Clustering

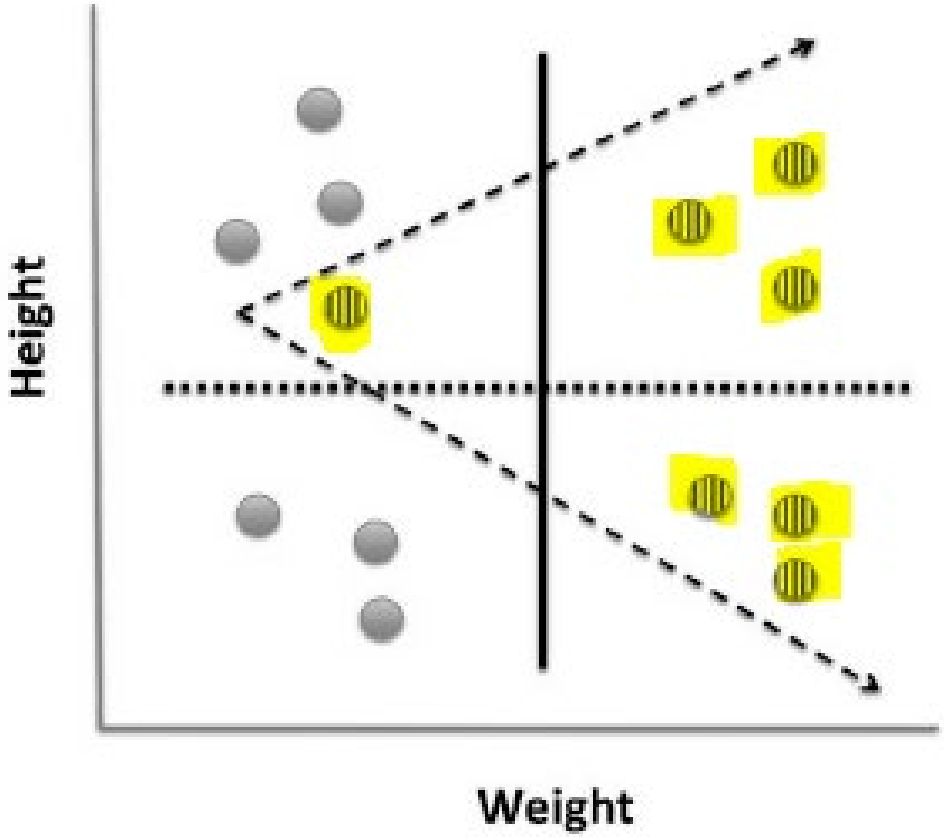
Chapter 25

Unsupervised learning techniques

- Clustering
  - K-means
  - Fuzzy k-means
  - Hierarchical clustering
  - Mixture of Gaussians
- Anomaly detection
- Neural networks
- Latent variable learning techniques

What is clustering?

- Process of organizing objects into groups “with” similar members
- Finding “hidden structure”



- Optimization problem

What are we optimizing?

- Minimize **dissimilarity** of clusters while constraining the number of clusters to some maximum
  - Why? The minimum dissimilarity would be each element in its own cluster
- What is **dissimilarity**?

$$dissimilarity(C) = \sum_{c \in C} variance(c)$$

- What is the **variability** of the cluster?
  - Not normalized
  - A large cluster may have a large variability

$$variance(c) = \sqrt[2]{\sum_{e \in c} distance(mean(c), e)^2}$$

Hierarchical clustering
(not in text)

- **Agglomerative** or “bottom up”
  - Start by assigning each node to its own cluster
  - Find the *closest* pair of clusters and merge them
  - Repeat until the desired number of clusters is reached
- **Divisive** or “top down”
  - Start with all nodes in one cluster
  - Split clusters until desired number is reached

What do we mean by the closest neighbor

- **Single-linkage** is the shortest distance between of any element in one cluster to any element in another cluster
- **Complete-linkage** is the longest distance between of any element in one cluster to any element in another cluster
- **Average-linkage** considers the average distance of any element in one cluster to any element of another cluster

Example

	BOS	NY	CHI	DEN	SF	SEA
BOS	0	206	963	1949	3095	2979
NY		0	802	1771	2934	2815
CHI			0	966	2142	2013
DEN				0	1235	1307
SF					0	808
SEA						0

{NY} {BOS} {CHI} {DEN} {SF} {SEA}

{NY, BOS} {CHI} {DEN} {SF} {SEA}

{NY, BOS, CHI} {DEN} {SF} {SEA}

{NY, BOS, CHI} {DEN} {SF, SEA}

Single-linkage: {NY, BOS, CHI, DEN} {SF, SEA}

Complete-linkage: {NY, BOS, CHI} {DEN, SF, SEA}

Quick summary

- Hierarchical clustering
  - Dendogram (tree structure) shows structure of hierarchy
    - Used to select number of clusters
  - Flexible linkage metrics
  - Deterministic
  - Most useful on small sets with known number of target clusters
  - **Slow**
    - Performance: $O(n^3)$
    - Space: $O(n^2)$

More about **centroids**

- A centroid is an element, belonging to the cluster, whose feature vector contains the mean of feature vectors of all members in the cluster
- “center of mass” for the cluster
- That tells us that feature vectors must be numeric

K-means clustering

- Partition examples into **k** clusters
- Each example is in the cluster where it is closest to the centroid
- Dissimilarity of the set of clusters is minimized

```
Randomly choose k examples as centroids
While true
   Create k clusters by assigning each example to the closest
   centroid
   Compute new centroids for all clusters
   If no changes to centroids from previous iteration
      STOP
```