Unsupervised learning *

Principal component analysis

- PCA seeks a small number of dimensions that are where the observations vary along each dimension. Each dimension in PCA is a linear combination of p features $z_1 = \phi_{11} x_1 + \dots + \phi_{p1} x_p$ where $\sum_{j=1}^{p} \phi_{j1}^2 = 1$

$\Rightarrow$ Maximize $\{ \frac{1}{n} \sum_{i=1}^{n} (\sum_{j=1}^{p} \phi_{j1} x_{ij})^2 \}$ subject to $\sum_{j=1}^{p} \phi_{j1}^2 = 1$ (sample variance of the n values of z_{i1}.

$\phi_{11}, \dots \phi_{p1}$

The loading vector ϕ_1 with elements $\phi_{11}, \phi_{21}, \dots \phi_{p1}$ defines a direction in feature space along which the data vary the most. If we project the n data points $x_1, \dots x_n$ onto this direction, the projected values are the principal component scores $z_{11}, \dots z_{n1}$ themselves.

- The first M principal component score vectors and the first M principal component loading vectors provide the best M-dimensional approximation (in terms of Euclidian distance) to the ith observation x_{ij}. This representation can be written $x_{ij} \approx \sum_{m=1}^{M} z_{im} \phi_{jm}$

Here the $\neq$ between original data and approximation is $\sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - \sum_{m=1}^{M} z_{im} \phi_{jm})^2$ = approximation error

- $\underbrace{\sum_{j=1}^{p} \frac{1}{n} \sum_{i=1}^{n} x_{ij}^2}_{\text{Var of data}} = \underbrace{\sum_{m=1}^{M} \frac{1}{n} \sum_{i=1}^{n} z_{im}^2}_{\text{Var of first } M \text{ PCs (to be max)}} + \underbrace{\frac{1}{n} \sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - \sum_{m=1}^{M} z_{im} \phi_{jm})^2}_{\text{MSE of } M\text{-dimensional approximation (to be min)}}$

We see that PCA can be seen as minimizing the approximation error as ~~minimizing~~ maximizing the variance

- Proportion of variance explained of the mth principal component is

$$1 - \frac{\sum_{j=1}^{p} \sum_{i=1}^{n} (x_{ij} - \sum_{m=1}^{M} z_{im} \phi_{jm})^2}{\sum_{j=1}^{p} \sum_{i=1}^{n} x_{ij}^2} = 1 - \frac{RSS}{TSS}$$

where TSS is the total sum of squared elements of X, and RSS represents the residual sum of squares of the M-dimensional approximation given by the principal components. Recalling the definition of R^2, this means we can interpret the PVE as the R^2 of approximating X given the first M principal components.

- Before performing PCA, the variables should be centered to have zero mean. Furthermore they should have been individually scaled. This is in contrast to some other supervised and unsupervised techniques such as linear regression in which scaling has no effect.

Matrix completion

1. create a complete data matrix $\tilde{x}$ of dimension $n \times p$ of which the (i,j) elements equals

$$\tilde{x}_{ij} \begin{cases} x_{ij} & \text{if } (i,j) \in O \\ \bar{x}_j & \text{if } (i,j) \notin O \end{cases}$$

2. Repeat steps (a) - (c) until the objective fails to decrease:

(a) Solve

$$\underset{A \in \mathbb{R}^{n \times M}, B \in \mathbb{R}^{p \times M}}{\text{minimize}} \left\{ \sum_{j=1}^{p} \sum_{i=1}^{n} (\tilde{x}_{ij} - \sum_{m=1}^{M} a_{im} b_{jm})^2 \right\}$$

(b) For each element $(i,j) \in O$ set $\tilde{x}_{ij} \leftarrow \sum_{m=1}^{M} \hat{a}_{im} \hat{b}_{jm}$

(c) compute the objective

$$\sum_{(i,j) \in O} (x_{ij} - \sum_{m=1}^{M} \hat{a}_{im} \hat{b}_{jm})^2$$

3. Return the estimated missing entries $\hat{x}_{ij}, (i,j) \notin O$

Clustering Methods

In general, cluster observations on the basis of the features in order to identify subgroups among the observations. Or we can cluster features on the basis of the observations in order to discover subgroups among the features (transposing the matrix).

K-means clustering:

The idea behind K-means clustering is that a good clustering is one for which the within cluster variation is as small as possible. minimize $\{\sum_{k=1}^{K} W(C_k)\}$ with $W(C_k) = \frac{1}{|C_k|}\sum_{i,i' \in C_k}\sum_{j=1}^{p}(x_{ij}-x_{i'j})^2$

1. Randomly assign a number from 1 to K to each of the observations. These serve as initial cluster assignment for the observations.
2. Iterate until the cluster assignment stop changing:
   - (a) For each cluster, compute the cluster centroid: the vector the p features means for the observation in the k^{th} cluster
   - (b) Assign each observation to the cluster whose centroid is closest (Euclidian distance).

K mean find a local optima, it should then be repeated → then select the best solution.

Hierarchical clustering

bottom-up or agglomerative clustering (dendogram - upside down tree). The term hierarchical refers to the fact that clusters obtained by cutting the dendogram at a given height are necessarily nested with the clusters. However, on an arbitrary data set, this assumption of hierarchical structure might be unrealistic. i.e the true structure are not nested, in the sense that the best division into three groups doesn't result from taking the best division into two groups and splitting up one of those groups

1. Begin with n observations and a measure (Euclidean) of all $\binom{n}{2} = \frac{n(n-1)}{2}$ pairwise dissimilarities. Treat each observation as its own cluster.
2. For $i = n, n-1, \dots, 2$
   - (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that least dissimilar. Fuse them.
   - (b) Compute the new pairwise inter-cluster dissimilarities among the i-1 remaining clusters.

Linkage

Complete → Maximal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B and record the largest of the dissimilarities

Single → Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observation in cluster A and the observation in cluster B, and record the smallest of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are one at a time

Average → Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observation in cluster B, and record the average of these

Centroid → Dissimilarity between the centroid for cluster A (a mean of vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable inversion

Instead of Euclidian distance for dissimilarity we can use correlation based distance / similarity (if correlated).

- Validating the model → compute p-values but not that great, no consensus on the method.
- Suppose outlier that do not belong into each cluster, how to cluster the rest without outlier → mixture model → soft version of K-means.