Introduction to Statistical/Machine Learning

Unsupervised Machine Learning

Anthony Strittmatter

Literature

- James, Witten, Hastie, and Tibshirani (2013): ”An Introduction to Statistical Learning”, Springer, Chapters 6.3.1, 10, download.

- Hastie, Tibshirani, and Friedman (2009): ”Elements of Statistical Learning”, 2nd ed., Springer, Chapters 14.2, 14.5, download.

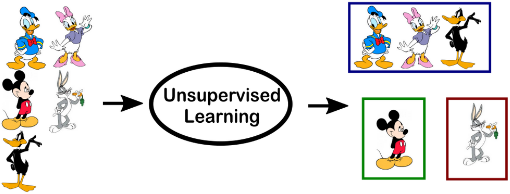
Supervised vs. Unsupervised Machine Learning

Supervised Machine Learning:

- We observe data on Y and X and want to learn the mapping $\widehat{Y} = \widehat{f}(X)$
- Classification when Y is discrete, regression when Y is continuous

Unsupervised Machine Learning:

- We observe only data on X and want to learn something about the data structure



Unsupervised Machine Learning

- Explorative data analysis.
- Discovering subgroups among observations or variables.
- No easy way to assess model accuracy.
- Visualization of X data.

$\Rightarrow$ We discuss **Principal Component Analysis (PCA)** and **K-Means Clustering**.

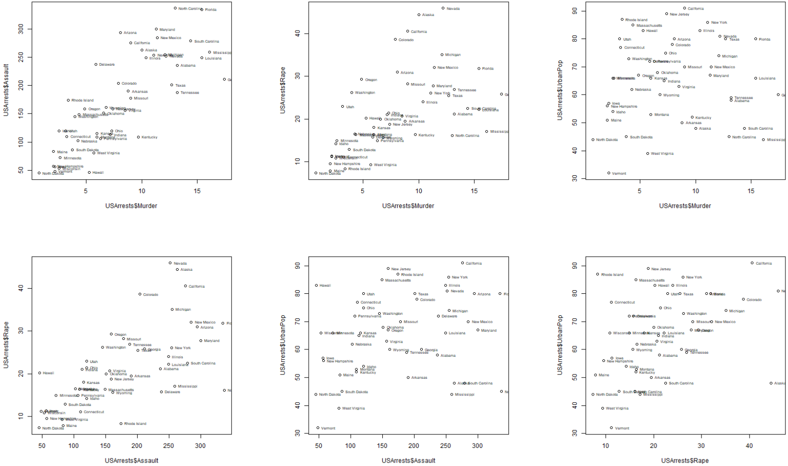
Violent Crime Rate by US States

Variables:

- Murder arrests (per 100,000)
- Assault arrests (per 100,000)
- Percent urban population (UrbanPop)
- Rape arrests (per 100,000)

	Murder	Assault	UrbanPop	Rape
Alabama	13.2	236	58	21.2
Alaska	10	263	48	44.5
Arizona	8.1	294	80	31
Arkansas	8.8	190	50	19.5
California	9	276	91	40.6
Colorado	7.9	204	78	38.7
⋮	⋮	⋮	⋮	⋮

Scatterplots



→ PCA finds low dimensional representation of data that captures as much information as possible.

Principal Components

- We observe the features X_1, X_2, ..., X_p.
- Principal components are normalized linear combinations of the features

$$\begin{aligned} Z_1 =& \phi_{11}X_1 + \phi_{21}X_2 + ... + \phi_{p1}X_p, \\ Z_2 =& \phi_{12}X_1 + \phi_{22}X_2 + ... + \phi_{p2}X_p, \\ &\vdots \\ Z_p =& \phi_{1p}X_1 + \phi_{2p}X_2 + ... + \phi_{pp}X_p, \end{aligned}$$

  that maximize the variance of Z_1, Z_2, ... Z_p.
- The factor loadings of the principal component k are $\phi_k = \phi_{1k}, \phi_{2k}, ...\phi_{pk}$.
- Normalized means $\sum_{j=1}^{p} \phi_{jk}^2 = 1$ for all $k = 1, ..., p$.

Objective Function

- First Principal Component:

$$max_{\phi_{11},\dots,\phi_{p1}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{p} \phi_{j1} x_{ij} \right)^2 \right\} \text{ s.t. } \sum_{j=1}^{p} \phi_{j1}^2 = 1.$$

- Second Principal Component:

$$max_{\phi_{12},\dots,\phi_{p2}} \left\{ \frac{1}{N} \sum_{i=1}^{N} \left(\sum_{j=1}^{p} \phi_{j2} x_{ij} \right)^2 \right\} \text{ s.t. } \sum_{j=1}^{p} \phi_{j2}^2 = 1$$

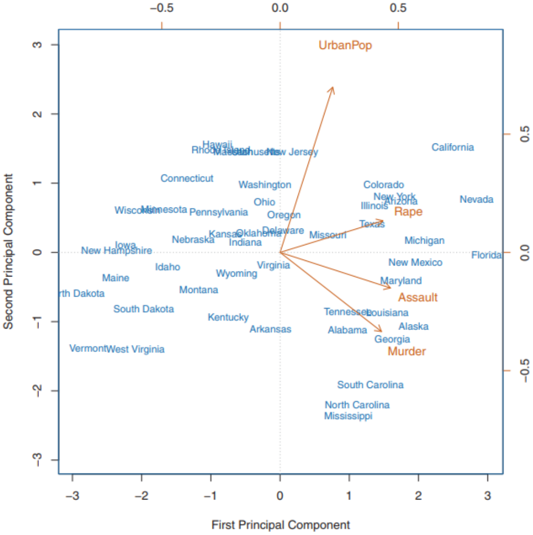
  and ϕ_2 is orthogonal to ϕ_1.

- etc.

Principal Component Loading Vectors

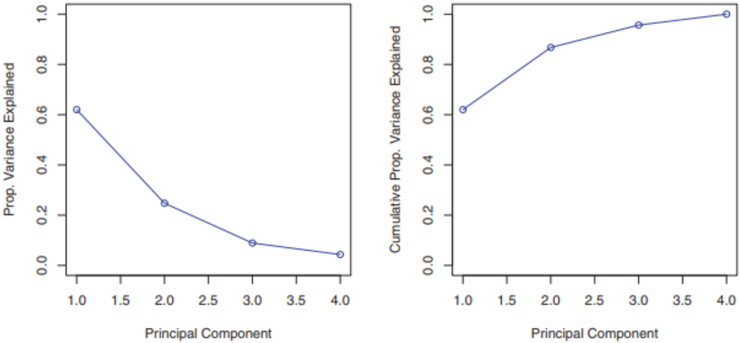
	Principal Component 1	Principal Component 2
	ϕ_1	ϕ_2
Murder	0.536	-0.418
Assault	0.583	-0.188
Urban Population	0.278	0.873
Rape	0.543	0.167

Visualization of Principal Components



Source: James, Witten, Hastie, Tibshirani (2013)

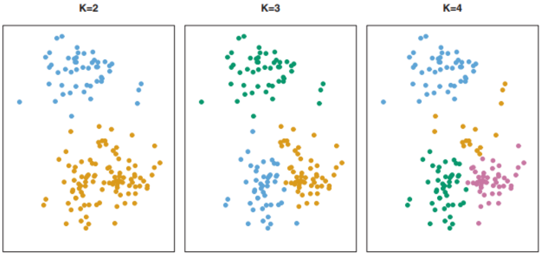
Proportion Variance Explained



Source: James, Witten, Hastie, Tibshirani (2013)

Difference between PCA and Clustering

- Principal Component Analysis (PCA) looks to find a low-dimensional representation of the observations that explain a good fraction of the variance.
- Clustering looks to find homogeneous subgroups among the observations.



Source: James, Witten, Hastie, Tibshirani (2013)

Objective Function K-means Clustering

- **Squared Euclidean distance:**

$$W(C_k) = \frac{1}{|C_k|} \sum_{i,i' \in C_k} \sum_{j=1}^{p} (x_{ij} - x_{i'j})^2$$

with $|C_k|$ being the number of observations in the kth cluster.

- **Optimization problem:**

$$\min_{C_1,\ldots,C_K} \left\{ \sum_{k=1}^{K} W(C_k) \right\}$$

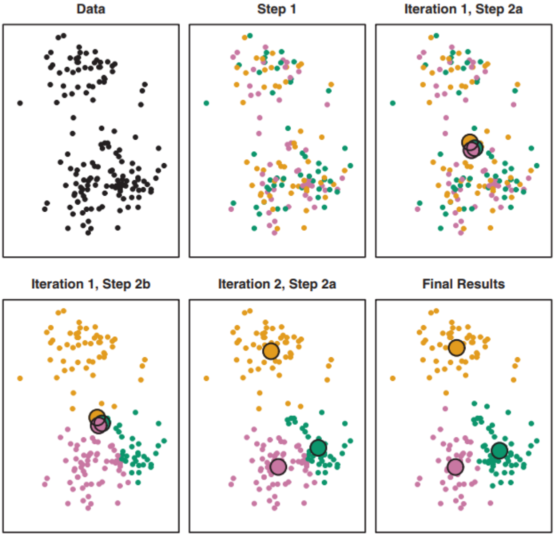
→ Minimize the within cluster squared Euclidean distance.

Optimization Algorithm K-Means Clustering

Algorithm

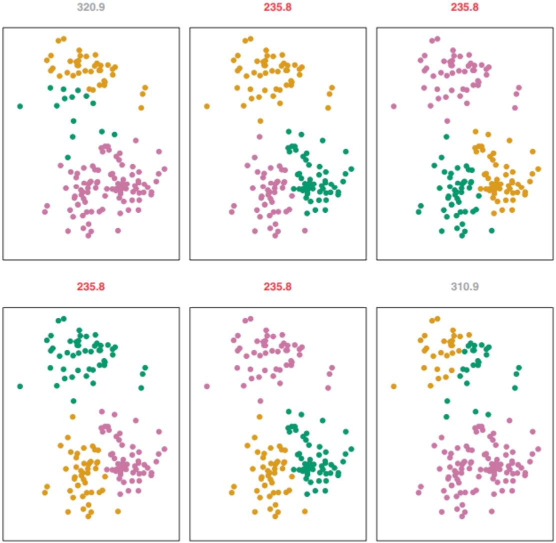
1. Randomly assign a number, from 1 to K, to each of the observations. These serve as initial cluster assignments for the observations.
2. Iterate until the cluster assignments stop changing:
    2.1 For each of the K clusters, compute the cluster centroid. The kth cluster centroid is the vector of the p feature means for the observations in the kth cluster.
    2.2 Assign each observation to the cluster whose centroid is closest (where closest is defined by using the squared Euclidean distance)

Graphical Illustration of Optimization Algorithm



Source: James, Witten, Hastie, Tibshirani (2013)

Initialisation of the Algorithm (Step 1)



Source: James, Witten, Hastie, Tibshirani (2013)

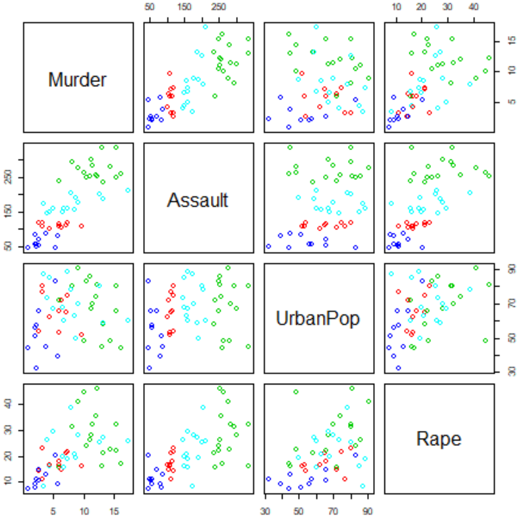
4-Means Clustering for Crime Data

- **Cluster 1:** low crime
  Connecticut, Idaho, Indiana, Kansas, Kentucky, Montana, Nebraska, Ohio, Pennsylvania, Utah
- **Cluster 2:** very high crime
  Alabama, Alaska, Arizona, California, Delaware, Florida, Illinois, Louisiana, Maryland, Michigan, Mississippi, Nevada, New Mexico, New York, North Carolina, South Carolina
- **Cluster 3:** low pop, low crime
  Hawaii, Iowa, Maine, Minnesota, New Hampshire, North Dakota, South Dakota, Vermont, West Virginia, Wisconsin
- **Cluster 4:** high crime
  Arkansas, Colorado, Georgia, Massachusetts, Missouri, New Jersey, Oklahoma, Oregon, Rhode Island, Tennessee, Texas, Virginia, Washington, Wyoming

Descriptives by Cluster

	Mean			
	Cluster 1	Cluster 2	Cluster 3	Cluster 4
Murder	5.59	11.81	2.95	8.214
Assault	112.4	272.6	62.7	173.3
UrbanPop	65.6	68.31	53.9	70.64
Rape	17.27	28.38	11.51	22.84

Scatterplot of Clusters


Murder
Assault
UrbanPop
Rape