CLUSTERING ALGORITHM

Shingchern D. You

Clustering

- Want to find out male/female based on weights and heights (credit: https://towardsdatascience.com/everything-you-need-to-know-about-scatter-plots-for-data-visualisation-924144c0bc5)

Clustering algorithm

- Goal of clustering algorithm: To find the cluster every data point belongs
- What need to know before doing clustering
  - **Number of clusters must be known** (in this case $k = 2$)
  - Sometimes it is tricky to choose a good value of k without *a priori* knowledge

Clustering algorithm

- The algorithm can be divided into two sub-problems
- If centroid points for each class is known
  - Assign an observation (sample) $\boldsymbol{x} \in R^p$ to the class with shortest distance (to centroid)
- If all observations in a cluster is known
  - Easy to find centroid of the cluster (simple average)

Clustering algorithm

- But, we don't know either one
- We formulate the problem in math
- Let $S = \{s_1, s_2, \cdots, s_k\}$, and s_j be a set of data points with centroid $\boldsymbol{\mu}_j$
- We want to find

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{\boldsymbol{x} \in s_i} \|\boldsymbol{x} - \boldsymbol{\mu}_j\|$$

- In the above Eq., we need a **distance** function

Clustering algorithm

- Finding the optimal solution is **NP-hard** for Euclidean space (even for 2 clusters)
- Therefore, we seek for **heuristic** algorithms (likely obtain a local optimum)
- One well-known algorithm is **k-means** (the term "mean" here is average)
- As k-means is an iterative algorithm, its solution is related to initial conditions

K-means algorithm

Input: cluster number k and input samples $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_N$

Initialize $\boldsymbol{\mu}_1, \ldots, \boldsymbol{\mu}_k$ (randomly pick k $\boldsymbol{x}_i$ out of N)

Repeat

For $i = 1$ to N // Assignment step

if $\left\|\boldsymbol{x}_i - \boldsymbol{\mu}_j\right\|$ is minimal distance $b(i,j) \leftarrow 1$
else $b(i,j) \leftarrow 0$

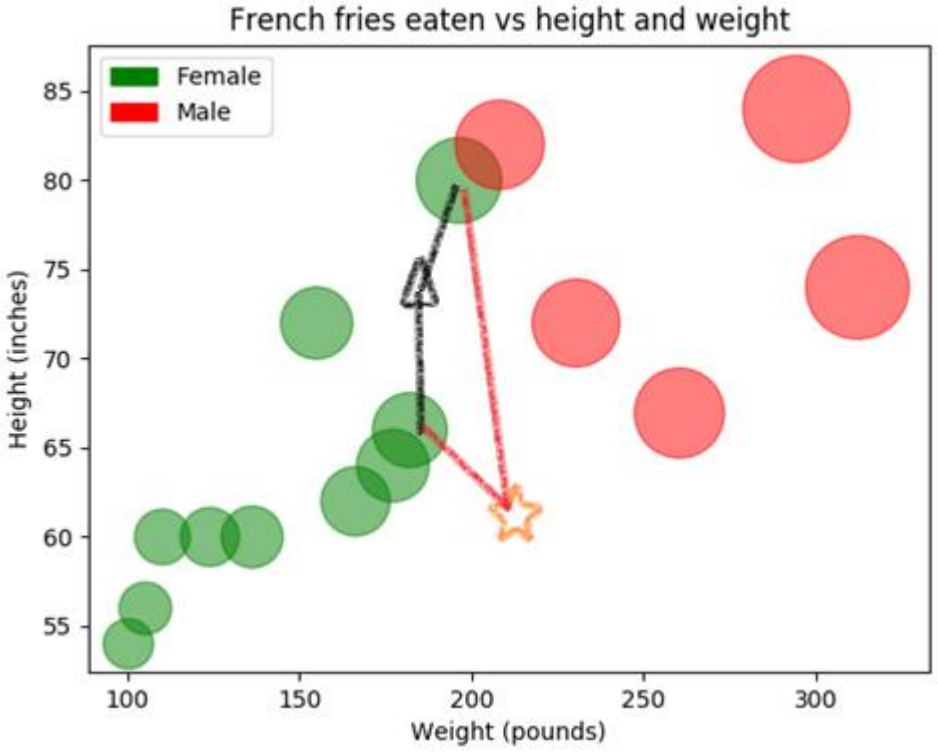
For $j = 1$ to k // Update step

$$\boldsymbol{\mu}_j \leftarrow \frac{\sum_{i=1}^{N} b(i,j)\boldsymbol{x}_i}{\sum_{i=1}^{N} b(i,j)}$$

Until converge

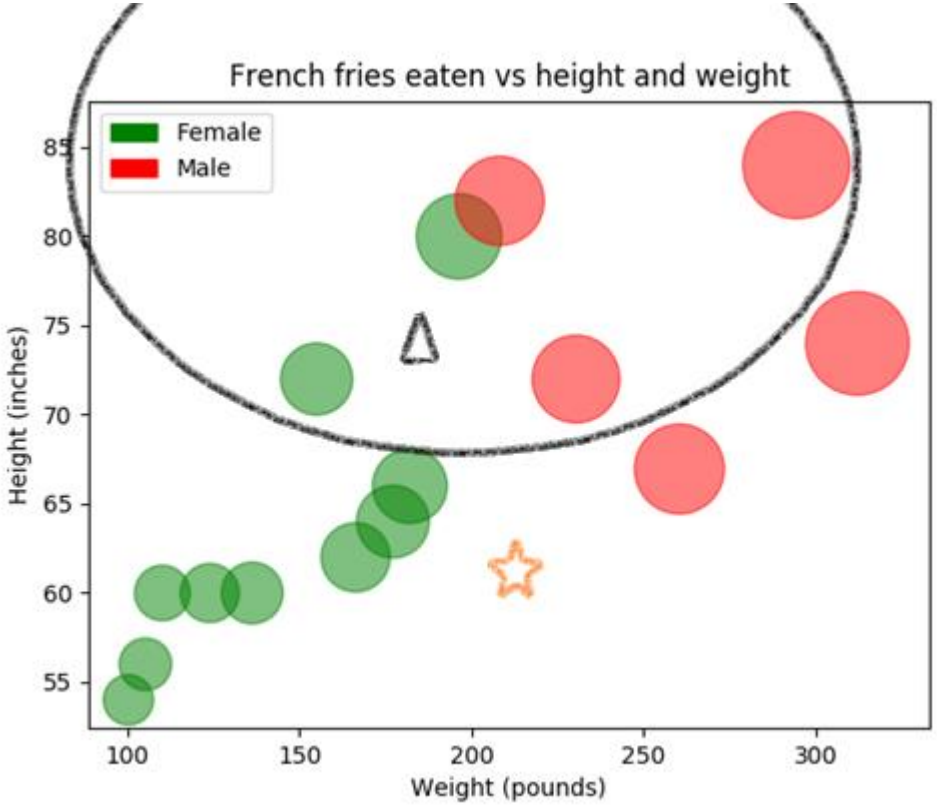
K-means algorithm

- Triangle & star are two centroids
- Use centroids to assign each sample to a class



K-means algorithm

- Compute centroid for each cluster



K-means algorithm

- The two steps in k-means algorithm are assignment and update
- The algorithm is a variation of generalized EM (expectation maximization) algorithm
- Recall each step can be solved easily
- With iteration, solution is found (may not be optimal solution)
- We can prove that k-means algorithm converges

K-means algorithm

- Some problems remains
  - How to determine k
  - Any better method to find initial centroids
  - Empty cluster
- Previous algorithm is a **batch k-means** algorithm, but **online k-means** algorithm also available
- An extension of k-means is ISODATA (Iterative Self-Organizing Data Analysis Technique), dynamically determine *k* (need additional hyper-parameters)

K-means algorithm applications

- VQ (Vector Quantization) in signal processing
  - Meaning of quantization
  - Scalar quantization vs vector quantization
  - VQ for **data compression**
- Cluster analysis
- A step in **feature learning** (or dictionary learning)
  - Check keyword: **bag of feature**

Applications

- How to avoid local minimum?
  - Use multiple runs with smallest cost function
- How to choose a good value of K
  - Plot cost function (choose elbow point, may not visually find a good one)
  - Think of the purpose of running k-means, and then we can do further analysis (say, cost vs performance of having more clusters or less clusters)
  - Usually manually chosen (not auto chosen)
  - By visualization (e.g., use PCA to draw 2-D plot)

Alternative clustering algorithms

- Adaptive k means
- Neural networks
  - ART (adaptive resonant theory)
  - SOFM (self organized feature map)
- More