boldsymbol{U}\sum_{i}^{N}\boldsymbol{z}_{\boldsymbol{i}} + \boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N}\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{\mu} + \boldsymbol{\mu} = \boldsymbol{\mu}$

The optimazation problem turns to:

$$\max_{\boldsymbol{U}} \sum_{i}^{N} ||\tilde{\boldsymbol{x}}_{\boldsymbol{i}} - \boldsymbol{\mu}||^2$$

$$s.t. \quad \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$$

As $\tilde{\boldsymbol{x}}_{\boldsymbol{i}} = \boldsymbol{U}\boldsymbol{z}_{\boldsymbol{i}} + \boldsymbol{\mu}$, the objective function can be written as:

$$\sum_{i}^{N} ||\boldsymbol{U}\boldsymbol{z}_{\boldsymbol{i}}||^2$$

Then, for $||\boldsymbol{U}\boldsymbol{z}_{\boldsymbol{i}}||^2 = (\boldsymbol{U}\boldsymbol{z}_{\boldsymbol{i}})^\top(\boldsymbol{U}\boldsymbol{z}_{\boldsymbol{i}}) = \boldsymbol{z}_{\boldsymbol{i}}^\top\boldsymbol{U}^\top\boldsymbol{U}\boldsymbol{z}_{\boldsymbol{i}} = \boldsymbol{z}_{\boldsymbol{i}}^\top\boldsymbol{z}_{\boldsymbol{i}} = ||\boldsymbol{z}_{\boldsymbol{i}}||^2$, the objective function can be written as:

$$\sum_{i}^{N} ||\boldsymbol{z}_{\boldsymbol{i}}||^2$$

As $\boldsymbol{z}_{\boldsymbol{i}} = \boldsymbol{U}^\top(\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{\mu})$, the objective function can be written as:

$$\sum_{i}^{N} ||\boldsymbol{U}^\top(\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{\mu})||^2 = \sum_{i}^{N} Trace(\boldsymbol{U}^\top(\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{\mu})(\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{\mu})^\top\boldsymbol{U})$$

Note that $\sum_{i=1}^{N}(\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{\mu})(\boldsymbol{x}_{\boldsymbol{i}} - \boldsymbol{\mu})^\top$ is N times the empirical covariance matrix $\boldsymbol{\Sigma}$, so the objective function can be written as:

$$Trace(\boldsymbol{U}^\top\boldsymbol{\Sigma}\boldsymbol{U})$$

Until now, the optimization problem turns to:

$$\max_{\boldsymbol{U}} \ Trace(\boldsymbol{U}^\top\boldsymbol{\Sigma}\boldsymbol{U})$$

$$s.t. \quad \boldsymbol{U}^\top\boldsymbol{U} = \boldsymbol{I}$$

The Lagrangian of the optimization problem is:

$$L(\boldsymbol{U}, \boldsymbol{\Lambda}) = Trace(\boldsymbol{U}^\top \boldsymbol{\Sigma} \boldsymbol{U}) + Trace(\boldsymbol{\Lambda}^\top (\boldsymbol{U}^\top \boldsymbol{U} - \boldsymbol{I}))$$

where $\boldsymbol{\Lambda}$ is the Lagrange multiplier.

Take first order derivative of $L(\boldsymbol{U}, \boldsymbol{\Lambda})$ with respect to $\boldsymbol{U}$ and set it to zero, we have:

$$\boldsymbol{\Sigma} \boldsymbol{U} + \boldsymbol{U} \boldsymbol{\Sigma}^\top - \boldsymbol{U} \boldsymbol{\Lambda} - \boldsymbol{\Lambda}^\top \boldsymbol{U} = 0$$

$$\boldsymbol{\Sigma} \boldsymbol{U} = \boldsymbol{U} \boldsymbol{\Lambda}$$

$$\boldsymbol{\Sigma} \boldsymbol{u_k} = \lambda_k \boldsymbol{u_k}$$

As we can see, $\boldsymbol{u_k}$ is the eigenvector of $\boldsymbol{\Sigma}$, and λ_k is the corresponding eigenvalue.

Futher, we can turn the optimization problem to:

$$\max_{\boldsymbol{U}} \sum_{k=1}^{K} \boldsymbol{u_k}^\top \boldsymbol{\Sigma} \boldsymbol{u_k}$$

$$s.t. \quad \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$$

From simplicity, we can use SVD to decompose $\boldsymbol{\Sigma}$ to $\boldsymbol{Q} \boldsymbol{\Lambda} \boldsymbol{Q}^\top$, where $\boldsymbol{U}$ is the eigenvector matrix and $\boldsymbol{\Lambda}$ is the eigenvalue matrix.

Then, the optimization problem turns to:

$$\max_{\boldsymbol{U}} \sum_{k=1}^{K} \sum_{d=1}^{D} \lambda_d \boldsymbol{u_k}^\top \boldsymbol{q_d} \boldsymbol{q_d}^\top \boldsymbol{u_k}$$

$$s.t. \quad \boldsymbol{U}^\top \boldsymbol{U} = \boldsymbol{I}$$

To maximize the objective function, we need to pick K eigenvectors with top-K eigenvalues of $\boldsymbol{\Sigma}$.

The way that I implement the PCA algorithm is as following:

1. Calculate the empirical covariance matrix $\boldsymbol{\Sigma} = \frac{1}{N} \sum_{n=1}^{N} \left(\mathbf{x}^{(n)} - \boldsymbol{\mu}\right) \left(\mathbf{x}^{(n)} - \boldsymbol{\mu}\right)^\top$, where $\boldsymbol{\mu}$ is a vector contianing the mean of each attribution.

2. Do SVD decomposition of $\boldsymbol{\Sigma}$ to obtain its D eigenvalues $\{\lambda_i\}_{i=1}^{D}$ and eigenvectors $\{\mathbf{q}_i\}_{i=1}^{D}$, and rank them from large to small according to the eigenvalues.

3. Pick the top-K eigenvectors to form the matrix $\mathbf{U} = [\mathbf{q}_1, \ldots, \mathbf{q}_K] \in \mathbb{R}^{D \times K}$.

4. The new representation of $\mathbf{x}^{(n)}$ is $\mathbf{U}^\top \left(\mathbf{x}^{(n)} - \boldsymbol{\mu}\right)$.

1.2 K-means

Denote the ith data item to x_i, the kth cluster center to c_k. If x_i belongs to c_k, $r_{ik} = 1$, otherwise 0.

The object of K-means is to minimize the variance between the data items and the cluster center. The variance is defined as:

$$\sum_{i=0}^{n}\sum_{k=0}^{K} r_{ik}(x_i - c_k)^2$$

The optimazation perspective of K-means is solve the following objective function:

$$\min_{\boldsymbol{c},\boldsymbol{r}} \sum_{i=0}^{n}\sum_{k=0}^{K} r_{ik}(x_i - c_k)^2$$
$$s.t. \quad \begin{cases} \boldsymbol{r} \in \{0,1\}^{n\times K} \\ \sum\limits_{k=1}^{K} r_{ik} = 1, \forall i \end{cases}$$

We can solve this problem by using the following iterative algorithm:

1. Initialize the cluster centers $\boldsymbol{c} = \{c_1, c_2, \ldots, c_K\}$.
2. Repeat the following steps until convergence:
   (a) Given the cluster centers $\boldsymbol{c}$, update $\boldsymbol{r}$. So, the optimazation problem in this part is as follows:

$$\min_{\boldsymbol{r}} \sum_{i=0}^{n}\sum_{k=0}^{K} r_{ik}(x_i - c_k)^2$$
$$s.t. \quad \begin{cases} \boldsymbol{r} \in \{0,1\}^{n\times K} \\ \sum\limits_{k=1}^{K} r_{ik} = 1, \forall i \end{cases}$$

   r_{ik} can be solved independently in the above problem, so we can solve it by:

$$r_{ik^*} = 1$$

   where $k^* = argmin\sum_{i=1}^{n}\{(x_i - c_k)^2\}_{k=1}^{K}$

(b) Given the cluster centers $\boldsymbol{r}$, update $\boldsymbol{c}$. So, the optimazation problem in this part is as follows:

$$\min_{\boldsymbol{c}} \sum_{i=0}^{n} \sum_{k=0}^{K} r_{ik}(x_i - c_k)^2$$

c_k can be solved independently in the above problem, so we can solve it by taking derivative of c_k and set it to 0, then we get:

$$c_k = \frac{\sum_{i=1}^{n} r_{ik} x_i}{\sum_{i=1}^{n} r_{ik}}$$

The way that I implement the K-means algorithm is as following:

1. Randomly pick K (K is the cluster number) points as the initial centroids.
2. Go into a loop, in each iteration, do the following:
   (a) Assign each point to the nearest centroid.
   (b) Update the centroids by calculating the mean of all points assigned to it.
3. Stop the loop when the centroids do not change or the iteration items goes to the maximum iteration.

2 Results

3 Evaluation to the results

3.1 Silhouette Coefficient

The silhouette coefficient is a measure of cluster cohesion and separation. It quantifies how well a data point fits into its assigned cluster based on two factors:

$$a(i) = \frac{\sum_{j \in C_i, j \neq i} d(i,j)}{|C_i - 1|}, \forall i \in C_i$$

$$b(i) = \min_{k \neq i} \frac{\sum_{j \in C_k} d(i,j)}{|C_k|}$$

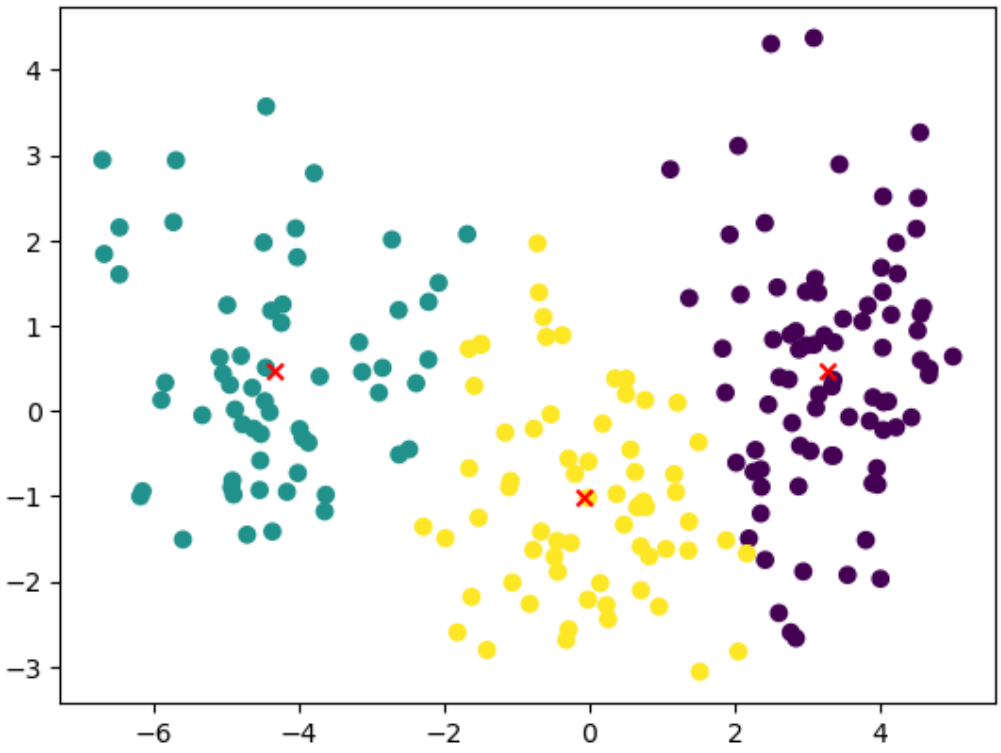

Figure 1: The result of K-means clustering for the data after PCA

where $d(i, j)$ is the distance between point i and j, C_i is the cluster that point i belongs to.

The $s(i)$ for a single point is then given as:

$$s(i) = \begin{cases} \dfrac{b(i) - a(i)}{\max(a(i), b(i))}, if|C_i| > 1 \\ 0, if|C_i| = 1 \end{cases}$$

For cluster k, $\bar{s}(k)$ is the mean of all the points in the cluster. The silhouette coefficient for the clustering is the maximum of $\bar{s}(k)$ for all clusters.

$$SC = \max \bar{s}(k)$$

3.2 Rand Index

The Rand index is a measure of the similarity between two data clusterings. Given a set of elements S and two partitions of S to compare, X and Y, define the following:

In this example, X is the observation and Y is the predicted result.
a, the number of pairs of elements in S that are in the same subset in X and in the same subset in Y.
b, the number of pairs of elements in S that are in different subsets in X and in different subsets in Y.
c, the number of pairs of elements in S that are in the same subset in X and in different subsets in Y.
d, the number of pairs of elements in S that are in different subsets in X and in the same subset in Y.

$$IR = \frac{a + b}{a + b + c + d}$$

3.3 Performance of the clustering

Here is the performance of the above clustering:

- Silhouette Coefficient: 0.5463
- Rand Index: 0.8744