

MATH466/MATH766

Math of machine learning

02/26 Lecture 14 Clustering

References:

-

Todays contents:

- k means
- Review of linear algebra

Important concepts:

- nearest neighbor

Recommend reading:

-

Unsupervised Learning

data : $\{x_i\}_{i=1}^n$ unlabeled

common tasks : ① clustering - separate objects into groups

(+ classification)

② dimension reduction, feature extraction

representation learning, manifold learning

(can be used for pre-processing of
supervised learning and other tasks)

③ generative model

common challenges :

intrinsic $\begin{pmatrix} \text{high dimension data} \\ \text{high volume of data} \\ \text{scarcity of data} \end{pmatrix}$ computation

Plan for this module

clustering : K-means

spectral clustering

Expectation Maximization (if time permits)

dim reduction

Principal Component Analysis (PCA)

Multidimensional Scaling (MDS)

closely
related

& manifold learning :

graph
 \downarrow

Page Rank

Iso map

spectral embedding (Laplace eigenmap)

diffusion map (HW)

t-distribution Stochastic Neighbor Embedding (t-SNE)

Clustering

Data: $\{x_i\}_{i=1}^n$

Goal: find an assignment map π s.t.

(hard clustering) $\pi(x_i) \in \{1, 2, \dots, k\}$

(soft membership) $\pi(x_i) \subseteq \{1, 2, \dots, k\}$

k-means (hard clustering)

Find k cluster centers $\{\mu_1, \dots, \mu_k\}$

and the assignment map π

$$\text{s.t. } \min_{\pi, \mu} \sum_{i=1}^n \|x_i - \mu_{\pi(x_i)}\|^2$$

- Generally NP-hard
- A heuristic algorithm: Lloyd's Algorithm (1957)

randomly initialize μ , loop until μ converges

Fix center μ , update π

Fix assignment π , update μ

The algorithm usually converges

But ① different initialization \rightarrow different converged solutions

② difficult case \rightarrow easier to get different results

- Strategies

① run the algorithm multiple times
and pick the best solution

② k-mean ++ initialization
(spread out the k initial cluster centers)

choose μ_1 uniformly randomly from $\{x_i\}_{i=1}^n$
for $l = 2, \dots, k$
 for x_i not chosen, compute $D(x_i)$: the dist between x_i and its nearest center
 choose μ_l randomly from remaining x_i with a weight proportional to $D^2(x_i)$.

- Demo : difficult case for k-means

① unknown value of $k \leftarrow$ silhouette analysis

② unbalanced clusters \leftarrow multiple runs

③ anisotropically distributed blobs

④ unequal variance

⑤ nonconvex clusters

intrinsic difficulties

Eigen - Decomposition

1. $A \in \mathbb{R}^{n \times n}$

eigenvalue λ , eigenvector x $Ax = \lambda x, \lambda \in \mathbb{C}, x \in \mathbb{C}^n$
 \downarrow n linearly independent eigenvectors

diagonalizable

$$A = P \Lambda P^{-1}$$

$$P = [x_1 \ x_2 \ \dots \ x_n]$$

$$\Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}$$

- ① Not all matrices are diagonalizable
 - ② The diagonalizing matrix P is not unique
 - ③ If A is diagonalizable, $B = SAS^{-1}$ for some S
 (B is similar to A), then B is diagonalizable
-

2. $A \in \mathbb{R}^{n \times n}, A = A^T$

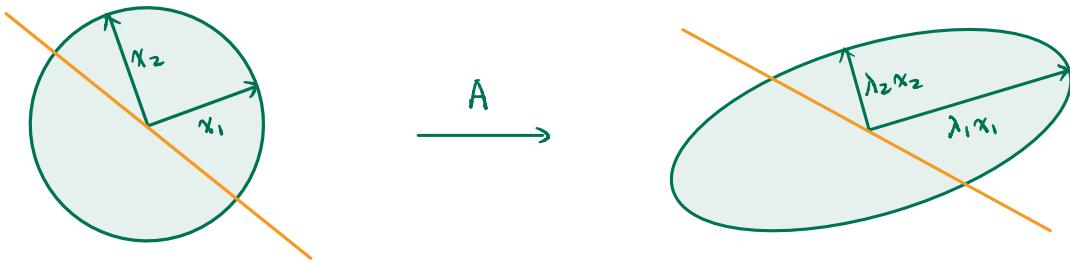
(Spectral Thm) A can be factored into $A = Q \Lambda Q^T$
 where $Q = [x_1, x_2, \dots, x_n]$ is orthogonal

- ① Always diagonalizable
 - ② eig-vals and eig-vecs are real
 - ③ $A = \sum_{i=1}^n \lambda_i \frac{x_i x_i^T}{\|x_i\|^2}$ projection operator
-

WLOG, assume $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_n$

Define Rayleigh quotient $R(x) := \frac{x^T A x}{x^T x}$

* Courant - Fischer - Weyl min-max principle



$$\lambda_1 = \max_x R(x),$$

$$x_1 = \operatorname{argmax}_x R(x)$$

$$\lambda_n = \min_x R(x)$$

$$x_n = \operatorname{argmin}_x R(x)$$

$$\lambda_2 = \max_{x^T x = 0} R(x) \quad \text{"max" when } x \in \text{span}\{x_2\}$$

S_j : j-dim subspace

$$\lambda_2 = \max_{S_2} \left[\min_{x \in S_2} R(x) \right] \quad \text{"max" when } S_2 = \text{span}\{x_1, x_2\}$$

$$\lambda_j = \max_{S_j} \left[\min_{x \in S_j} R(x) \right] \quad \text{"max" when } S_j = \text{span}\{x_1, x_2, \dots, x_j\}$$

$$= \min_{S_{n-j+1}} \left[\max_{x \in S_{n-j+1}} R(x) \right] \quad \text{"min" when } S_{n-j+1} = \text{span}\{x_j, x_{j+1}, \dots, x_n\}$$

$$3. \quad A \in \mathbb{R}^{n \times n}, \quad A = A^T, \quad A \text{ p.s.d.}$$

$$\text{Def. (Frobenius norm)} \quad \|A\|_F^2 = \sum_{i,j=1}^n A_{ij}^2$$

Low-rank approximation:

$$\min_{\tilde{A}} \|\tilde{A} - A\|_F \quad \text{s.t. } \operatorname{rank}(\tilde{A}) \leq j$$

$$\text{Solution: } \tilde{A} = \sum_{i=1}^j \lambda_i x_i x_i^T$$

Q: min-max principle and low-rank approximation for general real matrix?