Machine Learning

2. week

- Distance-based grouping
  - K-means clustering
  - K-NN classifier

Distance-based Grouping

The most important aim of machine learning is to find similar data points and to group them together in the same cluster or class.

The similarity term is often expressed by inverse of distance.

Many classification and clustering algorithms use distance measure to group data.

Similarity

Similarity vs. Distance

The most commonly used distance measure is Euclidean distance. Distance between two points with n-dimension is calculated as follows:

$$d_{ab} = \|a-b\| = \sqrt{(a_1-b_1)^2+(a_2-b_2)^2+\ldots+(a_n-b_n)^2}$$

The similarity is usually inverse of distance.

$$s_{ab} = \frac{1}{1+\|a-b\|}$$

Similarity vs. Distance

Similarity measure does not have to associated with the distance.

In the literature, there are many similarity measures such as distance-based, probabilistic and feature-based.

There are also many other distance measures such as Mahalanobis, Manhattan and Chebyshev.

Distance-based Clustering

Clustering is a unsupervised grouping process for most similar data points.

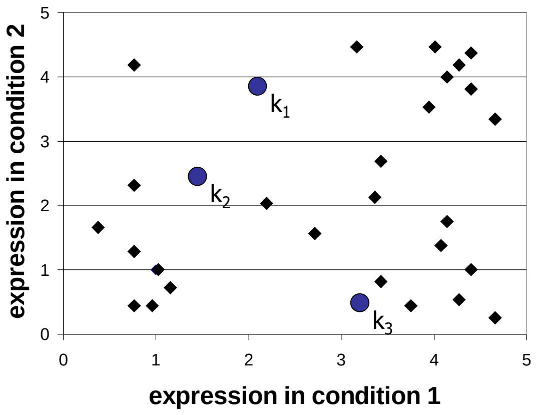
Here, we can change “most similar” term with "least distance".

K-means algorithm is the most popular and used method in distance-based clustering.

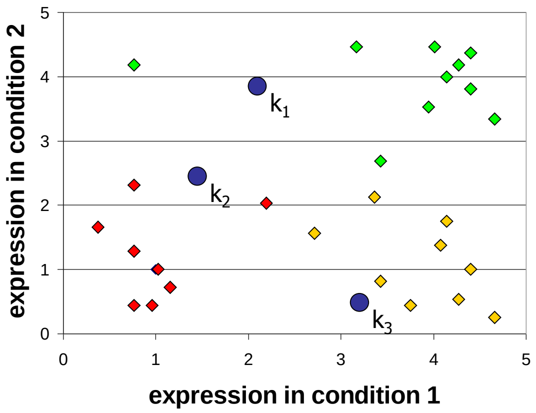
K-means Algorithm

- Pick a number (K) of cluster centers
- Assign every data point to its nearest cluster center
- Move each cluster center to the mean of its assigned data points
- Repeat 2-3 until convergence

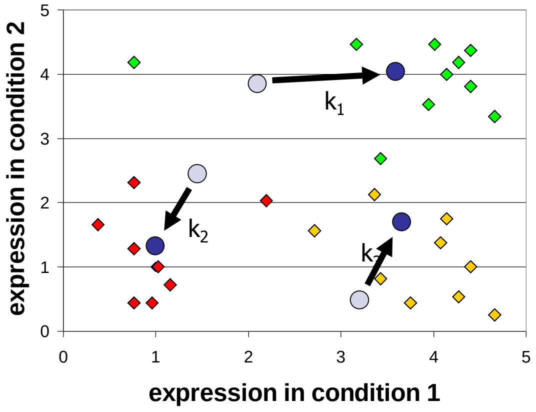
K-means Algorithm



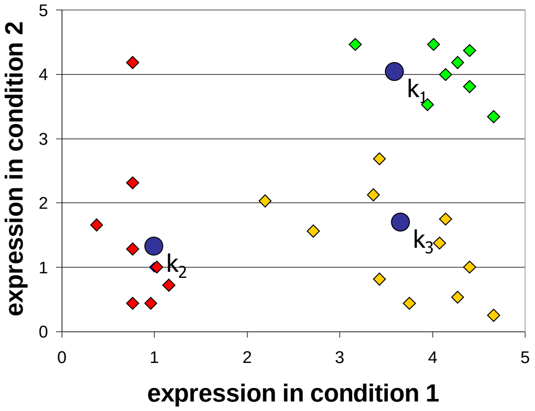
K-means Algorithm



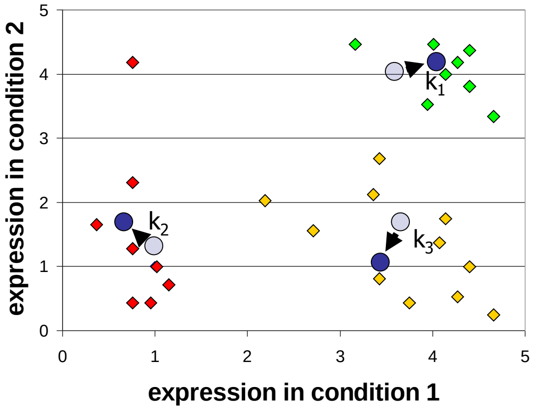
K-means Algorithm



K-means Algorithm



K-means Algorithm



K-means

Convergence often depends on a maximum iteration number or J cost function given as:

$$J = \sum_{i=1}^{K} \left(\sum_{k} \left\| x_k - c_i \right\|^2 \right)$$

The cost function should be its minimum value when all cluster centers are at the optimum position.

K-means Disadvantages

- Because it starts with random points, each result obtained is not always optimum.
- The number of clusters is requested from an outsource as in many clustering algorithms.
- When clusters look like concave (non-convex) shape, it is not successful.
- It uses hard clustering, but in practice, real data clusters are usually overlapping (fuzzy).
- It is sensitive to outliers.

MATLAB Application

```
>edit Kmeans_ornek.m
```

In this code, a K-means experiment can be performed on a randomly generated dataset.

Presentation Task

- AGNES
- DIANA
- BIRCH
- STING
- OPTICS
- CURE
- DBSCAN
- CLARANS

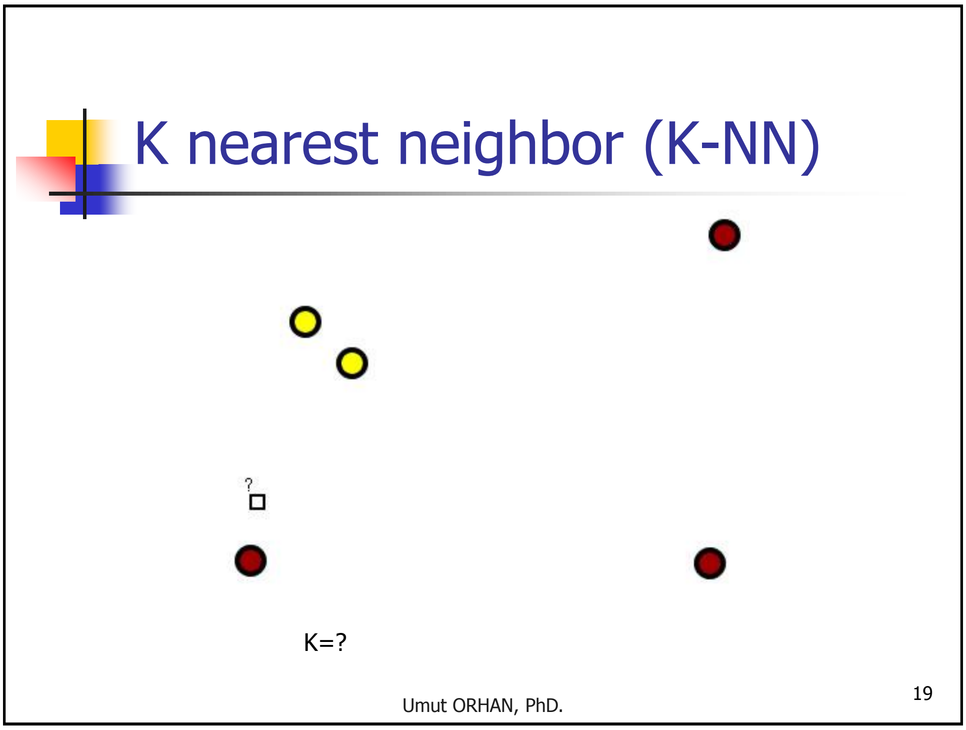
Distance-based Classification

Classification is the grouping process which uses similarities in both feature space and class information.

K nearest neighbor (K-NN) method is the most popular classification method using distance between data points.

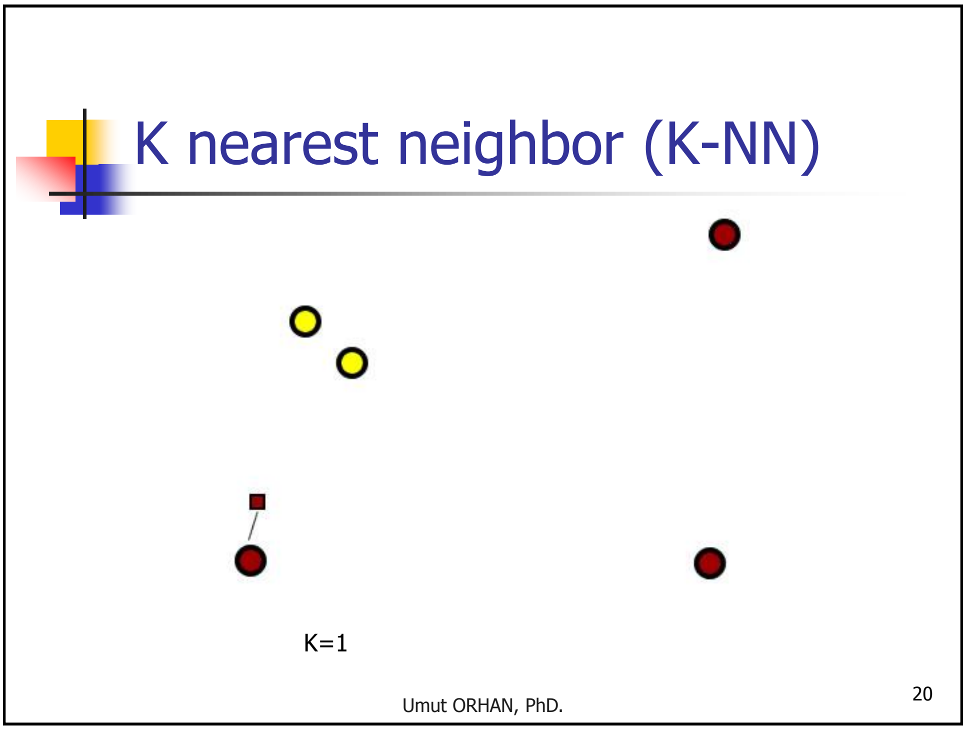
K-NN Algorithm

- K value is requested from the outside
- Distances are calculated from unknown point to known data points.
- Calculated distances are arranged in ascending order, K distances at top of the list are chosen (nearest ones).
- Among classes of the selected data points, dominant class is determined as class of the unknown point.



K nearest neighbor (K-NN)

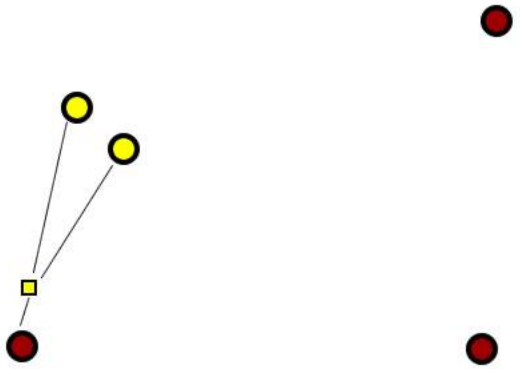
K=?



K nearest neighbor (K-NN)

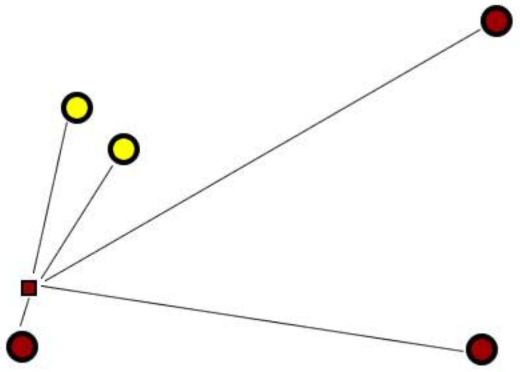
K=1

K nearest neighbor (K-NN)



K=3

K nearest neighbor (K-NN)



K=5

K-NN Disadvantages

- It is a memory based classifier. The data points should be kept in memory continuously. When the data set is very large, the computation time is getting worse.
- Because all properties are included in calculation, redundant or irrelevant features may adversely affect classification.
- In terms of performance, it usually remains behind of the advanced classification techniques such as artificial neural networks.

MATLAB Application

```
>edit KNN_ornek.m
```

The K-NN algorithm tests are performed on a randomly generated dataset. Student should see that different K values can produce different results in some dataset.

Application Task

Compare classification performances of the distance measures given below.

- Euclidean
- Manhattan
- Chebyshev
- Mahalanobis