Lecture 9: Unsupervised Learning and Clustering

INFO 1998: Introduction to Machine Learning

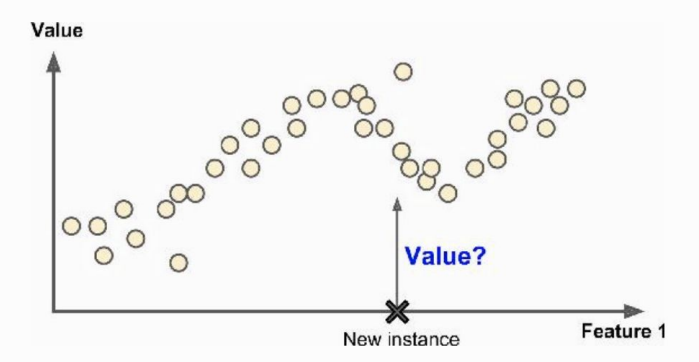
CDS Education

Agenda

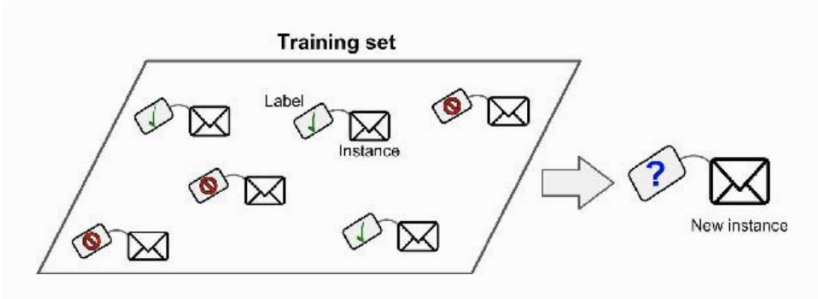
1. Review of Supervised Learning
2. Unsupervised Learning
3. Clustering Algorithms

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

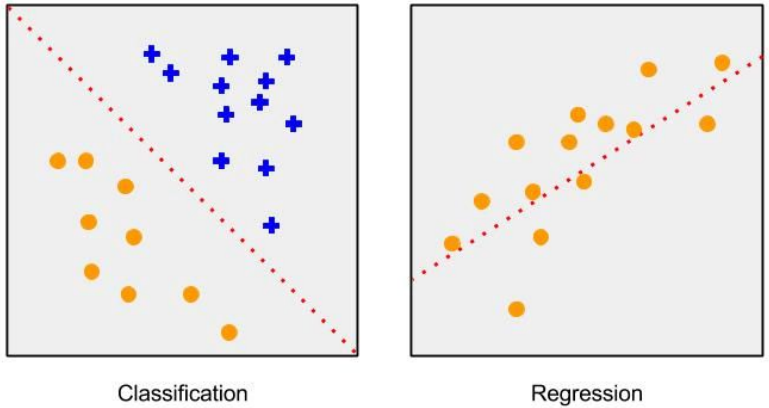


Example 2: Classifier
Predicts label

Recap: Supervised Learning

Supervised learning algorithms we have covered so far:

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



Classification

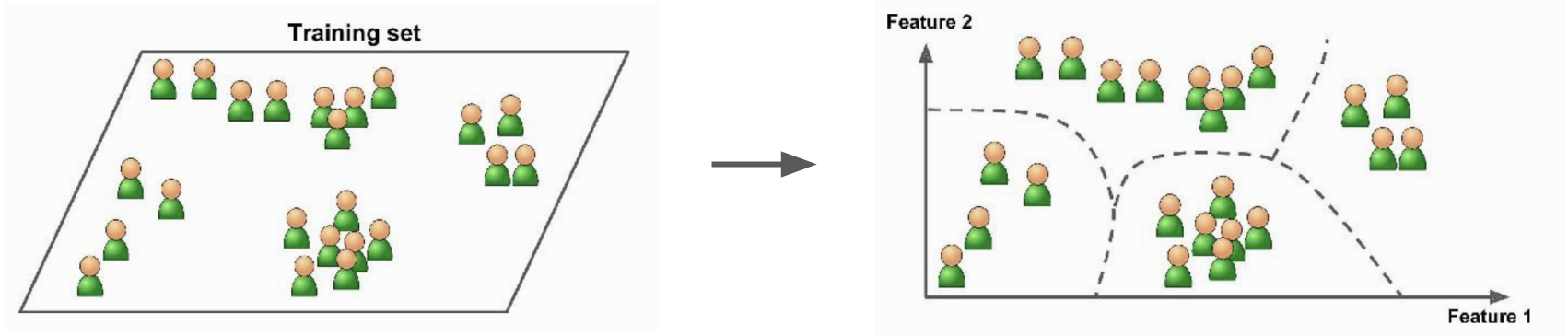
Regression

Which of these are classifiers? Which are regressors?

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:

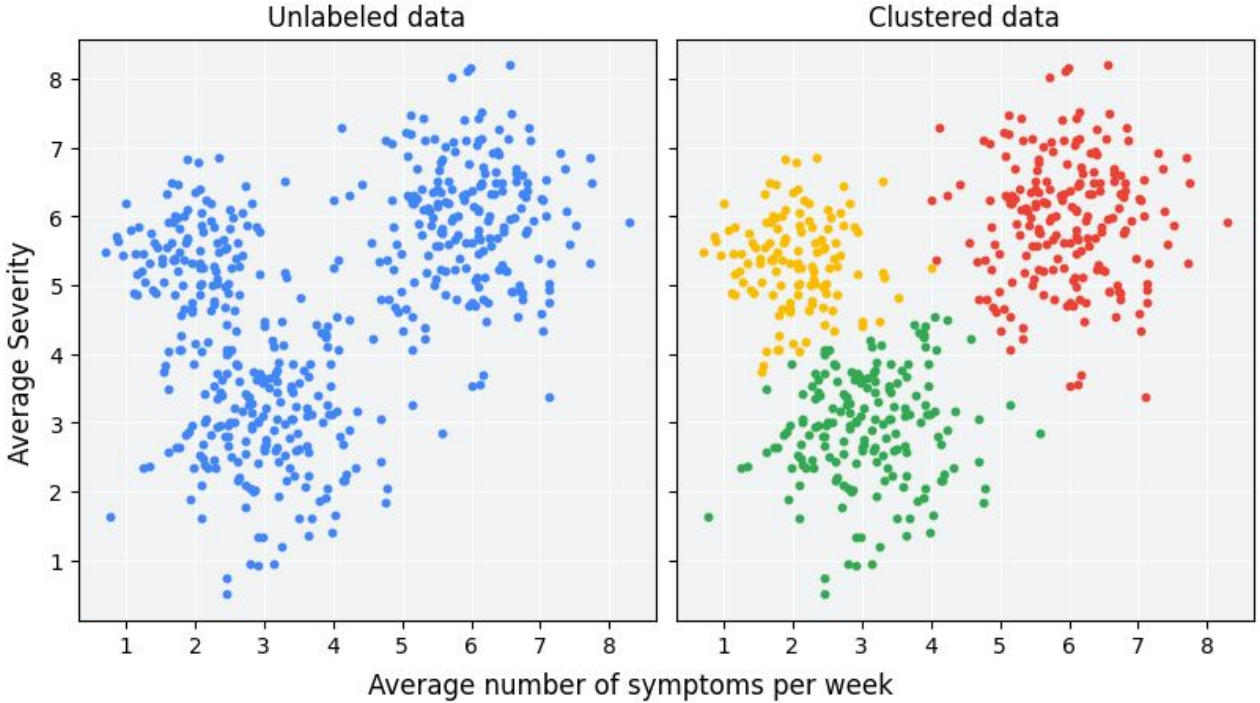
1. **Clustering**
   k-Means, Hierarchical Cluster Analysis (HCA), Gaussian Mixture Models (GMMs), etc.
2. **Dimensionality Reduction**
   Principal Component Analysis (PCA), Locally Linear Embedding (LLE)
3. **Association Rule Learning**
   Apriori, Eclat, Market Basket Analysis

… *More*

Cluster Analysis

- **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

Cluster Analysis



Clustering Application

Recommender Systems

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

A Bunch of Cool Logos

Running Example: Recommender Systems

Use 1: Collaborative Filtering

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

Pros

If cluster behavior is clear, can yield good insights

Cons

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	Limited to existing data about content
Scalable	Difficult to suggest for new users

How do we actually perform this “cluster analysis”?

Defining ‘Similarity’

- Remember from K Nearest Neighbors Discussion
- 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

Popular Clustering Algorithms

Hierarchical Cluster Analysis (HCA)

k-Means Clustering

Gaussian Mixture Models (GMMs)

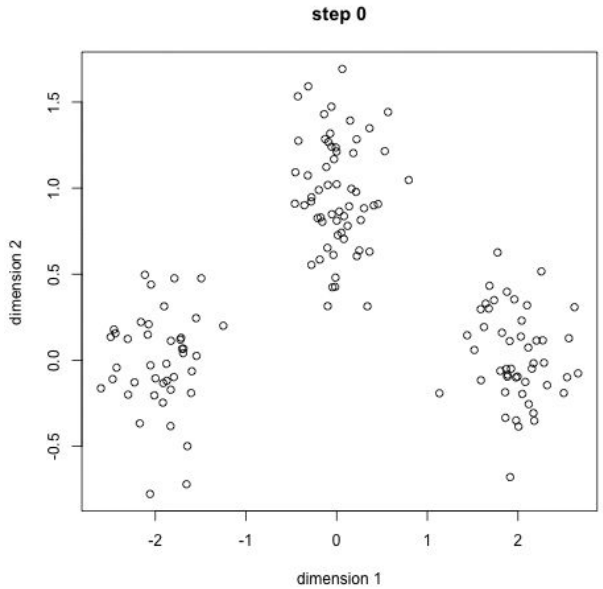
Algorithm 1: k-Means Clustering

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: **https://www.naftaliharris.com/blog/visualizing-k-means-clustering/**

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
  - Major assumptions about distribution of data!

Demo 2

Popular Clustering Algorithms

Hierarchical Cluster Analysis (HCA)

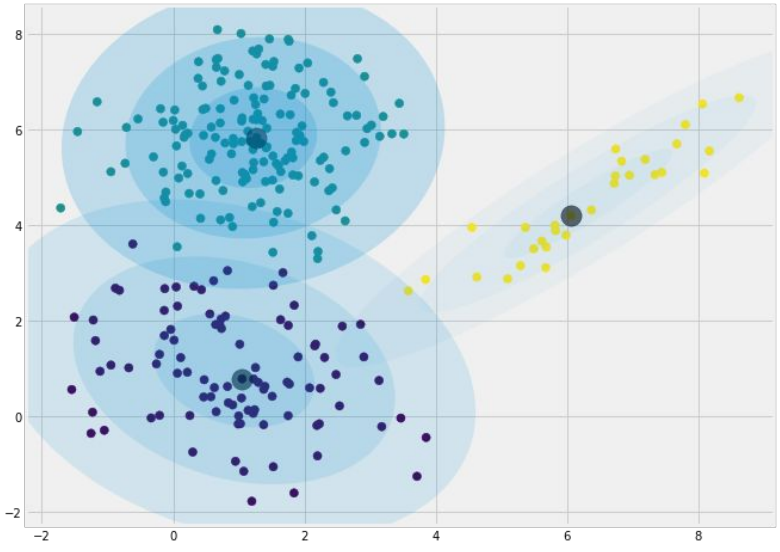
k-Means Clustering

Gaussian Mixture Models (GMMs)

Algorithm 2: Gaussian Mixture Models

Input parameter: *k*

- Starts with *k* Gaussian distributions
- Train on data to find the appropriate means and covariances for each cluster
- Compute probability of each test point lying inside each distribution and predict the one with the highest probability.



Demo 3

Coming Up

- **Assignment 8** due **tonight** at midnight!
- Assignment 9 due next week Wednesday
  - *Last coding assignment!*
- Next Week: **Data Science in the Real World**
- **Deep Learning Workshop** (coming up!)