May 15, 2023

Machine Learning Methods and Applications

Week 10. Unsupervised learning | Clustering and dimension reduction

Remember

- Boosting methods handle the problem with bias, while bagging methods handle the problem with variance in the models.
- The main idea of boosting is to add new models to ensemble sequentially.
- Learning rate is a kind of hyperparameter of boosting models that controls the size of step of optimization algorithm to find the optimal minimum value of the loss function.

Unsupervised learning

Unsupervised learning

- The goal of unsupervised learning is to find some patterns in unlabeled data.
- Supervised learning is used when you want to make predictions on labeled data..
- There are two application areas in unsupervised learning:
    - **Clustering** is the process of finding homogeneous subgroups.
    - **Dimension reduction** is a method to decrease the number of features.
- Unfortunately, there is no way to check the work in unsupervised learning like in supervised learning because we do not know the true answer.

Clustering

Clustering

- Clustering refers to a very broad set of techniques for finding distinct subgroups, or clusters in a data set.
- The observations within each group are quite similar to each other, while observations in different groups are quite different from each other.
- Some clustering methods: k-means, k-medoids, hierarchical clustering.

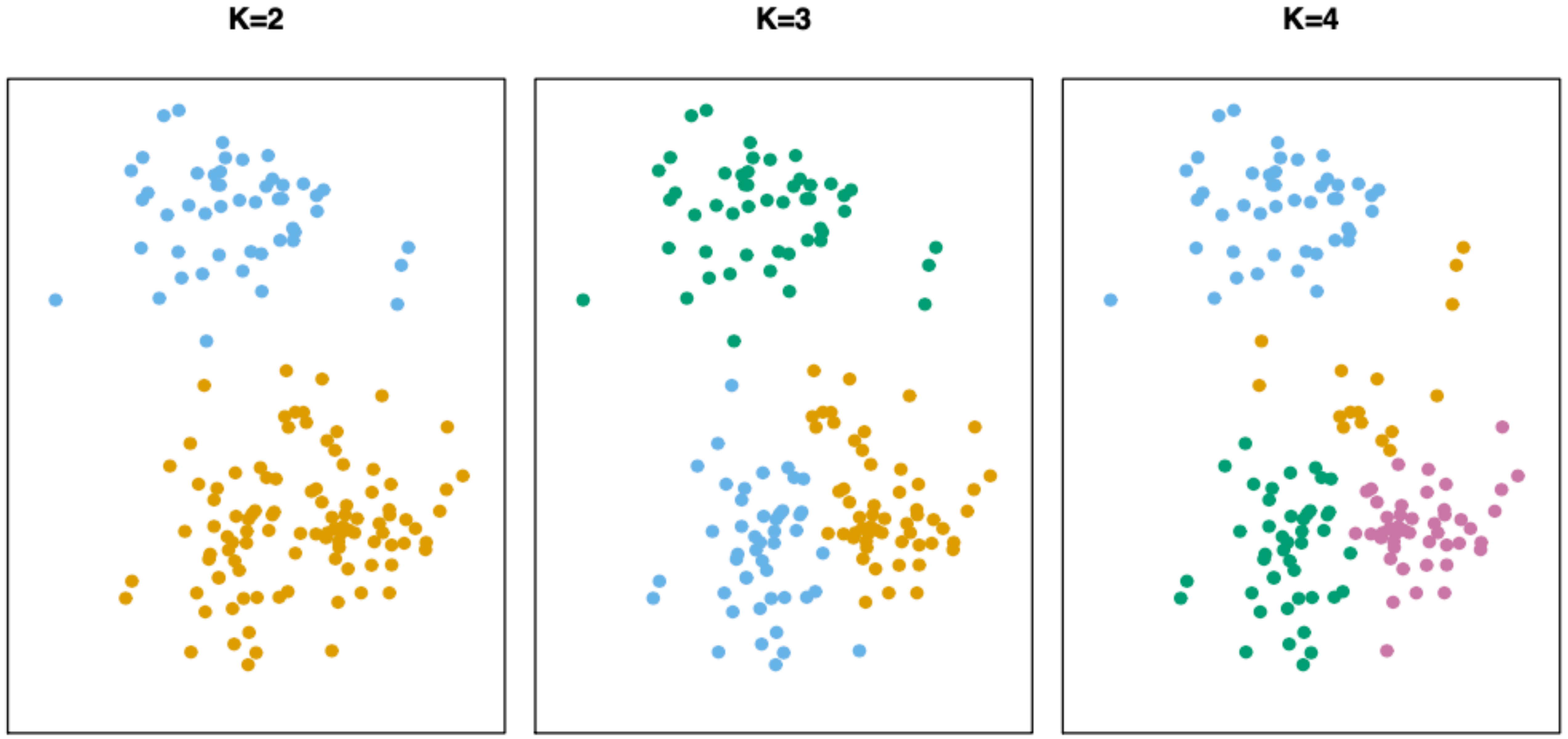
Examples

- **Market segmentation** by defining subgroups of people who might be more receptive to a particular from of advertising, or more likely to purchase a particular product.
- Determining natural groups of houses for sale based on size, number of bedrooms, etc.

k-means method

k-means method

k-means is a method used to find pre-specified number of non-overlapping clusters within a population.



k is the number of clusters
It is a ………….

k-means mechanism

1. First specify the desired number of clusters k.

2. Let $C_1, C_2, \ldots, C_k$ denote sets containing the indices of the observations in each cluster.

    - $C_1 \cup C_2 \cup \ldots \cup C_k = \{1,2,...,n\}$
    - $C_k \cap C_{k'} = \varnothing$ for all $k \neq k'$

3. Minimize $\sum_{k=1}^{K} W(C_k) = \sum_{k=1}^{K} \sum_{x_i \in C_k} (x_i - \mu_k)^2$ where $W(C_k)$ is within cluster variation by using a distance metric, e.g. Euclidean, Manhattan, …

k-means algorithm

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 chancing:

    - 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.

    - Assign each observation to the cluster whose centroid is closest (where closest is defined using Euclidean distance).

Dimension reduction

Dimension reduction

- Dimension reduction is used to reduce the dimension of the data with minimum loss of information.
- It is need for finding the structure in features (feature extraction), aiding in visualization.

Curse of dimensionality

- **Dimension**: Columns in the dataset that represent features of the row points.
- **Dimensionality**: The number of features/columns characterizing the dataset.
- **Curse of dimensionality**: As the dimensionality of the data grow, the feature space grows rapidly.

Curse of dimensionality

Cons

- Higher computational cost to handle high-dimensional data.
- Correlated and irrelevant features may degrade performance of ML models.
- Difficult interpretation and visualization of the data.

Curse of dimensionality

Solutions

- **Feature engineering** requires the domain knowledge.
- **Dimension reduction methods** such as Principal component analysis.

Principle component analysis

Principle component analysis

PCA is used to reduce the dimension of the data and to make smaller dimension for less risk of overfitting.

Application

See the R codes on the course GitHub repository!

The materials of today's lecture will be available on **GitHub**.
Feel free to contact me via e-mail: mustafacavus@eskisehir.edu.tr