MATH 748: Weekly Report

Dec 6 - Dec 12, 2021

Nayeong Kim (nkim10@sfsu.edu)

17 Unsupervised Learning

17.1 Introduction

(1) Supervised Learning vs. Unsupervised Learning
The difference of supervised learning problems and unsupervised learning problems comes from some feature of a given dataset. Unlike supervised learning, in unsupervised learning problems, the labels are not given. Hence the goal is different. In supervised learning, the goal is to find the label for a new dataset. On the other hand, the goal of unsupervised learning is to discover outstanding feature of a dataset. For example, clustering is a kind of unsupervised learning. It groups observations into subgroups with similar features.

(2) What is Unsupervised Learning?
It is learning patterns from data without pre-given labels. Only features are given. There are various unsupervised learning problems: dimension reduction, clustering, and density estimation problems. Since there is no pre-given labels, it is very hard to evaluate the result.

17.2 Dimension Reduction

The goal of dimension reduction is to describe the given data set with less variables. For example, when observations are distributed on a sphere which in 3-dimensional space, it can be reduced into 2-dimensional space by using spherical coordinate system. Principle component analysis(denoting PCA) is one of the methods for dimension reduction.

(1) Motivation of PCA
The main purpose is to reduce data dimension. Define the random vector and its mean vector(*population mean*) as $X = (X_1, \cdots, X_p)^T, \mu = E(X)$. Then the covariance matrix(*population variance-covariance matrix*) of X is the $\Sigma = Cov(X) = E(X-\mu)(X-\mu)^T$, its ij-th entry $\sigma_{ij} = E(X_i - \mu_i)(X_j - \mu_j)$. With the definition of σ_{ij}, we have Σ is symmetric. In practice, μ, Σ are unknown and estimated from the data.

(2) Linear combination of inputs
Consider the linear combinations

$$Z_j = v_j^T X = v_{j1}X_1 + \cdots v_{jp}X_p \quad j = 1, \cdots p.$$

With the constraints:

$$||v_i|| = \sum_i v_{ij}^2 = 1.$$

it is called normalizaed linear combinations. Also, $Var(Z_j) = v_j^T \Sigma v_j$, and $Cov(Z_j, Z_k) = v_j^T \Sigma v_k (j \neq k)$.

(3) PCA
PCA is a statistical procedure that

- uses an orthogonal transformation to convert a set of observations of correlated variables into a set of linearly uncorrelated variables (called *principal components*).
- finds directions with maximum variability.

The conditions for principal components are i) *uncorrelated, orthogonal* linear combinations $Z_1, \cdots, Z_p$ which maximizes the variance. Therefore we will find a new coordinate system by rotating (and translating) original observations.

(4) Mathematical Formulation

1. Find v_1 maximizes $Var(v_1^T X)$ with constraint: $||v_1|| = 1$
2. Inductively,
   find v_j maximizes $Var(v_j^T X)$ with constraints: $||v_j|| = 1, Cov(v_j^T X, v_k^T X) = 0$ for all $1 \leq k \leq j-1$.

Here, $Cov(v_j^T X, v_k^T X) = 0$ means the orthogonality of all v_js. The procedure inductively finds a unit vector v_j which maximizes the variance and orthogonal to previous vectors. To calculate $Cov(v_j^T X, v_k^T X) = v_j^T \Sigma v_k$, we need Σ so we use S_n as its estimator when it is unknown.

(5) Eigen Decomposition of Σ
We can assume that there are p pairs of eigenvalues and (unit) eigenvectors of Σ, denoting (λ_j, e_j), such that e_js form an orthonormal basis. (Actually, this makes sense because diagonalizable matrices are dense. I.e. in a small enough neighborhood of a matrix, we can find a diagonalizable matrix.) With spectral decomposition, we can write:

$$\Sigma = \sum_j \lambda_j e_j e_j^T$$

We can write $Var(Z_j) = e_j^T \Sigma e_j = \lambda_j$ because e_js form an orthonormal basis. Furthermore, the magnitude of e_{jk} measures the importance of the kth variable to jth principal component.

(6) Number of PC
It is determined based on

- the amount of total sample variance.
- the *scree* plot.

(7) Remarks
It is very useful in exploratory data analysis. It provides a simple description of the covariance structure and it is a good tool for a visualization of high-dimensional data. This analysis doesn't require any pre-assumptions for the distribution of the dataset(e.g. multivariate normal distribution).

17.3 Density Estimation

The goal is to estimate the density distribution of $X_1, \cdots, X_p$. The biggest challenge is to deal with high dimensional problems.

17.4 Clustering

(1) Goal
The goal of clustering is grouping or segmenting the dataset into subsets such that those within each subset are more closely related to on other than those assigned to different subsets.

(2) Challenges
It is hard to evaluate because clusters can be ambiguous and subjective (common problems in unsupervised learning: we don't know the true answer).

(3) K-means Clustering
The number K: pre-specified number of clusters is a tuning parameter. We will find a partition of $S = \{1, \cdots, n\}$: $C_1, \cdots, C_K$ which are disjoint subsets of S having a union S. It can be expressed in an optimization problem tha minimize

$$\sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

where

$$\frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 = 2 \sum_{i \in C_k} \sum_{j=1}^{p} (x_{ij} - \bar{x}_{kj})^2 \quad \text{where } \bar{x}_{kj} \text{ is a mean of the cluster } k.$$

The K-means clustering can be done with the following algorithm.

1. Randomly assign each observation to one of the K-clusters.
2. Iterate until the cluster assignments stop changing:
    - Find the centroid of each cluster.
    - Assign observations to the cluster of the closest centroid.

It can be stuck in a local optimum. Multiple trials of the algorithms can resolve the problem.

(4) Hierarchical Clustering
Hierarchical clustering doesn't require a hyper parameter K. After clustering, we can tune the K so we can choose the best (which can be subjective) clusters. It produces a tree based representation of the observations called *Dendrogram*. Using Dendrogram, we can decide the number of cluster: K. We can apply bottom-up algorithm as below.

1. Start with n clusters having one point. Each point is a cluster.
2. Identify the closest clusters and merge them. When j clusters exist, $j(j-1)$ comparisons are made.
3. Iterate the step 2 until a single cluster remains.

From the root to the leaves, it means the 1 cluster to n clusters. There are two important settings in hierarchical clustering. First one is dissimilarity measure. With different measures, the order of distances can differ: e.g. cosine, Euclidean, etc. Second one is defining dissimilarity between two clusters: called *linkage methods*. The following is examples of linkage methods.

- Complete Linkage: Largest distance between observations.
- Single Linkage: Smallest distance between observations.
- Average Linkage: Average distance between observations.
- Centroid Linkage: Distance between centroids.

Among them, complete linkage and average linkage are preferred. Centroid linkage has a benefit in time but it can cause inversion which means that two clusters merged are more similar than the pair of clusters that were merged in a previous step. It can happen because centroid linkage is not monotonic.
For the choice of dissimilarity measures, Euclidean distance and correlation-based distance are preferred.

(5) Challenges in Clustering
There are some challenges in clustering:

- Should the features be standardized?
- Hierarchical Clustering: What dissimilarity measure? What linkage? Where to cut the Dendogram?
- K-means Clustering: What K?

In practice, we try several different choices and compare them.