Data Mining and Learning

Jarek Szlichta
http://data.science.uoit.ca/

What is Data Mining?

- Approximate terminology, though there is some overlap:
  - **Data(base) operations**
    - Executing specific operations or queries over data
  - **Data mining**
    - Looking for patterns in data
  - **Machine Learning**
    - using data to make inferences or predictions

Big Data, Big World..

- Early data mining success stories
  - Victoria's Secret
  - Walmart
  - "Beer and diapers"

Data Mining Techniques

- We will cover data mining on market-basket data
  - with patterns being frequent itemset and finding association rules
- Examples of other types of data:
  - graphs (of the node-and-link variety),
  - streams,
  - text (known as “text mining”)
- Examples of other types of patterns:
  - looking for similar items,
  - looking for structural patterns in large networks
  - looking for clusters and/or anomalies

Market Basket Analysis

Market-Basket Data

- **Originated with retail data, specifically grocery stores, where a market basket is a set of items purchased together**
- More generally, market basket data is any data where there is
  - a fixed (possibly very large) set of items,
  - and a (usually large) number of transactions consisting of one or more of the items

Market Basket Data Examples

- Items: groceries, Transaction: grocery cart
- Items: online goods, Transaction: (virtual) shopping cart
- Items: college courses, Transaction: student transcript
- Items: students, Transaction: party
- Items: movies, Transaction: person
- Items: symptoms, Transaction: patient
- Items: words, Transaction: document

Frequent Itemsets

- **Sets of items that occur together frequently in transactions**
  - How large is a "set"?
  - What does frequently mean?
- Look for sets containing at least min-set-size items, may also constrain max-set-size
  - **Support: # transactions containing set / total # transactions**
  - **Look for sets with support > support-threshold**

Frequent Itemsets Example

- Transactions
  - T1: beer, eggs, milk
  - T2: beer, diapers, milk
  - T3: chips, eggs
  - T4: eggs, milk
  - T5: beer, chips, diapers, milk
- Assume min-set-size = 2, support-threshold = 0.3
  - Frequent itemsets?

Frequent Itemsets Example

- Transactions
  - T1: beer, eggs, milk
  - T2: beer, diapers, milk
  - T3: chips, eggs
  - T4: eggs, milk
  - T5: beer, chips, diapers, milk
- Assume min-set-size = 2, support-threshold = 0.3
  - Frequent itemsets?
  - Answer: beer/milk (0.6), beer/diapers (0.4), diapers/milk (0.4), eggs/milk (0.4), beer/diapers/milk (0.4)

Computing Frequent Itemsets with SQL

- Assume Table Shop(TID, item)
  - Frequent itemsets of two, support-threshold = 0.3
  - S1 and S2 are aliases to the same table Shop
    - Technique is based on self-join over table Shop

```
Select S1.item, S2.item
From Shop S1, Shop S2
Where S1.TID = S2.TID
   and S1.item < S2.item
Group by S1.item, S2.item
Having count(*) >
    (Select count(distinct TID)*0.3
   From Shop)
```

Computing Frequent Itemsets with SQL

- Table Shop(TID, item)
  - Frequent itemsets of three, support-threshold = 0.3

```
Select S1.item,S2.item,S3.item
From Shop S1,Shop S2,Shop S3
Where S1.TID = S2.TID And S2.TID = S3.TID
   And S1.item < S2.item
   And S2.item < S3.item
Group By S1.item,S2.item,S3.item
Having count(*) >
    (Select count(distinct TID)*0.3
   From Shop)
```

Association Rules

- **Set1 → Set2: when Set1 occurs in a transaction, Set2 often occurs in the same transaction**
- Commonly limit to looking for rules where Set2 is a single item
  - How large is Set1?
  - What does “often” mean?

Association Rules

- Look for sets Set1 containing at least min-set-size items, may also constrain max-set-size
- **Confidence: # transactions containing Set1 and Set2 / # transactions containing Set1**
  - Look for sets with confidence > confidence threshold
- Still consider Support: # transactions containing Set1 / total # transactions
  - Look for sets with support > support threshold (i.e., Set1 should be frequent itemset)

Association Rules Example

- Transactions
  - T1: beer, eggs, milk
  - T2: beer, diapers, milk
  - T3: chips, eggs
  - T4: eggs, milk
- min-set-size = 1, max-set-size = 1, confidence-threshold = 0.5, support-threshold = 0.5
  - Association rules?

Association Rules Example

- Transactions
  - T1: beer, eggs, milk
  - T2: beer, diapers, milk
  - T3: chips, eggs
  - T4: eggs, milk
- min-set-size = 1, max-set-size = 1, confidence-threshold = 0.5, support-threshold = 0.5
  - Association rules?
    - For instance, Beer → Diapers (0.5; 0.5), Beer → Milk (1;0.5), Eggs → Milk (0.66;0.75), Milk → Beer (0.66;0.75), Milk → Eggs (0.66;0.75), …

Classification and Clustering

Supervised and Unsupervised Machine Learning

- **Supervised: Create a model from well-understood training data, use it for inference or prediction about other data.**
  - Examples: regression, classification
- **Unsupervised: Try to understand the data, look for patterns or structure.**
  - Examples: data mining, clustering
  - Also in-between approaches, such as semi-supervised and active learning

Classification

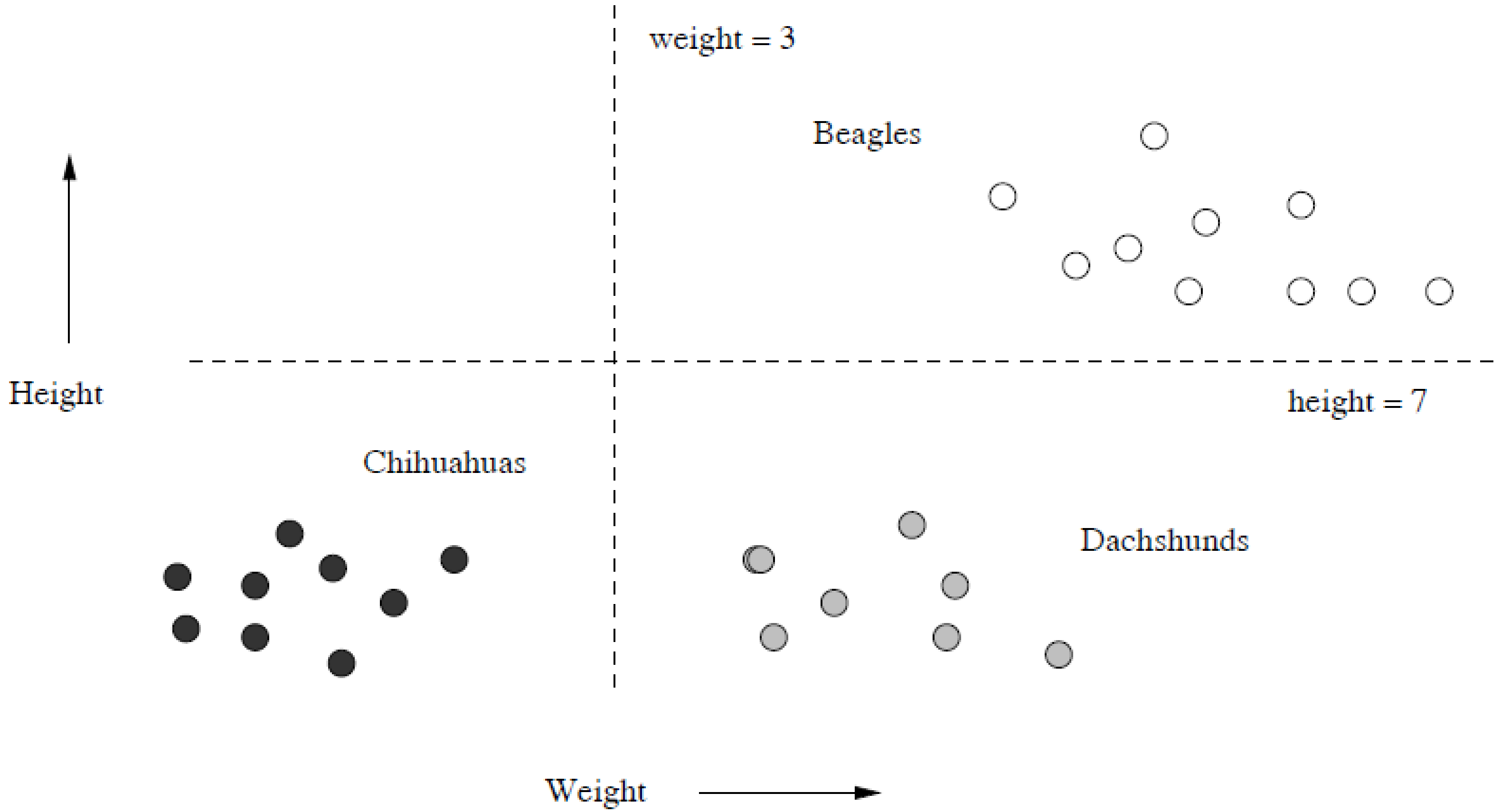
- **Goal: Given a set of feature values for an item not seen before, decide which one of a set of predefined categories the item belongs to**
  - Customer purchases
    - features: age, income, gender, zipcode, profession;
    - categories: likelihood of buying (high, medium, low)
  - Medical diagnosis
    - features: age, gender, history, symptom-1-severity, symptom-2-severity, test1result, test2result;
    - categories: diagnosis
  - Fraud detection in online purchases features:
    - item, volume, price, shipping, address;
    - categories: fraud or okay

Chihuahua, Beagles, Dachshunds..

Learning Illustrative Example

- Plot the *height and weight* of dogs in three classes: **Beagles, Chihuahuas, and Dachshunds.**
- Each pair (**x**, y) in the training set consists of:
  - Feature vector **x** of the form **[height, weight].**
  - The associated label *y* is the variety of the dog.
- An example of a training-set pair would be **([5 inches, 2 pounds], Chihuahua)**.

Heights and Weights of Certain Dogs



Decision Function

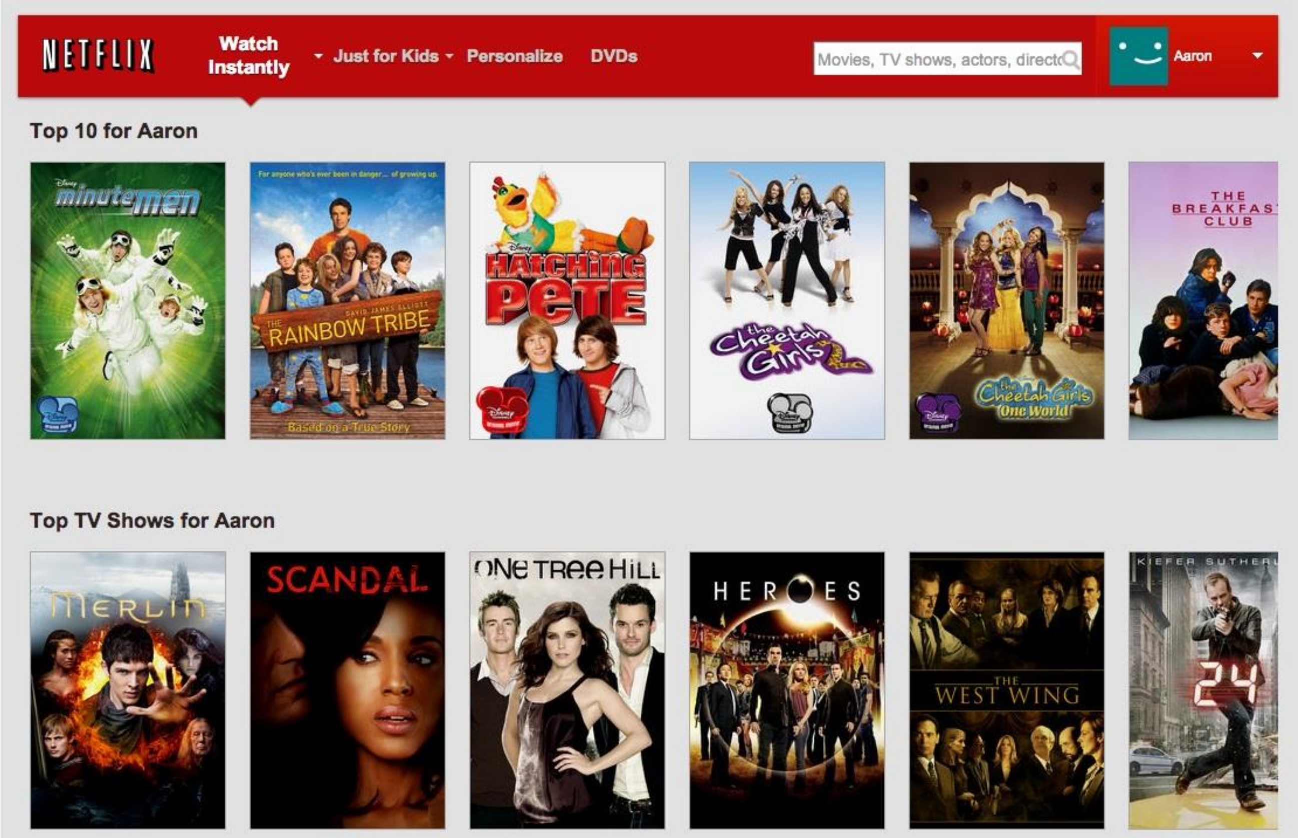
- The algorithm that implements function *f* is:

```
if (height > 7) print Beagle
else if (weight < 3) print Chihuahua
else print Dachshund;
```

- Is it supervised on unsupervised learning?

Netflix Suggestions

- Computed based on watched movies



K Nearest Neighbors - Classification

- K nearest neighbors is an algorithm that stores all available cases and classifies new cases based on a similarity measure
  - (e.g., distance function).

KNN and Distance Functions

- A case is classified by a majority vote of its neighbors, with the case being assigned to the class most common amongst its K nearest neighbors measured by a distance function.
  - If K = 1, then the case is simply assigned to the class of its nearest neighbor.

Distance functions

Euclidean $$\sqrt{\sum_{i=1}^{k}(x_i - y_i)^2}$$

Manhattan $$\sum_{i=1}^{k}|x_i - y_i|$$

Hamming Distance

- It should also be noted that all two distance measures are only valid for continuous variables
  - In the instance of categorical variables the Hamming distance must be used (outputs 0 or 1)

Hamming Distance

$$D_H = \sum_{i=1}^{k} \left| x_i - y_i \right|$$

$$x = y \Rightarrow D = 0$$

$$x \neq y \Rightarrow D = 1$$

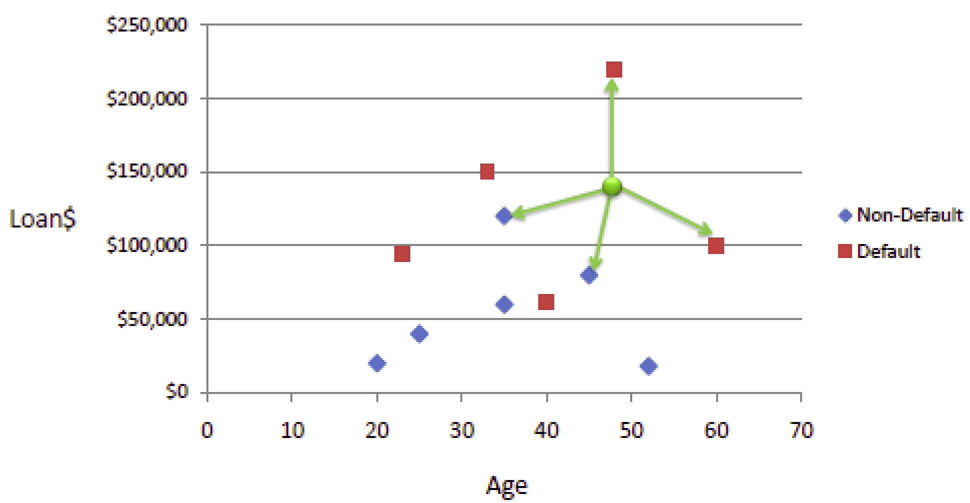
X	Y	Distance
Male	Male	0
Male	Female	1

Choosing K

- **Choosing the optimal value for K is best done by first inspecting the data**
  - In general, a larger K value is more precise as it reduces the overall noise but there is no guarantee
  - Historically, the optimal K for most datasets has been between 3-10. That produces much better results than 1NN

KNN Example

- Consider the following data concerning credit default. Age and Loan are two numerical variables (predictors) and Default is the target.
  - We can now use the training set to classify an unknown case (Age=48 and Loan=$142,000) using Euclidean distance



KNN Example Solution

- If K=1 then the nearest neighbor is the last case in the training set with Default=Y

D = Sqrt[(48-33)^2 + (142000-150000)^2] = 8000.01 >> Default=Y

Age	Loan	Default	Distance	
25	$40,000	N	102000	
35	$60,000	N	82000	
45	$80,000	N	62000	
20	$20,000	N	122000	
35	$120,000	N	22000	2
52	$18,000	N	124000	
23	$95,000	Y	47000	
40	$62,000	Y	80000	
60	$100,000	Y	42000	3
48	$220,000	Y	78000	
33	$150,000	Y	8000	1
48	**$142,000**	**?**		

Euclidean Distance

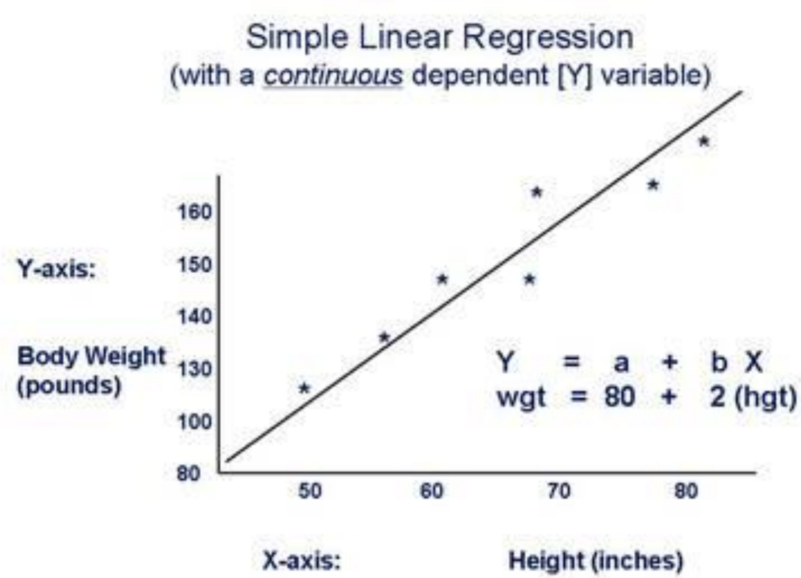
$$D = \sqrt{(x_1 - y_1)^2 + (x_2 - y_2)^2}$$

- With K=3, there are two Default=Y and one Default=N out of three closest neighbors. The prediction for the unknown case is again Default=Y

Classification using Logistic Regression

- Logistic regression uses training data to compute function f(x1 ,...,xn-1), where x1, ..., xn-1 are features, that gives probability of result xn being “yes”
  - Lots of hidden math..

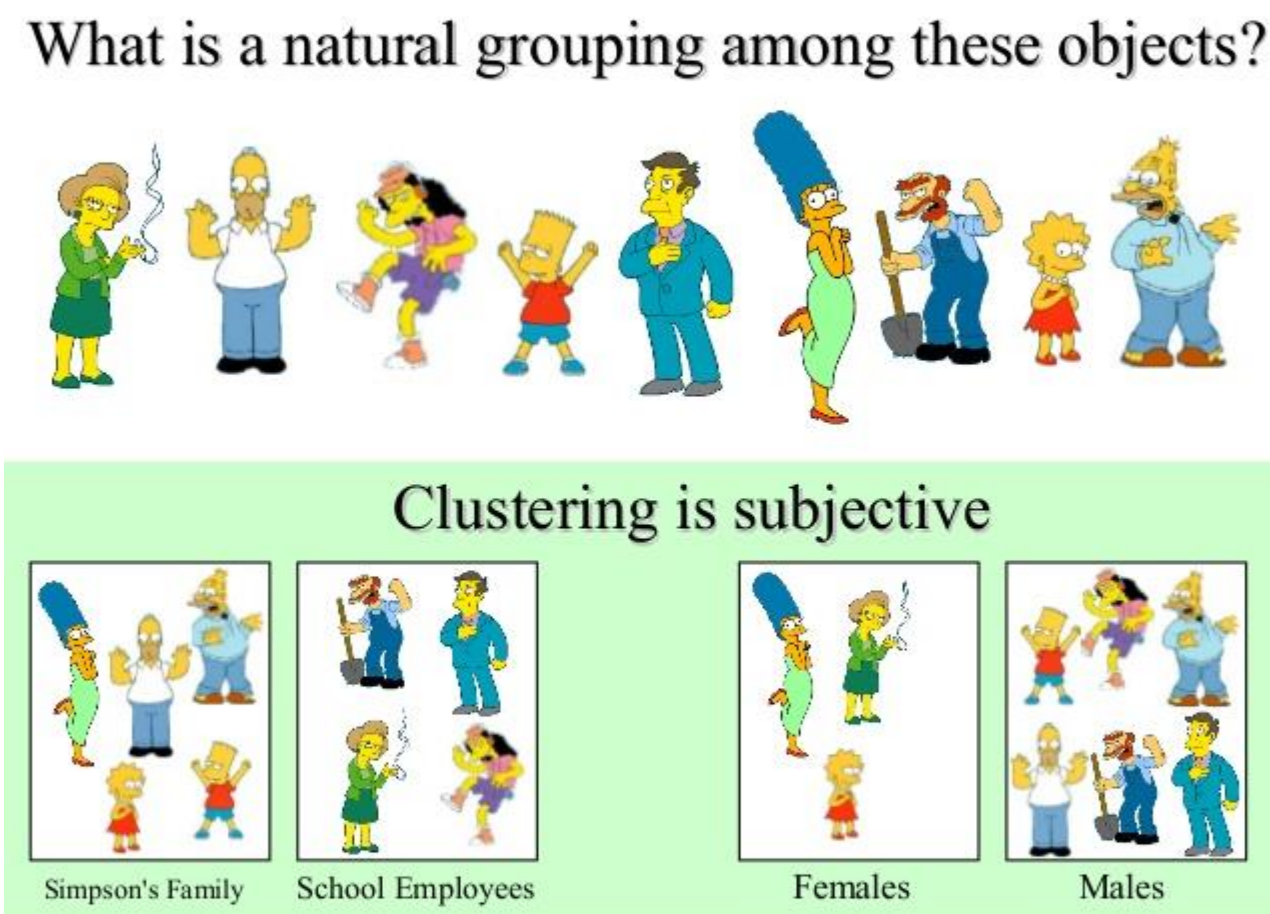
Linear Regression Example



Clustering

- Multidimensional feature space, distance metric between items
- **Goal: Partition set of items into k groups (clusters) such that items within groups are "close" to each other**
- Unsupervised, no training data

Clustering is Subjective



K-Means Clustering

- **K-Means clustering intends to partition *n* objects into *k* clusters in which each object belongs to the cluster with the nearest mean**
  - This method produces exactly *k* different clusters of greatest possible distinction
  - The best number of clusters *k* leading to the greatest separation (distance) is not known as a priori and must be computed from the data

K-Means Objective

- The objective of K-Means clustering is to minimize total intra-cluster variance, or, the squared error function:

number of clusters

number of cases

centroid for cluster *j*

case *i*

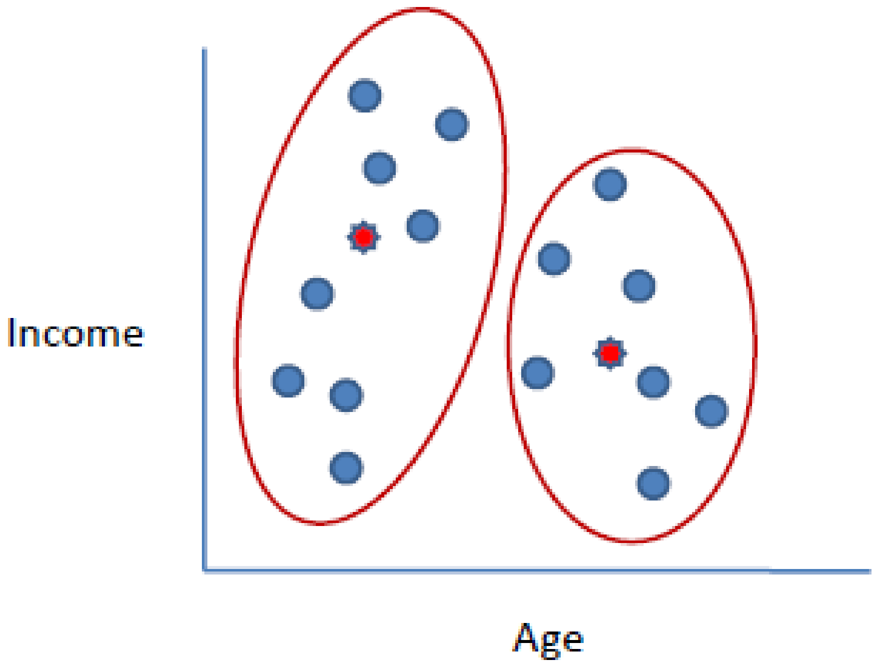
$$\text{objective function} \leftarrow J = \sum_{j=1}^{k} \sum_{i=1}^{n} \underbrace{\left\| x_i^{(j)} - c_j \right\|^2}_{\text{Distance function}}$$

K-Means Algorithm

- Clusters the data into *k* groups where *k* is predefined
  1. Select k points at random as cluster centers
  2. Assign objects to their closest cluster center according to the *Euclidean distance* function
  3. Calculate the centroid or mean of all objects in each cluster
  4. Repeat steps 2, 3 and 4 until the same points are assigned to each cluster in consecutive rounds

Clustering by Age and Income

- Data clustered by age and income



Example of Clustering

- Suppose we want to group the visitors to a website using just their age (a one-dimensional space) as follows:
  - 15,15,16,19,19,20,20,21,22,28,35,40,41,42,43,44, 60,61,65

Solution

- No change between iterations 3 and 4 has been noted.
- By using clustering, 2 groups have been identified 15-28 and 35-65
  - Initial centroids were chosen randomly

Initial clusters:

Centroid (C1) = 16 [16]
Centroid (C2) = 22 [22]

Iteration **1**:

C1 = 15.33 [15,15,16]
C2 = 36.25 [19,19,20,20,21,22,28,35,40,41,42,43,44,60,61,65]

Iteration **2**:

C1 = 18.56 [15,15,16,19,19,20,20,21,22]
C2 = 45.90 [28,35,40,41,42,43,44,60,61,65]

Iteration **3**:

C1 = 19.50 [15,15,16,19,19,20,20,21,22,28]
C2 = 47.89 [35,40,41,42,43,44,60,61,65]

Iteration **4**:

C1 = 19.50 [15,15,16,19,19,20,20,21,22,28]
C2 = 47.89 [35,40,41,42,43,44,60,61,65]

Quiz

- Provide a useful application of data mining.
- What K-Means algorithm is used for? Describe how K-Means algorithm works.
- What is KNN algorithm used for? How to chose the right K?

Reading List

- **Review Slides!**
- **Recommended**
  - Association rule learning
    - https://en.wikipedia.org/wiki/Association_rule_learning
    - http://www.theregister.co.uk/2006/08/15/beer_diapers
    - http://infolab.stanford.edu/~ullman/mining/assocrules.pdf
  - Classification and Clustering
    - http://www.saedsayad.com/k_nearest_neighbors.htm
    - http://www.saedsayad.com/mlr.htm
    - http://www.saedsayad.com/clustering_kmeans.htm
- **Optional**
  - http://infolab.stanford.edu/~ullman/mmds/bookL.pdf
  - This book is used in CSCI 4030, **Big Data Analytics**; course plug-in