CHAPTER 9 UNSUPERVISED LEARNING
What is unsupervised learning?

- We don’t observe outcome values/labels except feature variables $X$.
- The goal is to understand and characterize the structures/distributions within feature data (similar to density estimation).
- It is useful for data exploration and dimension reduction.
Principal component analysis

- It is one of the most popular methods examining intrinsic structure of $X$.
- Aims to identify the combination which explains most of data variances (principal directions).
- The method only uses the second moments of data.
Calculation of principal components

- Suppose $X_1, ..., X_n$ to be mean-zero observed feature values in $\mathbb{R}^p$.

- We look for a matrix $\tilde{V}_{p \times q}$ such that $\tilde{V} = (\tilde{V}_1, ..., \tilde{V}_q)$ have orthogonal unit columns and the projection of $X$ onto $\tilde{V}$-space contains the most variability, i.e.,

$$\sum_{i=1}^{n} \| X_i - \tilde{V} \tilde{V}^T X_i \|_2^2$$

is minimal.

- Write $X = (X_1, ..., X_n)^T$. This is equivalent to minimizing

$$\text{Trace} \left( X - X \tilde{V} \tilde{V}^T \right) \left( X - X \tilde{V} \tilde{V}^T \right)^T$$

$$= \text{Trace} \left( I - \tilde{V} \tilde{V}^T \right)^T X^T X \left( I - \tilde{V} \tilde{V}^T \right).$$
Solutions to PCA

– Consider singular-value decomposition:

\[ X_{n \times p} = U_{n \times p} \text{diag}\{d_1, ..., d_p\} V_{p \times p}, \]

\[ d_1 \geq d_2 \geq ... \geq d_p \geq 0. \]

– We aim to minimize

\[ \text{Trace}(I - \tilde{V}\tilde{V}^T)^T V^T D^T V (I - \tilde{V}\tilde{V}^T) = \text{Trace}(I - \tilde{V}\tilde{V}^T) V^T D^T V. \]

– It shows \( \tilde{V} \) consists of the first \( q \)-columns of \( V \).

– The first \( q \)-columns of \( UD^T \), the projection on the top \( q \) principal directions, are called principal components.
The choice of PC #

- Small # PC is believed to be sufficient.
- Useful for data visualization and dimension reduction.
- Normally, the top $q$ PCs are retained if their explained variation is at least 70% or 80%.
Latent component analysis

- Assume that the feature variables are multiple indirect measurements of a few independent latent sources.
- Recovering these latent sources understands feature structures and achieves dimension reduction.
- Two approaches: factor analysis and independent component analysis.
- Both assume
\[ X_{p \times 1} = A_{p \times q} S_{q \times 1} + \epsilon_{p \times 1}, \]
\[ S: \text{ latent components, } \epsilon: \text{ independent noise.} \]
Factor analysis

- $S = (S_1, ..., S_1)$ is assumed to be from Gaussian and uncorrelated.

- It is different from the PCA as FA looks for the common factors explaining maximal variability, while the PCA considers the total variability.

- Estimation is based on

$$Cov(X) = AA^T + diag(var(\epsilon_1), ..., var(\epsilon_n))$$

- SVD plays important role in estimating $A$; or the MLE is used for estimation.

- $A$ is subject to rotation.

- FA is most popular in social science.
Independent component analysis

- The difference from FA is that $S$ is not necessarily Gaussian but its components are independent.

- Actually, the solution to estimating $A$ and $S$ relies on minimizing some entropy so uses higher moments other than the second one.

- See ICA.pdf.
Multidimensional scaling

- This method projects original $X$ to a much lower-dimensional space.
- It is (probably only) useful for viewing $X$.
- The goal of the projection is to ensure pairwise distances before and after projections to be consistent as much as possible.
- Minimize (Stress function)

$$\left[ \sum_{i \neq j} (d(X_i, X_j) - \|Z_i - Z_j\|)^2 \right]^{1/2} .$$

- Gradient descent algorithm is used of minimization.
- Can be modified to add weights to each pair or just keep distance ranks to be consistent.
Variants of MDS

- Sammon mapping:
  \[ \sum_{i \neq j} \left( \frac{(d_{ij} - \|Z_i - Z_j\|)^2}{d_{ij}} \right). \]

- Ranks:
  \[ \sum_{i \neq j} (d_{ij} - g(\|Z_i - Z_j\|)) \]

  is minimized for both $Z$’s and $g$’s.

- Example of MDS.
Cluster analysis

– Search for clusters of subjects so that within-cluster subjects are most similar but between-cluster subjects are most different.

– Look for a map: \( \mathcal{C} : \{1, ..., n\} \rightarrow \{1, ..., K\} \) from subject ID to cluster ID.

– Within-cluster distance (loss):

\[
\frac{1}{2} \sum_{i,j=1}^{n} \sum_{k=1}^{K} I(\mathcal{C}(i) = \mathcal{C}(j) = k) d(X_i, X_j).
\] (1)

– Between-cluster distance (loss):

\[
\frac{1}{2} \sum_{i,j=1}^{n} \sum_{k=1}^{K} I(\mathcal{C}(i) = k, \mathcal{C}(j) \neq k) d(X_i, X_j).
\] (2)

– Either minimize (1) or maximize (2).
K-means cluster analysis

– Applies when the distance is the Euclidean distance.
– The within-cluster distance is equivalent to

\[
\sum_{i=1}^{n} \sum_{k=1}^{K} I(C(i) = k) \| X_i - m_k \|^2,
\]

where \( m_k \) is the \( k \)-cluster mean.
– An iterative procedure is used to update \( m_k \) and cluster membership.
K-medoids cluster analysis

- It applies to general proximity matrix.
- Replace mean $m_k$ by the point $X_i$ (medoid) in the same cluster which has the least summed distance from the other points in the cluster.
- Iteratively update the medoid and cluster membership.
Hierarchical clustering

– Either agglomerative (bottom-up) or divisive (top-down).
– At each level, either merge two clusters or split clusters in an optimal sense.
– The way of defining between-cluster distance includes single linkage, complete linkage and group average.
– The output is called a dendrogram.
Linkage definitions

- **Single linkage:**

  \[
  d(C_1, C_2) = \min_{i \in C_1, j \in C_2} d(X_i, X_j).
  \]

- **Complete linkage:**

  \[
  d(C_1, C_2) = \max_{i \in C_1, j \in C_2} d(X_i, X_j).
  \]

- **Group average:**

  \[
  d(C_1, C_2) = \frac{1}{n_1 n_2} \sum_{i \in C_1, j \in C_2} d(X_i, X_j).
  \]

- Single linkage can produce clusters with very large diameter and the complete linkage is oppositive; while the group average is a compromise between the two extremes.