inzva Applied AI Week 5

Uğur Ali Kaplan

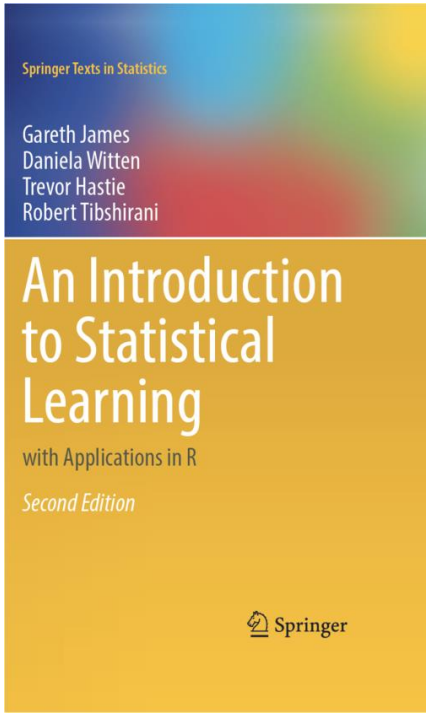
Contents

- Supervised and Unsupervised Learning Overview
- Determining Baselines
- Missing Value Imputation
- Class Imbalance
- Feature Selection and Extraction
- Ensemble Models
- Experiment Tracking with MLFlow
- Hyperparameter Optimization with Ray Tune

The Book

- Slides are based on the “An Introduction to Statistical Learning”, 2nd Ed.

- https://www.statlearning.com



Supervised and Unsupervised Learning

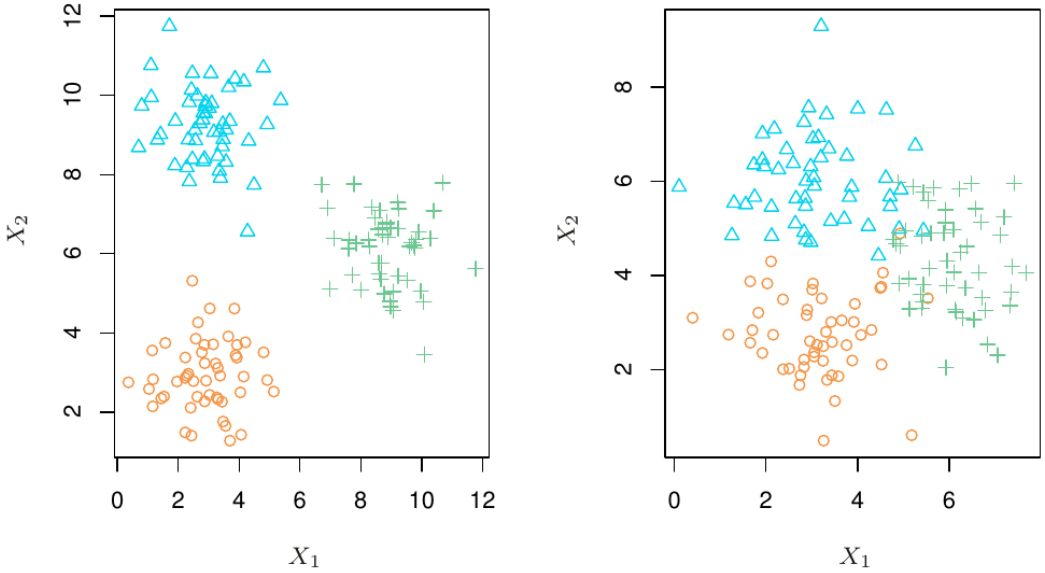
- Supervised
  - Goal: Predicting an output based on the input

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix} \qquad \mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad Y = f(X) + \epsilon.$$

- Unsupervised
  - Goal: Uncovering relationships and structure from data

$$\mathbf{X} = \begin{pmatrix} x_{11} & x_{12} & \dots & x_{1p} \\ x_{21} & x_{22} & \dots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \dots & x_{np} \end{pmatrix}$$

Clustering



K-Means Clustering

- Pre-defined number of clusters
    - K Clusters
- Each sample belongs to one cluster
- Clusters are non-overlapping

$$\underset{C_1,\ldots,C_K}{\text{minimize}} \left\{ \sum_{k=1}^{K} \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2 \right\}$$

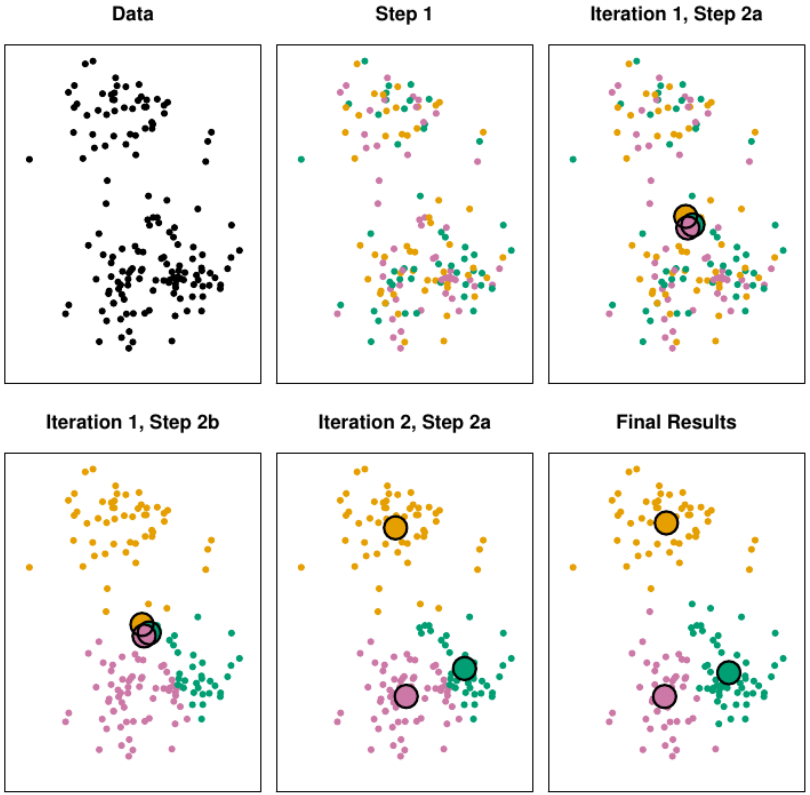
K-Means Clustering

- Running the following algorithm multiple times:

Algorithm 12.2 *K-Means Clustering*

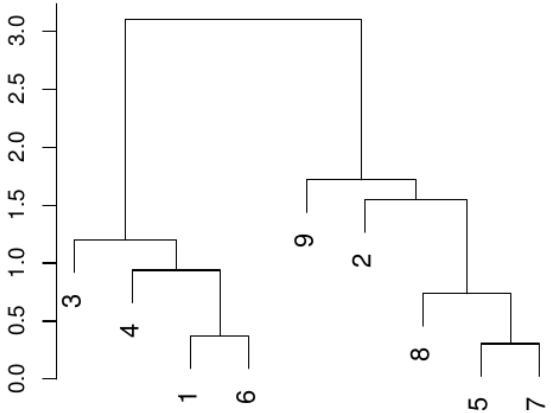
1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
   (a) For each of the K clusters, compute the cluster *centroid*. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
   (b) Assign each observation to the cluster whose centroid is closest (where *closest* is defined using Euclidean distance).

K-Means Clustering

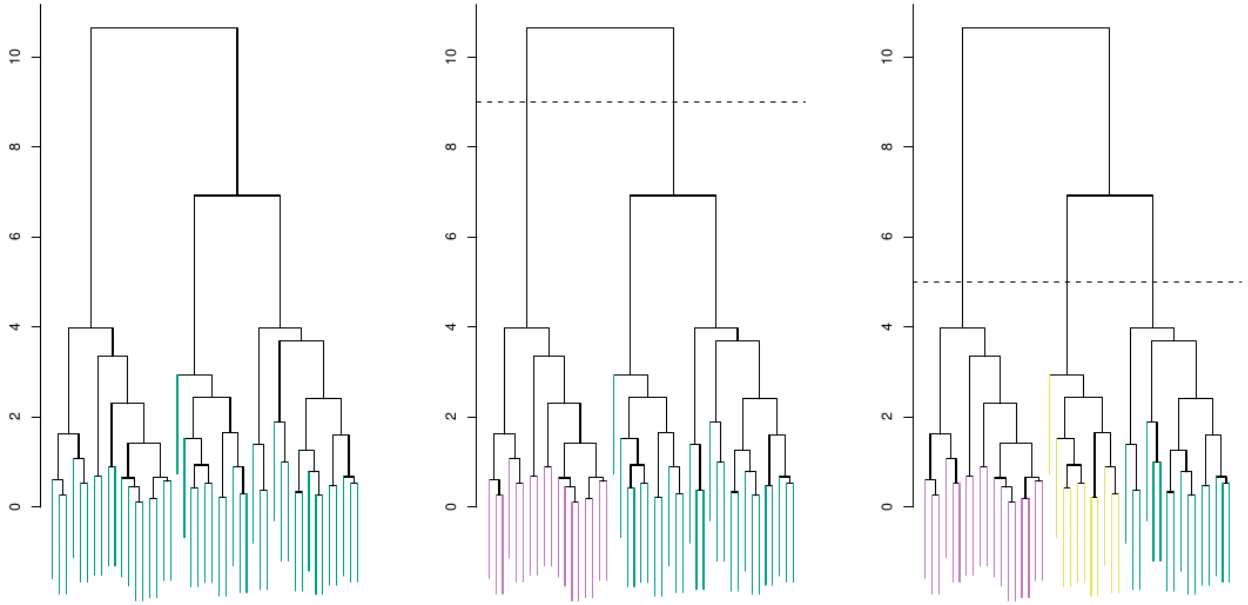


Hierarchical Clustering

- Unknown number of clusters
- Dendrogram



Hierarchical Clustering



Hierarchical Clustering

- Dissimilarity Measure
    - *Euclidean*
- Each sample is a cluster
- Similar clusters are fused together

Hierarchical Clustering

Algorithm 12.3 *Hierarchical Clustering*

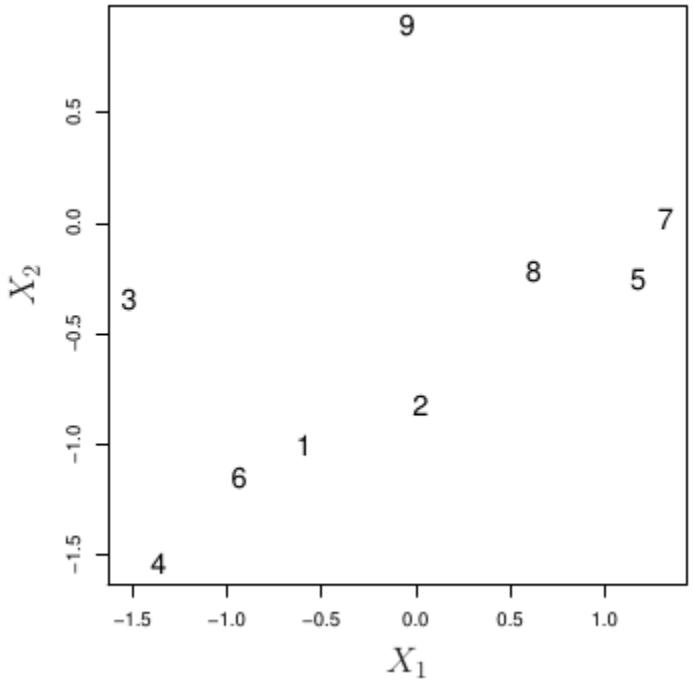
1. Begin with n observations and a measure (such as Euclidean distance) of all the $\binom{n}{2} = n(n-1)/2$ pairwise dissimilarities. Treat each observation as its own cluster.

2. For $i = n, n-1, \ldots, 2$:

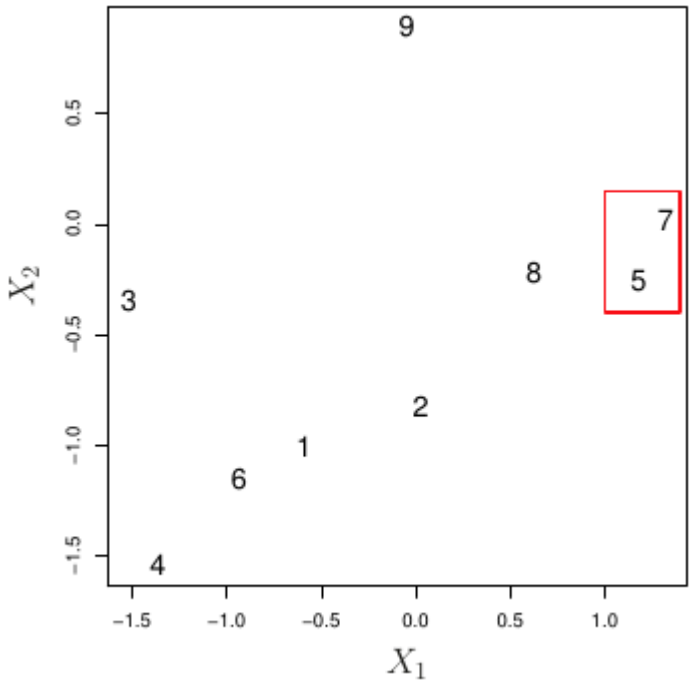
   (a) Examine all pairwise inter-cluster dissimilarities among the i clusters and identify the pair of clusters that are least dissimilar (that is, most similar). Fuse these two clusters. The dissimilarity between these two clusters indicates the height in the dendrogram at which the fusion should be placed.

   (b) Compute the new pairwise inter-cluster dissimilarities among the $i-1$ remaining clusters.

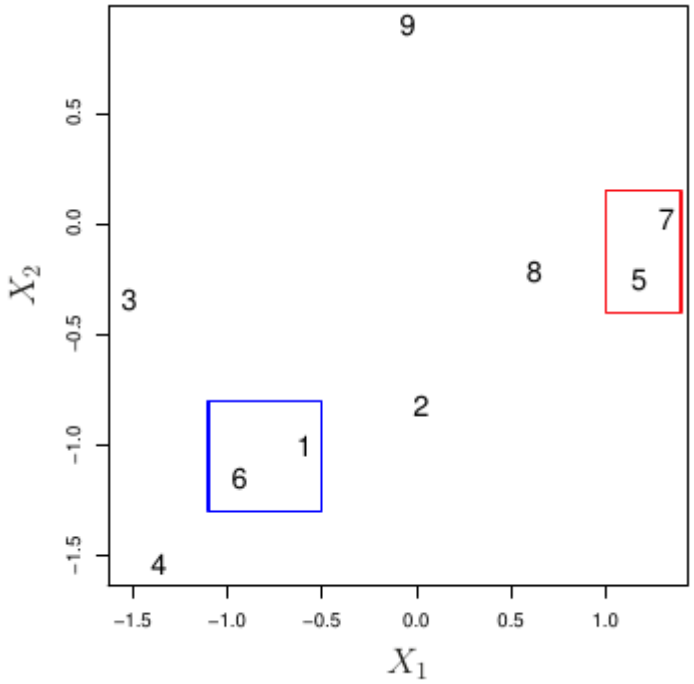
Hierarchical Clustering



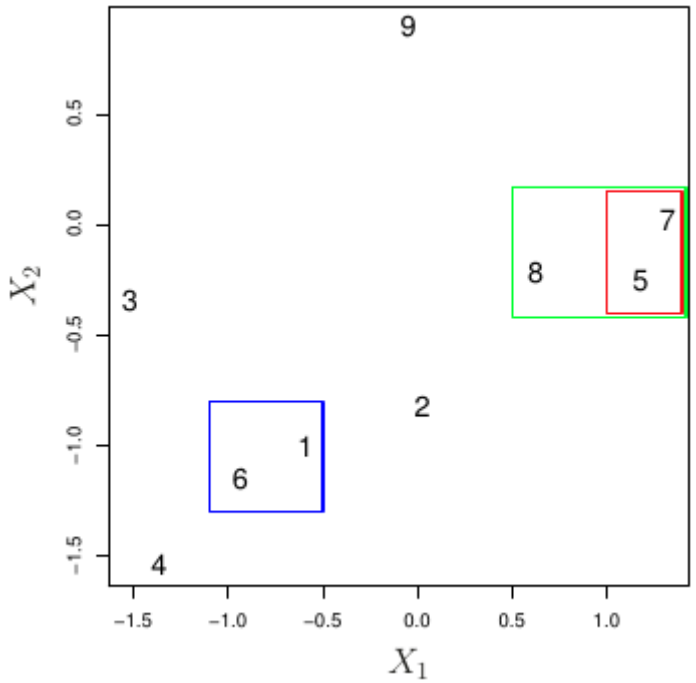
Hierarchical Clustering



Hierarchical Clustering



Hierarchical Clustering



Hierarchical Clustering

Linkage	*Description*
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 these dissimilarities.
Single	Minimal intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *smallest* of these dissimilarities. Single linkage can result in extended, trailing clusters in which single observations are fused one-at-a-time.
Average	Mean intercluster dissimilarity. Compute all pairwise dissimilarities between the observations in cluster A and the observations in cluster B, and record the *average* of these dissimilarities.
Centroid	Dissimilarity between the centroid for cluster A (a mean vector of length p) and the centroid for cluster B. Centroid linkage can result in undesirable *inversions*.

TABLE 12.3. *A summary of the four most commonly-used types of linkage in hierarchical clustering.*

Naïve Bayes

- Bayes Rule

$$P(y|\mathbf{x}) = \frac{P(\mathbf{x}|y)P(y)}{P(\mathbf{x})}$$

- Naïve Assumption

$$P(\mathbf{x}|y) = \prod_{\alpha=1}^{d} P(x_\alpha|y)$$

- For more details: http://www.cs.cornell.edu/courses/cs4780/2018fa/lectures/lecturenote05.html

Imputation

- Univariate Methods
  - Mean
  - Median
  - Mode
  - Constant
- Multivariate Methods
  - kNN
  - Regression

Imbalance

- Oversampling
  - SMOTE
- Undersampling
  - Tomek Links
  - Edited Nearest Neighbors
- Oversampling and Undersampling

Feature Extraction and Selection

- Feature Selection
  - Filter Methods
  - Wrapper Methods
  - Embedded Methods
- Feature Extraction
  - Principal Component Analysis

Principal Component Analysis

- Unsupervised
- Low-dimensional representation of data (Dimensionality Reduction)
- Principal components
  - Summarizing a set of correlated variables with less number of representative variables, which explain most of the variance in the features.
- Principal component directions
  - Directions in feature space where original features are highly variable

Practical

- Experiment Tracking
  - MLFlow
- Hyperparameter Tuning
  - Ray Tune
  - Ax
  - Optuna

Homework

- Use Optuna to optimize our model. Compare results with Ax.
- Find a small dataset from Kaggle. Compare XGBoost, CatBoost and LightGBM in terms of speed and performance. For performance, use relevant metrics, not just accuracy.