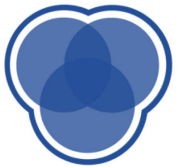
INFO 1998: Introduction to Machine Learning



CDS Education

We explore, learn, and educate big minds.

Lecture 9: Clustering and Unsupervised Learning

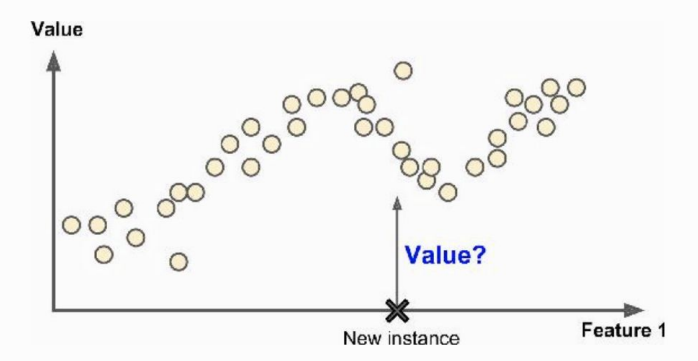
INFO 1998: Introduction to Machine Learning

CDS Education

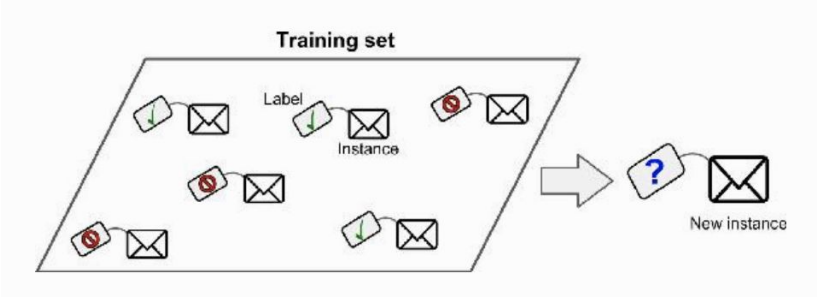
We explore, learn, and educate big minds.

Recap: Supervised Learning

- The training data you feed into your algorithm includes **desired solutions**
- Two types you've seen so far: **regressors and classifiers**
- In both cases, there are definitive "answers" to learn from



Example 1: Regressor
Predicts value

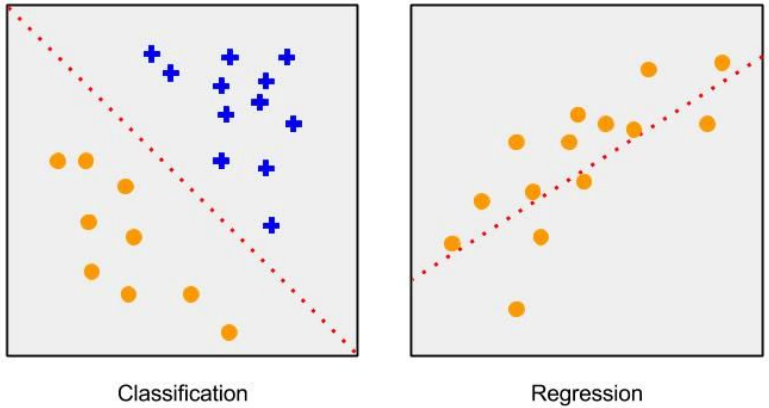


Example 2: Classifier
Predicts label

Recap: Supervised Learning

Supervised learning algorithms we have covered so far:

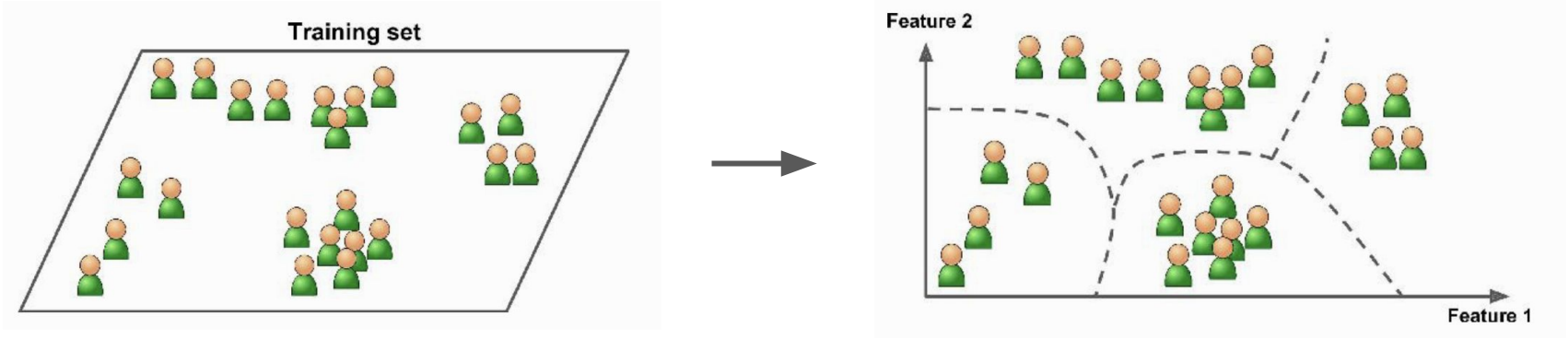
- k-Nearest Neighbors
- Perceptron
- Logistic Regression
- Decision Trees and Random Forest
- Linear Regression



What are some **limitations** of supervised learning?

Today: Unsupervised Learning

- In unsupervised learning, the training data is **unlabeled**
- Algorithm tries to learn by itself



An Example: Clustering

Unsupervised Learning

Some types of unsupervised learning problems:

1. **Clustering**
   k-Means, Hierarchical Cluster Analysis (HCA), Gaussian Mixture Models (GMMs), etc.
2. **Dimensionality Reduction**
   Principal Component Analysis (PCA), Locally Linear Embedding (LLE)
3. **Association Rule Learning**
   Apriori, Eclat, Market Basket Analysis

… *More*

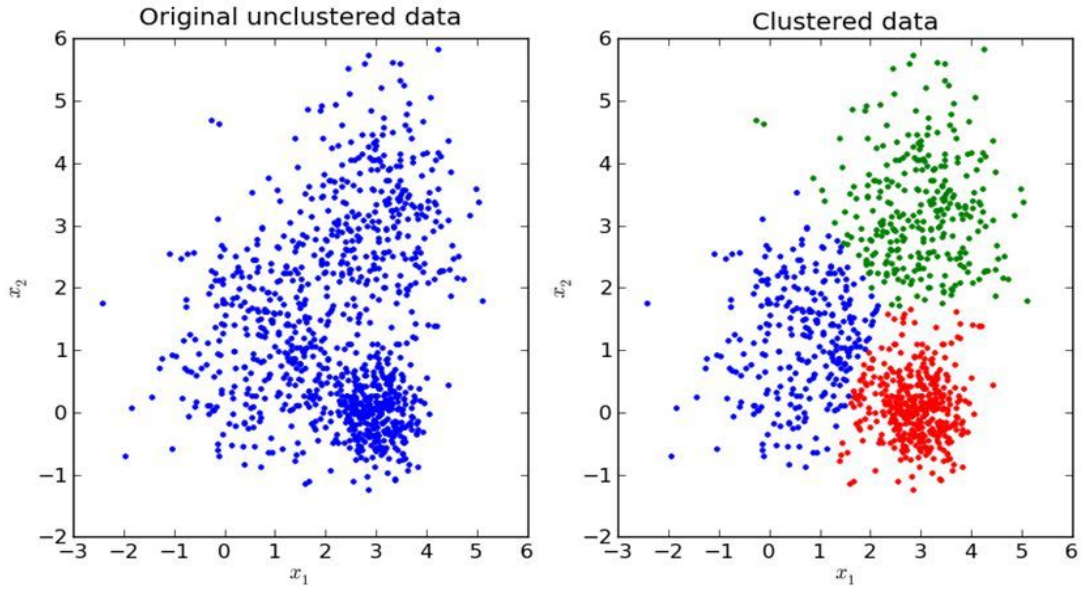
Unsupervised Learning

Some types of unsupervised learning problems:

1. **Clustering**
   k-Means, Hierarchical Cluster Analysis (HCA), Gaussian Mixture Models (GMMs), etc.
2. **Dimensionality Reduction**
   Principal Component Analysis (PCA), Locally Linear Embedding (LLE)
3. **Association Rule Learning**
   Apriori, Eclat, Market Basket Analysis

… *More*

Cluster Analysis



Cluster Analysis

- **Loose definition:** Clusters have objects which are “similar in some way” (and “dissimilar to objects in other clusters)
- Clusters are **latent variables (variables that are unknown)**
- Understanding clusters can:
  - Yield underlying trends in data
  - Supply useful parameters for predictive analysis
  - Challenge boundaries for pre-defined classes and variables

Clustering Application

Recommender Systems

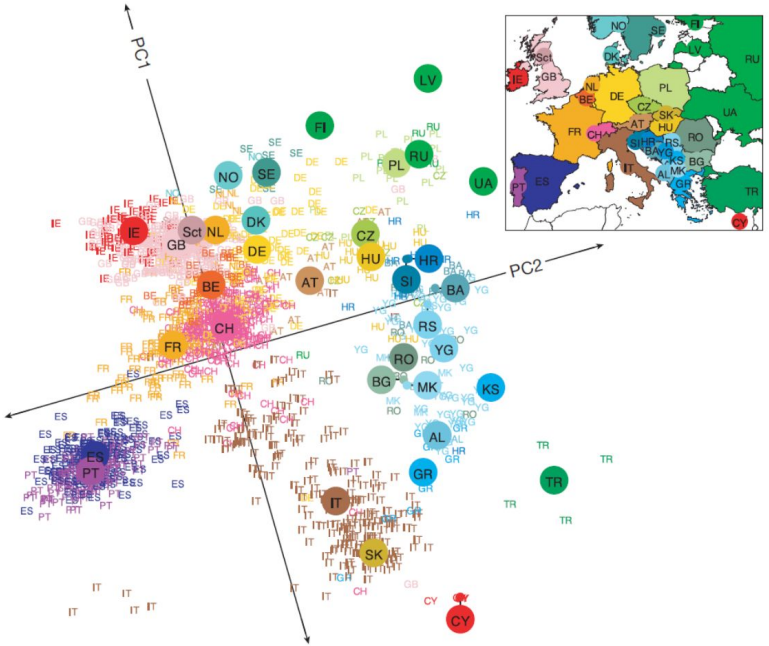
Intuition: People who are “similar”, will like the same things

A Bunch of Cool Logos

Clustering Application

Finding Population Structure in Genetic Data

From 1,387 European samples



Running Example: Recommender Systems

Use 1: Collaborative Filtering

- “People similar to you also liked X”
- Use other’s rating to suggest content

Pros	Cons
If cluster behavior is clear, can yield good insights	Computationally expensive Can lead to dominance of certain groups in predictions

Running Example: Recommend MOVIES

	Amy	Jef	Mike	Chris	Ken
The Piano	–	–	+		+
Pulp Fiction	–	+	+	–	+
Clueless	+		–	+	–
Cliffhanger	–	–	+	–	+
Fargo	–	+	+	–	+

Running Example: Recommender Systems

Use 2: Content filtering

- “Content similar to what YOU are viewing”
- Use user’s watch history to suggest content

Pros	Cons
Recommendations made by learner are intuitive Scalable	Limited in scope and applicability

Another Example: Cambridge Analytica

- Uses Facebook profiles to build psychological profiles, then use traits for target advertising
- Ex. has personality test measuring openness, conscientiousness, extroversion, agreeableness and neuroticism -> different types of ads

How do we actually perform this
“cluster analysis”?

Popular Clustering Algorithms

Hierarchical Cluster Analysis (HCA)

k-Means Clustering

Gaussian Mixture Models (GMMs)

Defining ‘Similarity’

- How do we calculate proximity of different data points?
- Euclidean distance:

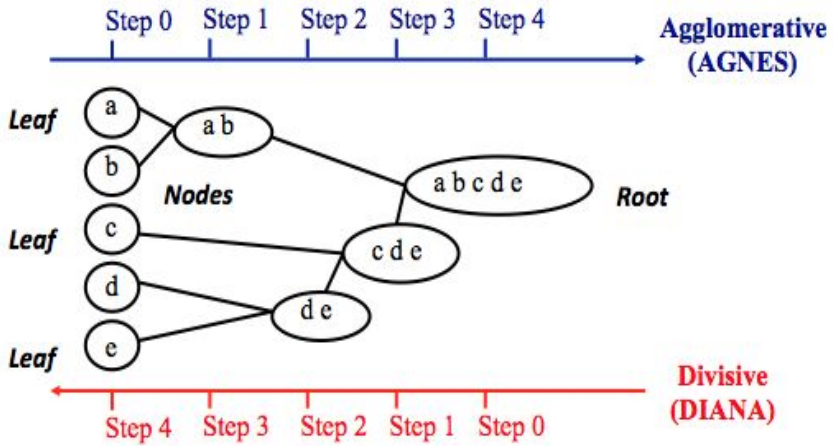
$$E(x, y) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$$

- Other distance measures:
  - Squared euclidean distance, manhattan distance

Algorithm 1: Hierarchical Clustering

Two types:

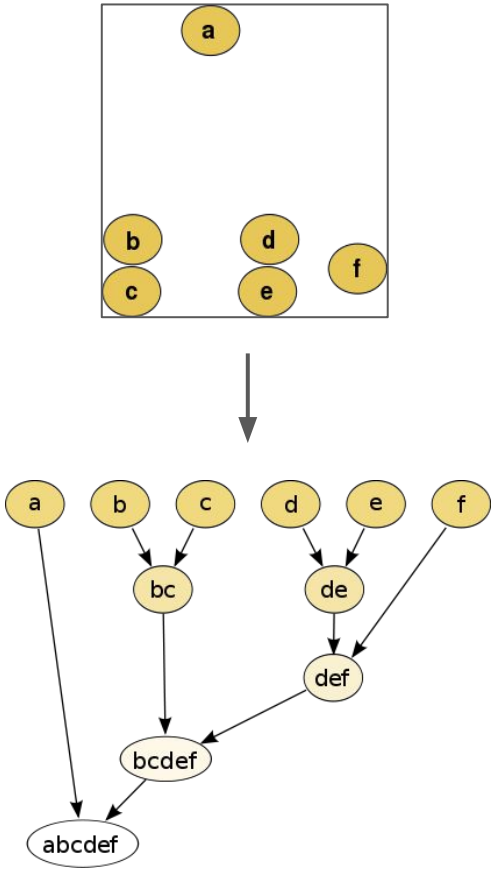
- Agglomerative Clustering
  - Creates a tree of **increasingly large** clusters *(Bottom-up)*
- Divisive Hierarchical Clustering
  - Creates a tree of **increasingly small** clusters *(Top-down)*



Agglomerative Clustering Algorithm

- Steps:
  - Start with each point in its own cluster
  - Unite adjacent clusters together
  - Repeat

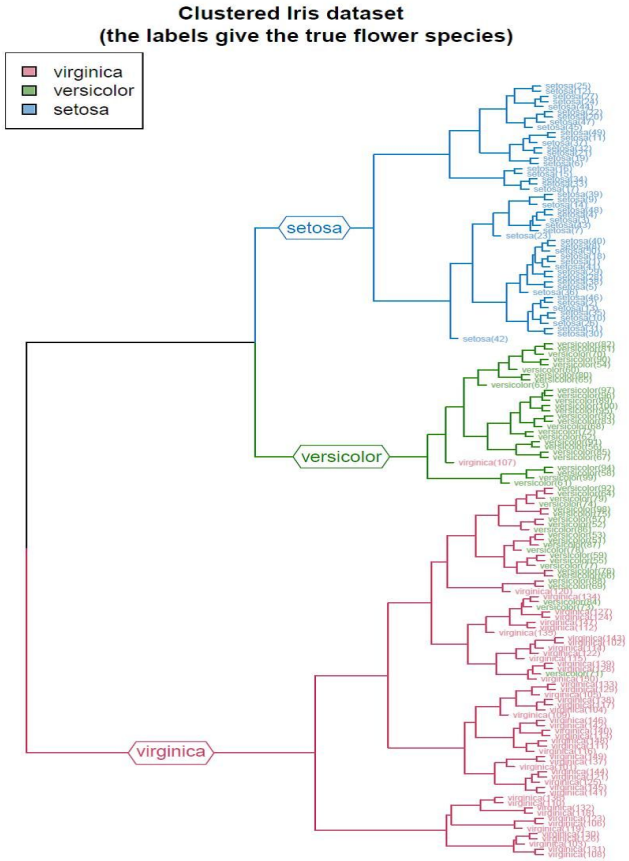
- Creates a tree of **increasingly large** clusters



Agglomerative Clustering Algorithm

How do we visualize clustering?
Using **dendrograms**

- Each width represents distance between clusters before joining
- Useful for estimating how many clusters you have



The iris dataset that we all love

Demo 1

Popular Clustering Algorithms

Hierarchical Cluster Analysis (HCA)

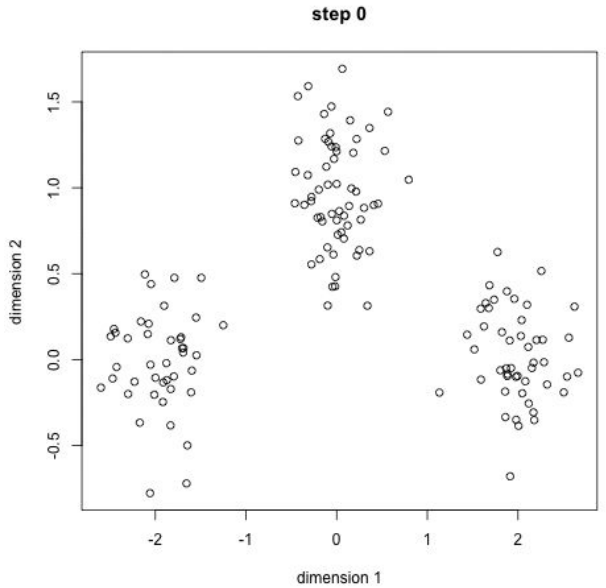
k-Means Clustering

Gaussian Mixture Models (GMMs)

Algorithm 2: k-Means Clustering

Input parameter: *k*

- ➢ Starts with *k* random centroids
- ➢ Cluster points by calculating distance for each point from centroids
- ➢ Take average of clustered points
- ➢ Use as new centroids
- ➢ Repeat until convergence



Interactive Demo: http://stanford.edu/class/ee103/visualizations/kmeans/kmeans.html

Algorithm 2: k-Means Clustering

- A **greedy** algorithm
- Disadvantages:
  - Initial means are randomly selected which can cause suboptimal partitions

    Possible Solution: Try a number of different starting points
  - Depends on the value of k

Demo 2

Coming Up

- **Assignment 9 is Optional**:
  - Will replace your second lowest score if you submit
  - Due at 5:30pm on December 16th, 2020
- **Last Lecture**: Real-world applications of machine learning (*December 16th, 2020*)
- **Final Project:** Due on December 16th, 2020

CDS Education

We explore, learn, and educate big minds.