CDS Education

Introduction to Machine Learning for Python

Unsupervised Learning

Recap: Supervised Learning

Supervised learning uses **regressors** and **classifiers**.

- We train a **learner** to predict a **dependent variable**, given independent variables
- There is a definitive “answer” to learn from

Use of “unlabeled” data

One common case of **unlabeled data**: data with missing values.

A column could have a lot of missing values that need to be approximated.

We can estimate true values by **imputing** the missing data.

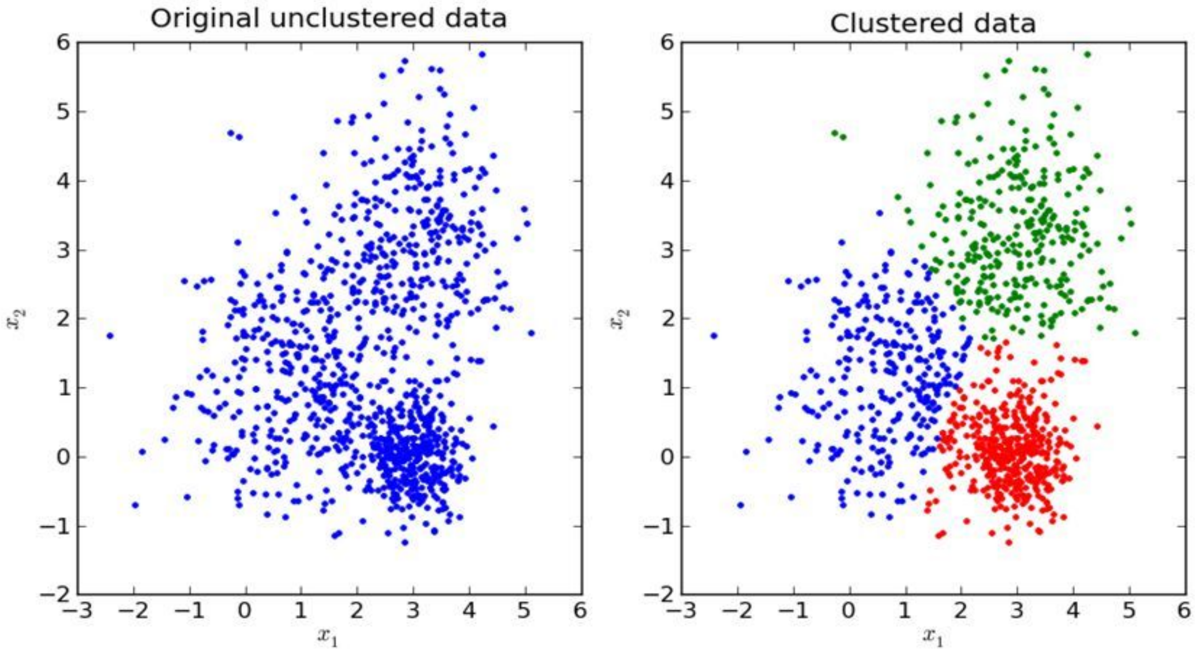
Source

Cluster Analysis

Clusters (close-knit groups of data in space) are **latent variables**.

Understanding clusters can:

- Yield underlying trends in data
- Supply useful parameters for predictive analysis
- Challenge the boundaries of predefined classes in variables


Original unclustered data
Clustered data
x_1
x_2
−3
−2
−1
0
1
2
3
4
5
6

Source

Recommendation Systems

Recommendations are the heart of many businesses

Google

YouTube

amazon

NETFLIX

Technique 1: Collaborative Filtering

Collaborative filtering: “people similar to you also liked X.”

- Example: Using other users’ ratings to suggest content.

If cluster behavior is clear, can yield good insights

Computationally expensive

Can lead to dominance of certain groups in predictions

Collaborative Filtering: Movie Rating Example

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

Technique 2: Content Filtering

Content filtering: “content similar to what you’re viewing”

- Example: Using other movies watched by user to recommend an unwatched movie.

Recommendations made by learner are intuitive

Scalable

Limited in scope and applicability

Content Filtering: Movie Suggestion Example

	Harry Potter	Indiana Jones	Back to the Future	Mean Girls	Black Swan
Genre	Adventure, Family, Fantasy	Action, Adventure	Adventure, Comedy, Sci-Fi	Comedy	Drama, Thriller
Won an Oscar?	no	yes	yes	no	yes
Female Lead?	no	no	no	yes	yes
Year Released	2001	1981	1985	2004	2010
Rating	PG	PG	PG	PG-13	R

Case Study: Cambridge Analytica

Cambridge Analytica

Goal: Use Facebook data to build psychological profiles that could then be used in targeted advertising

- Score people on a personality test measuring openness, conscientiousness, extroversion, agreeableness and neuroticism
- Correlate traits to a type of advertisement

Popular Clustering Algorithms

Hierarchical Clustering

k-means Clustering

Gaussian Mixture Model

Source

Proximity/Similarity

- Euclidean distance:

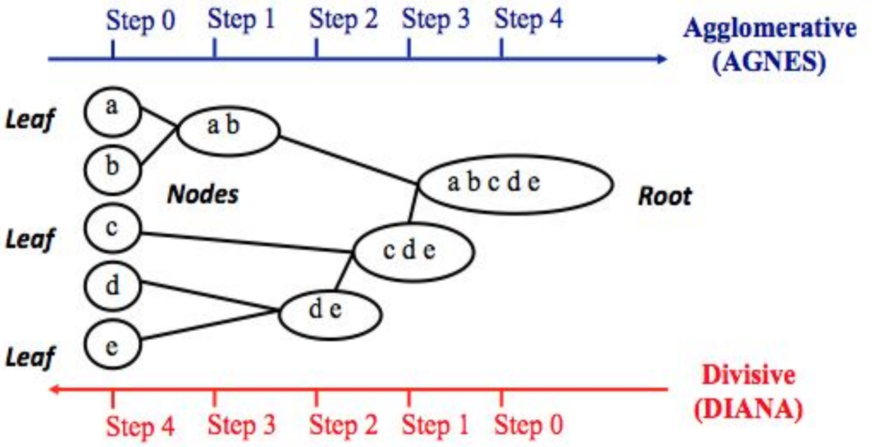
$$E(x, y) = \sqrt{\sum_{i=0}^{n} (x_i - y_i)^2}$$

- Other methods:
  - Squared euclidean distance, manhattan distance

Hierarchical Clustering

Two algorithms:

- Agglomerative clustering
  - Creates a tree of **increasingly** large clusters
- Divisive hierarchical clustering
  - Creates a tree of **decreasingly** large clusters

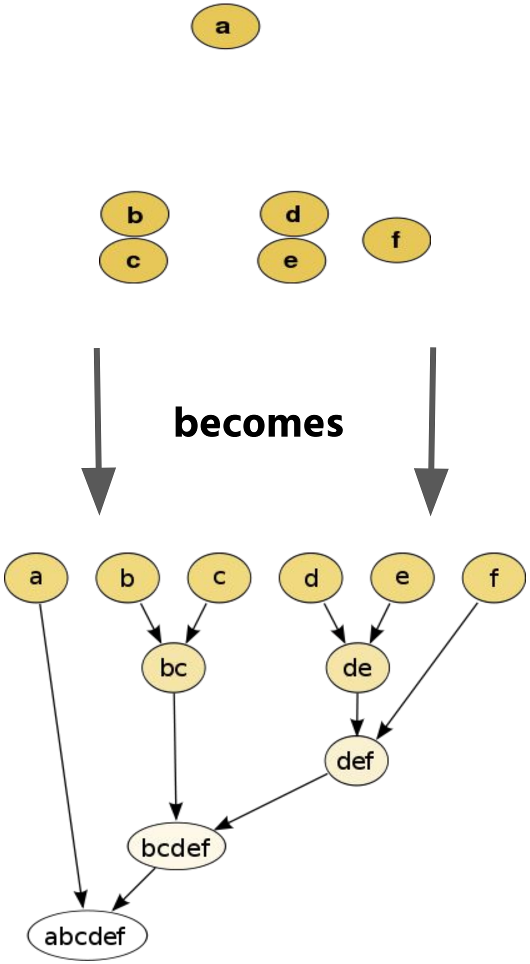


Hierarchical Clustering

Agglomerative Clustering Algorithm:

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

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



becomes

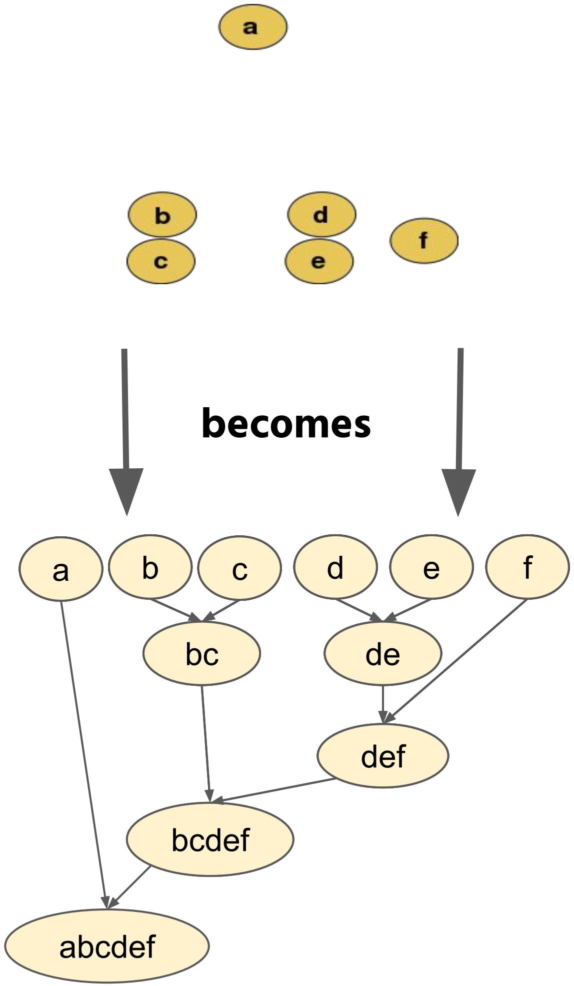
Source

Hierarchical Clustering

Agglomerative Clustering Algorithm:

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

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

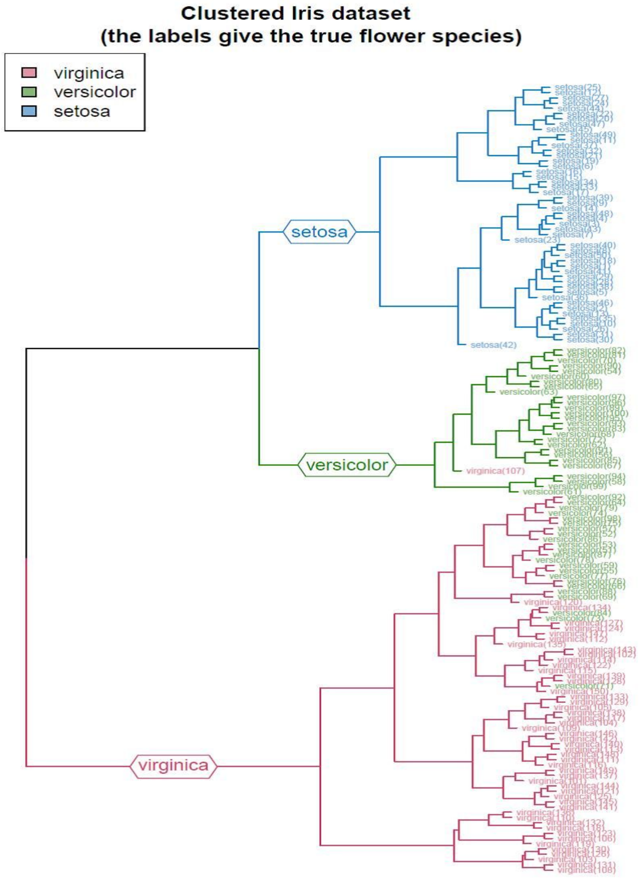


Source

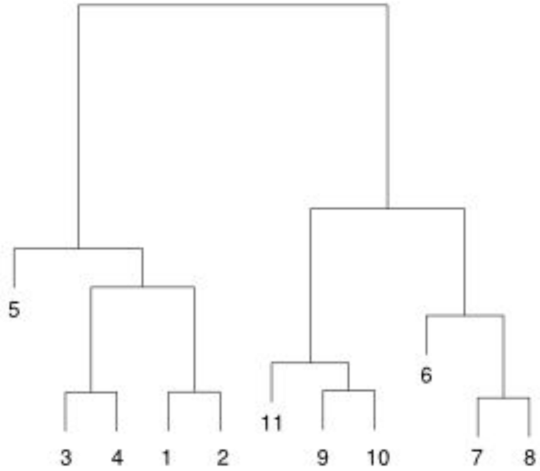
Dendrograms

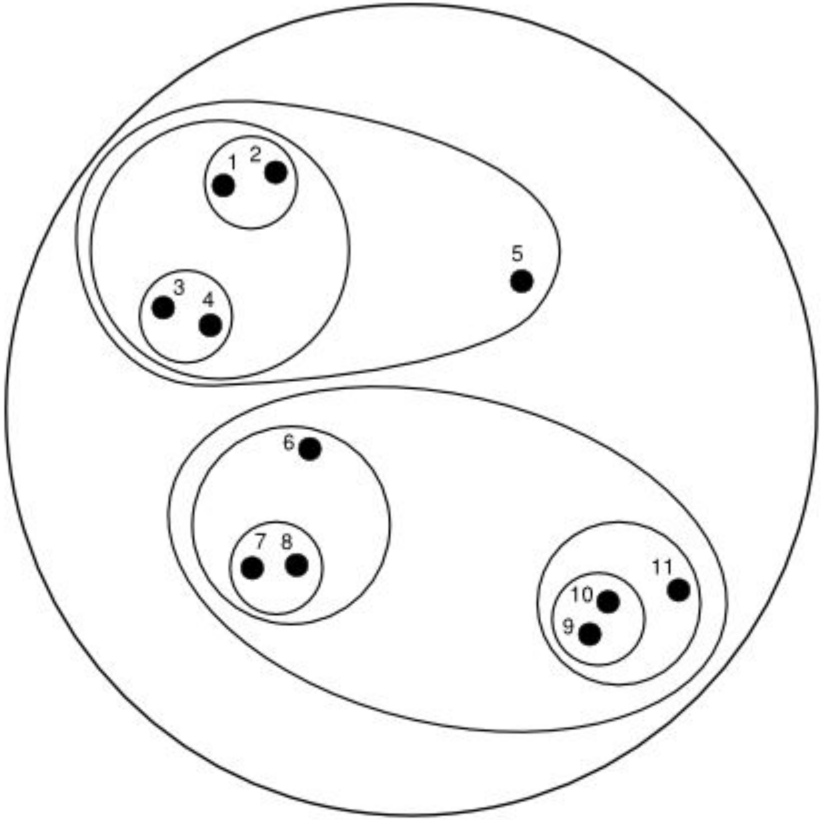
Visualizes hierarchical clustering

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



Source





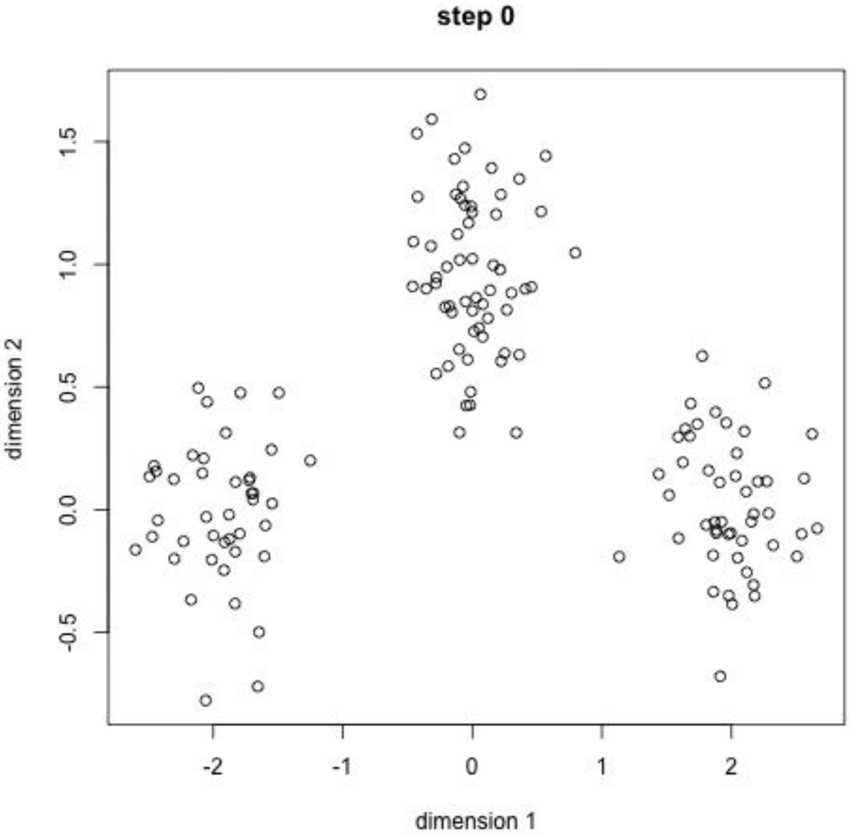
Source

Demo!

K-means Clustering

Simplest clustering algorithm. Input parameter: *k*

1. Starts with *k* random centroids
2. Cluster points using "centroids"
3. Take average of clustered points
4. Use as new centroids
5. Repeat until convergence



Source

K-means Clustering

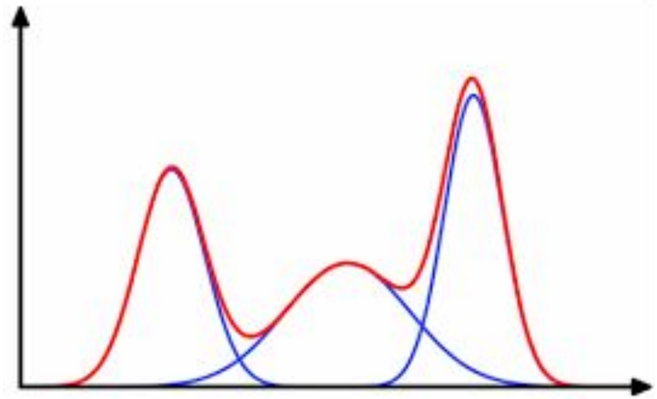
Greedy algorithm

- Partitions the n samples into k clusters
- Minimizes the sum of the squared distances to the cluster centers
- Weaknesses:
  - Initial means are randomly selected which can cause suboptimal partitions
    - Try a number of different starting points
  - Depend on the value of k

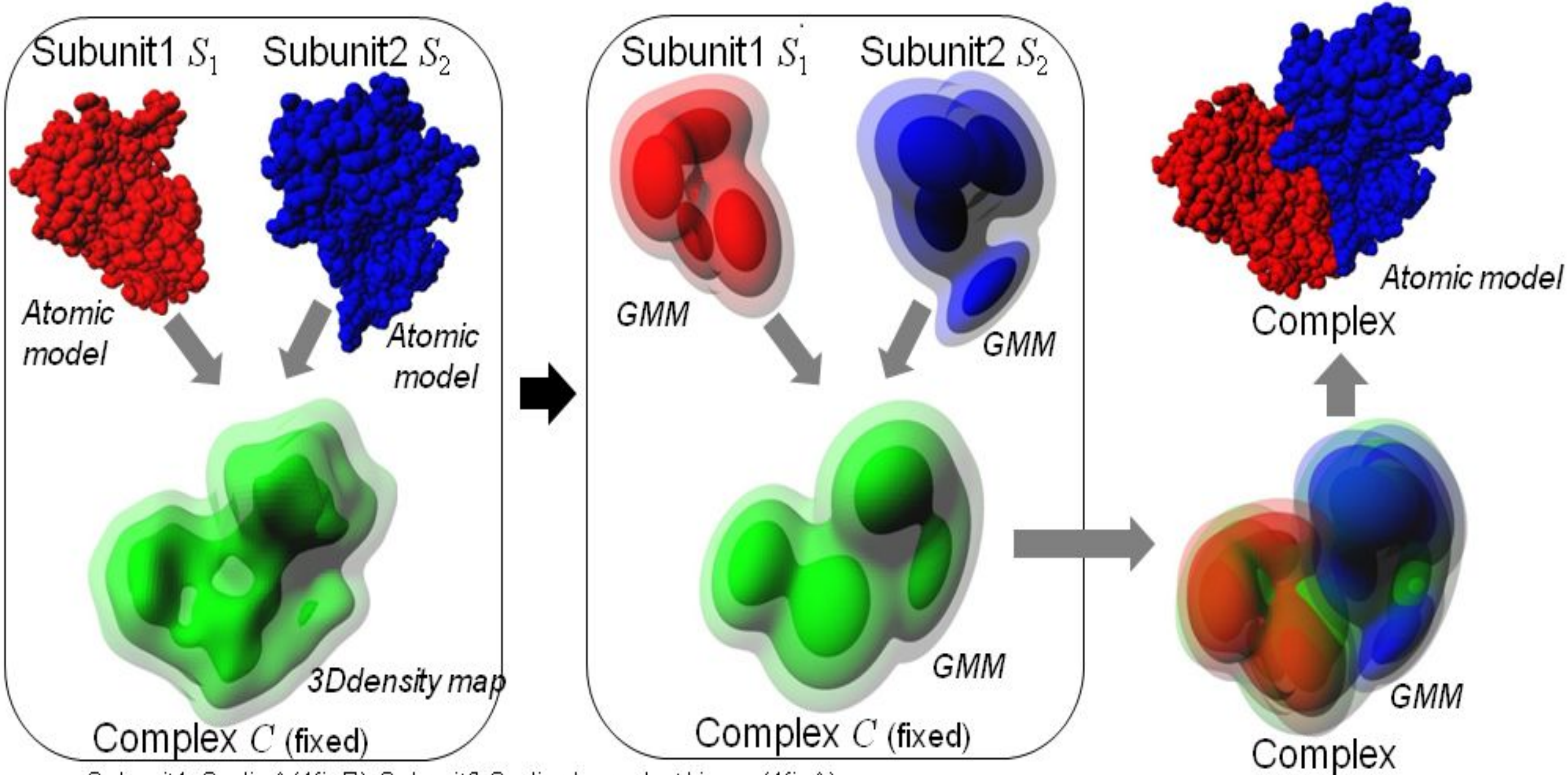
Gaussian Mixture Model (GMM)

Assumptions that the data is a <u>mixture</u> of clusters.

- Clusters may overlap
- Gaussian mixture models assume that each cluster is **normally distributed**



GMM may more accurately describe reality since boundaries are usually not clear cut.



Subunit1: Cyclin A (1finB) ,Subunit2:Cyclin-dependent kinase(1finA)

Maximum Likelihood Estimator (MLE)

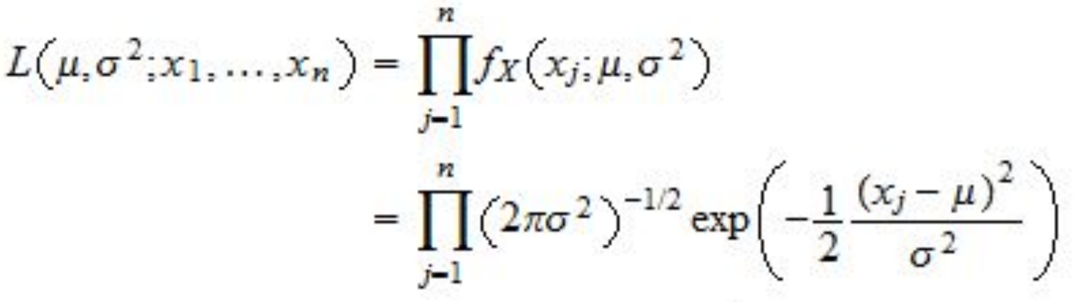
Application of an unsupervised learning problem

- How to fit the best model to a set of data
- Can either assume a certain distribution or estimate without knowing the distribution

Maximum Likelihood Estimator (MLE)

Given observations, how likely is a certain set of parameters?

- Assumptions must be made on the probability distribution
- Obtain a function of maximum likelihood
- Obtain local maxima, minima using calculus



$$L(\mu,\sigma^2;x_1,\ldots,x_n)=\prod_{j=1}^{n} f_X(x_j;\mu,\sigma^2)$$

$$=\prod_{j=1}^{n}(2\pi\sigma^2)^{-1/2}\exp\left(-\frac{1}{2}\frac{(x_j-\mu)^2}{\sigma^2}\right)$$

Source

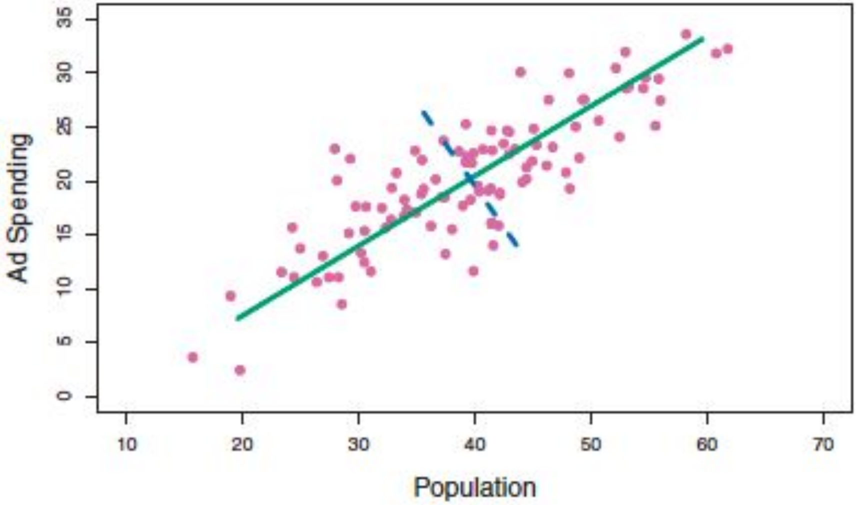
Expectation-Maximization Algorithm

A general unsupervised learning method for MLEs

1. Pick random values for parameters
2. Make predictions based on the parameters
3. Take these predictions as true, solve for most likely parameters.

   Repeat step 2 with these parameters
4. Repeat until convergence

Principal Component Analysis (PCA)

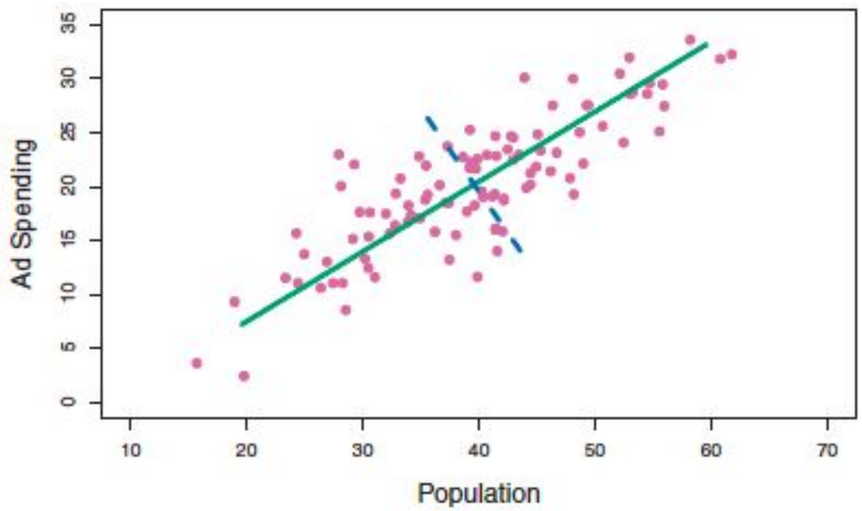


Want to understand the “direction” that our data goes in without storing whole data set.

1. Find the direction along which the data has the largest variance (projections of all data points are the largest). Called the **first principal component** (green line).

Hastie, Trevor, et al. “An Introduction to Statistical Learning.”

Principal Components



2. Find the direction which is orthogonal to the first principal component and has the largest variance (projections of points are largest).

This is the **second principal component** (blue dotted line).

Garath, James, et al. "An Introduction to Statistical Learning in R."

Principal Components

Generally, *n* dimensional data can have *n* principal components.

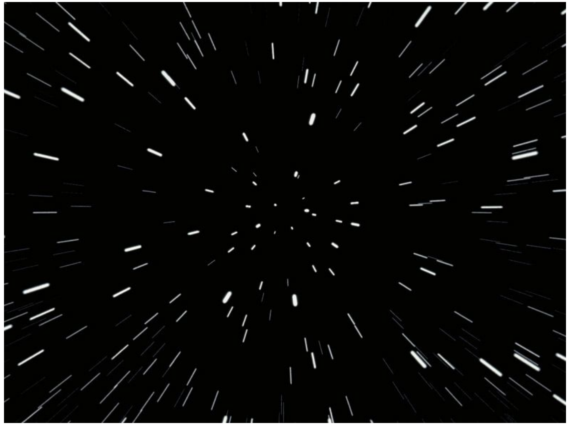
Principal component analysis - process of constructing these components (orthogonal directions of largest variance)

Why?

PCA is used for:

Exploratory data analysis for unsupervised learning (what are the general trends?)

Obtaining a low-dimensional **approximation** for high dimensional data (thousands of features)



Source

Coming Up

Your problem set: Project part C

Next week: Ensemble Learning

See you then!

CDS

Cornell Data Science