Info 251: Applied Machine Learning
Lab 12
4/22/2020

Topics

- Unsupervised Learning
- k-Means Clustering
- Dimensionality Reduction

Unsupervised Learning

- Absence of labeled data
- Pre-processing for supervised learning

k-Means Clustering

- The aim is to segregate groups with similar traits and assign them into clusters
- Input: $x_1, \ldots, x_N,\ x_i \in \mathbb{R}^n$
- Parameter: K clusters
- Distance metric: Euclidean
- Other metrics can be problematic

k-Means Clustering

- Objective

$$\min_S \sum_{i=1}^{K} \sum_{x \in S_i} ||x - \mu_i||_2^2$$

- $S_1, \ldots, S_K$ are the clusters
- μ_i is the mean of points in S_i
- NP-Hard

k-Means Algorithm

Initialize: $\mu_i,\ i = 1, \dots, K,\ t = 0$
while *Centroids change* **do**
  Assign:$S_i^{(t)} = \{||x_p - \mu_i^{(t)}||_2^2 \leq ||x_p - \mu_j^{(t)}||_2^2, \forall j, j = 1, \dots, K\}$ for each data point p
  Update:$\mu_i^{(t+1)} = \frac{1}{|S_i^{(t)}|} \sum_{x_j \in S_i^{(t)}} x_j$
  t=t+1
end

Algorithm 1: (Naive) k-Means

- Initialization of μ_is
- Either pick randomly K points from $x_i, i = 1, \dots, N$ or
- randomly assign each point x_i to $1, \dots, K$ cluster and then compute the μ_is

k-Means Algorithm

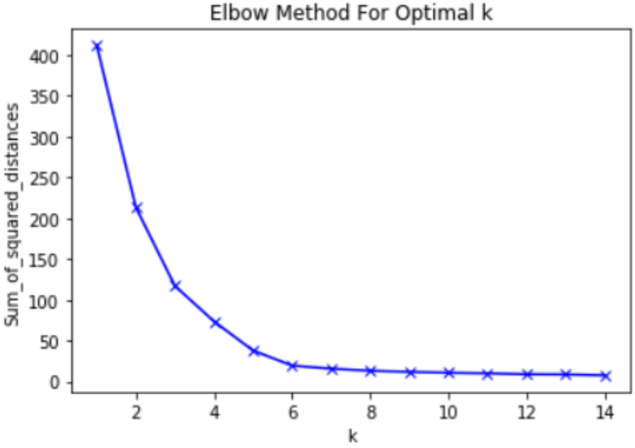
- Heuristic algorithm
- It converges
- No guarantees of optimality
- Standardize data beforehand
- How do we choose K?

k-Means Algorithm

- Heuristic algorithm
- It converges
- No guarantees of optimality
- Standardize data beforehand
- How do we choose K?
- Elbow rule
- Try for example $K = 1, 2, 3, \ldots 20$ and plot sum of squared errors vs K

Elbow Rule

- $SSE = \sum_{i=1}^{K} \sum_{x \in S_i} ||x - \mu_i||_2^2$

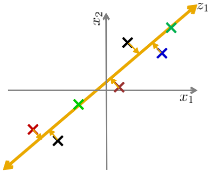


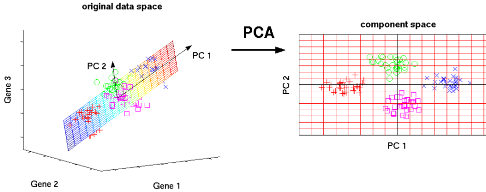
Dimensionality Reduction

- Curse of dimensionality
- Helps with overfitting
- Data visualization

Principal Component Analysis

- Karl Pearson (1901)
- $x_i \in \mathbb{R}^m \rightarrow z_i \in \mathbb{R}^k$ with $k << m$
- High level idea is to project high dimensional data to lower dimensions that explain a lot of variation





PCA

- How do we get the space along which projections have the largest variance?
- SVD of a matrix $X \in \mathbb{R}^{n \times m}$ gives decomposition $U \Sigma V^T$
- U is an $n \times n$ unitary matrix
- Σ is an $n \times m$ rectangular diagonal matrix with non-negative real numbers
- V is an $m \times m$ unitary matrix
- Columns of V are eigenvectors of $X^T X$
- Columns of U are eigenvectors of $X X^T$
- The elements of Σ are the square roots of eigenvalues of $X^T X$

PCA

- PCA using SVD
- Center X
- Compute $X = U\Sigma V^T$
- Principal components are given by $U\Sigma$
- To reduce dimensionality to k $X_{pca} = U_k\Sigma_k$
- U_k denotes first k columns of matrix U

PCA

- PCA using SVD
- Center X
- Compute $X = U\Sigma V^T$
- Principal components are given by $U\Sigma$
- To reduce dimensionality to k $X_{pca} = U_k\Sigma_k$
- U_k denotes first k columns of matrix U
- Train-test split and then PCA or PCA and then train-test split?

- Notebook