Master's Capstone Project:

Prediction Strength of K-Means Clustering

Zhongkai Wang

Department of Biostatistics, University of Florida

zkwang@ufl.edu

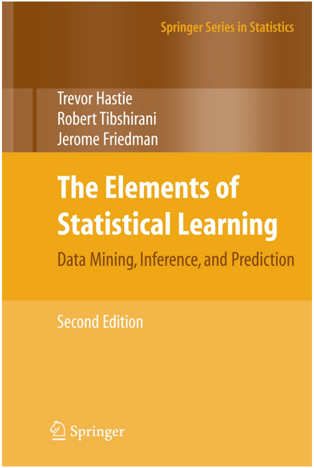
March 22, 2016

Overview

Alpha-Go by Google

From the Statistics side

Statistical learning is an applied field of statistics that includes vast amount of tools that emphasizing on building models, implementing algorithms, assessing uncertainties by learning from complex datasets. These tools can be categorized as supervised and unsupervised learning algorithms.



Supervised vs. Unsupervised Learning

Supervised Learning

Supervised learning refers to building models based on input (training) and predict/estimate outcome (testing). For example, regression and classification.

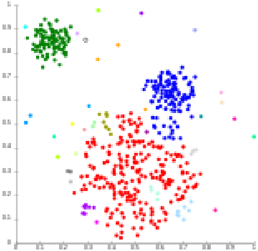
Unsupervised Learning

Unsupervised learning only has input and no (supervised) output. We usually use it to study the association and underlying patterns of input data. It includes cluster analysis, principal component analysis et al.

Overview

Cluster Analysis

Clustering is the task of grouping a set of objects in such a way that objects in the same group (called a cluster) are more similar (in some sense or another) to each other than to those in other groups (clusters). - Wikipedia



Cluster Analysis

Computer Science: grouping objects based on a set of features.
Statistics: grouping observations based on a set of variables.

For example,
Students: GPA, GRE, Age,...
Athletics: Speed, Hight, Weight,...
Computers: Model, Year, CPU,...

The k-means clustering

Squared Euclidean distance:

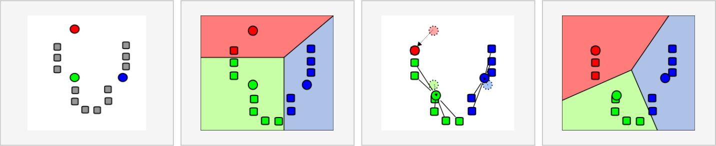
$$d(x_i, x_{i'}) = \sum_{j-1}^{p} (x_{ij} - x_{i'j})^2 = ||x_i - x_{i'}||^2 \tag{1}$$

The k-means clustering's objective is to find:

$$\arg\min_{S} \sum_{i=1}^{k} \sum_{x \in S_i} ||\mathbf{x} - \mu_{\mathbf{i}}||^2 \tag{2}$$

Where $\mu_{\mathbf{i}}$ is the mean of points in S_i.

The k-means clustering



1. *k* initial "means" (in this case *k*=3) are randomly generated within the data domain (shown in color).

2. *k* clusters are created by associating every observation with the nearest mean. The partitions here represent the Voronoi diagram generated by the means.

3. The centroid of each of the *k* clusters becomes the new mean.

4. Steps 2 and 3 are repeated until convergence has been reached.

(From Wikipedia)

The k-means clustering

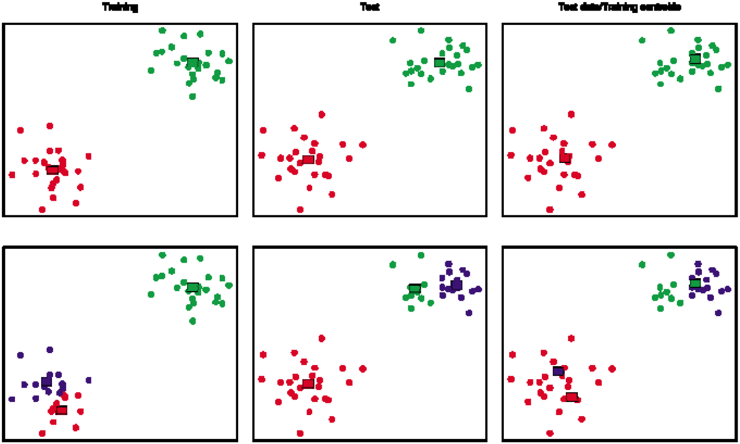
How do we know how many clusters?

Prediction Strength

The prediction strength measures the proportion of observation pairs in each of the testing cluster that are also assigned to the same cluster using the training centroids.

- For each i from 1 up to maximum number of clusters, divide the dataset into two groups, a training set and a testing set.
- Run a k-means algorithm on each set to find i clusters.
- For each testing cluster, count the proportion of pairs of points in that cluster that would remain in the same cluster, if each were assigned to its closest training cluster mean.
- The minimum over these proportions is the prediction strength for using i clusters.

Prediction Strength



Prediction Strength

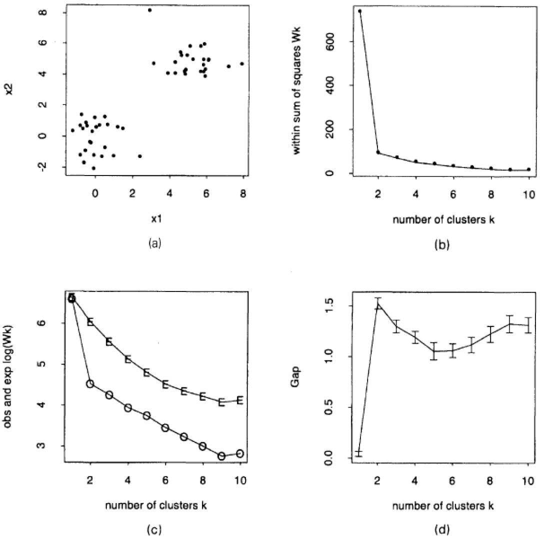
$$ps(k) = \min_{1 \le j \le k} \frac{1}{n_{kj}(n_{kj} - 1)} \sum_{i \neq i' \in A_{kj}} D[C(X_{tr}, k), X_{te}]_{ii'} \quad (3)$$

Training data $X_{tr} = x_{ij}$, a k-means clustering operation $C(x_{tr}, k)$.

A $n \times n$ square matrix where each element, $D[C(X_{tr}, k), X_{te}]_{ii'}$, is an indicator with 1 if observation i abd $i^{'}$ falls into the same cluster when applying the training's centroids to the testing data, otherwise is 0.

Threshold: Pick an arbitrary cut-off (usually is 0.8).

Gap Statistic



Gap Statistic

The within-cluster sum of squares W_k as (b) for each possible choice of cluster numbers $k = 1 - 10$:

$$W_k = \sum_{r=1}^{k} \frac{1}{2n_r} D_r = \sum_{r=1}^{k} \sum_{i,i' \in C_r} \frac{1}{2n_r} d_{ii'} \tag{4}$$

Where $d_{ii'}$ is an distance function, and might be a squared Euclidean distance for simplify.

Meanwhile, we introduce an expected log within-cluster sum of squares $\hat{E}^*\{log(W_k)\}$, shown as higher curve in (c), using the reference dataset. A simply way to generate the reference dataset is assuming a equal number of data follow a uniform distribution within the range of each original p dimensions.

Gap Statistic

Gap Statistic

$$GAP_n(k) = E_n^*[\log(W_k)] - \log(W_k) \tag{5}$$

Threshold: choose the smallest k that maximum the Gap and satisfy $Gap(k) \geq Gap(k+1) - sd_{k+1}$.

Overview

Simulation Study

Table 1: Simulation Setup Summary

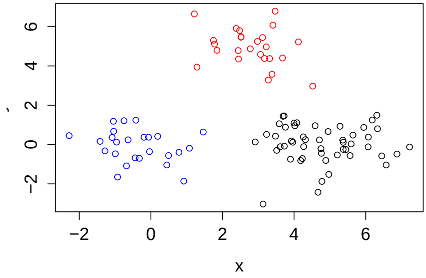
Sim. times	Name	Obs No.	Dimensions	True cluster No.
100	A	100	2	3
100	B	100	2	4
100	C	100	3	4
100	D	100	3	2
100	E	200	10	1

Datasets

- A. *Three cluster in two dimensions*: the clusters are standard normal variables with mean (0, 0), (3, 5), (5, 0).
- B. *Four cluster in two dimensions, not well separated*: each cluster has 25 standard normal variables with mean (0, 0), (0, 2.5), (2.5, 0), (2.5, 2.5).
- C. *For cluster in three dimensions, well separated*: each cluster has 25 standard normal variables with mean (-10, 0, -5), (3, 0, -5), (5, 0, 0), (-10, 0, 0).
- D. *Two cluster in two dimensions*: the two elongated clusters are stretching along the main deagonal of a cube.
- E. *One cluster in 10 dimensions*: the clusters are uniformly distributed in 10 dimensions).

Data A

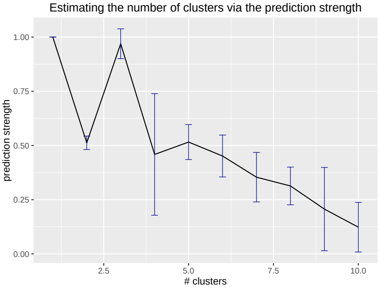
3 cluster dataset with well separated 100 data points in 2 dimensions. The points are generated as standard normal variables with (25, 25, 50) observations, centered at (0, 0), (3, 5), (5, 0).

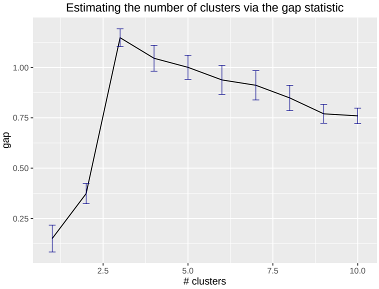


Data A Results

Table 2: Data A Simulation Result

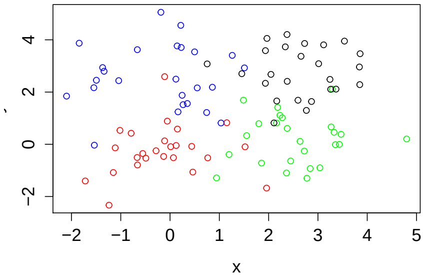
	1	2	3	4	5	6	7	8	9	10
PS	1		94	5						
GAP			100							





Data B

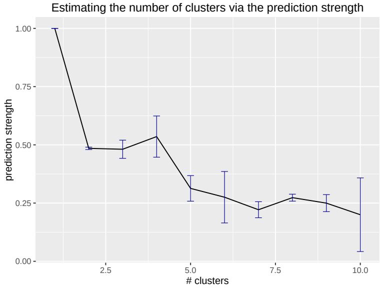
4 cluster in 2 dimensions that are mixed. Each cluster has 25 standard normal variables with mean (0, 0), (0, 2.5), (2.5, 0), (2.5, 2.5).

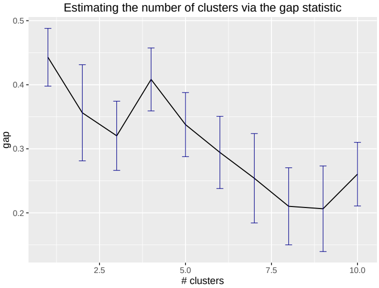


Data B Results

Table 3: Data B Simulation Result

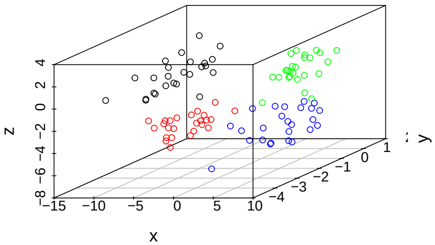
	1	2	3	4	5	6	7	8	9	10
PS	90	7	2	1						
GAP	88	1	8	1	2					





Data C

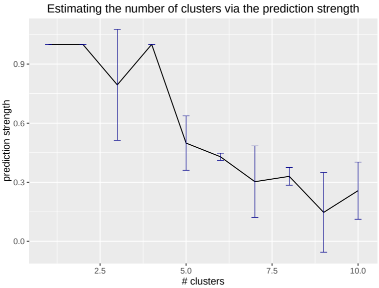
4 cluster in 3 dimensions that are well separated.

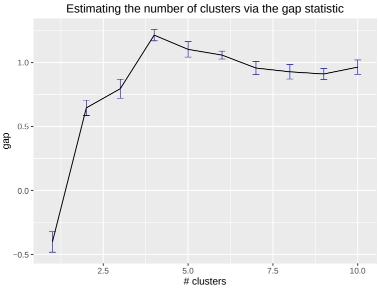


Data C Results

Table 4: Data C Simulation Result

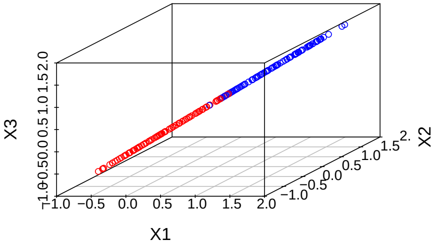
	1	2	3	4	5	6	7	8	9	10
PS		1		99						
GAP				100						





Data D

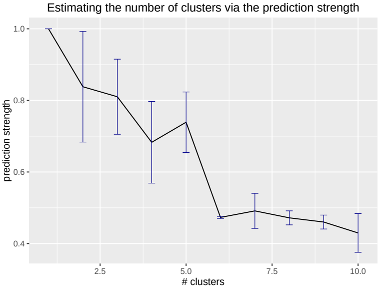
2 cluster in 3 dimensions that spread along the diagonal of a cube.

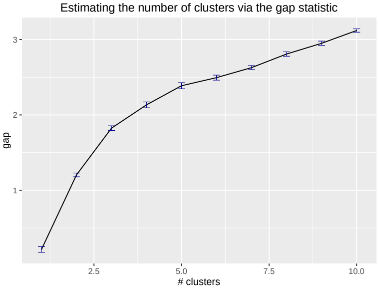


Data D Results

Table 5: Data D Simulation Result

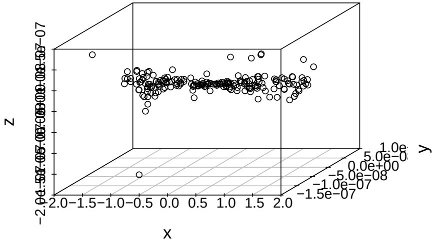
	1	2	3	4	5	6	7	8	9	10
PS		66	2	28	4					
GAP									1	99





Data E

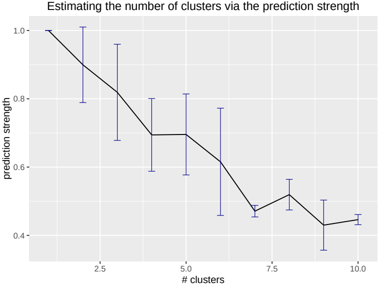
Single cluster in 10 dimensions that are mixed. Plot using a classical multi-dimensional scaling to reduce the data to three dimensions.

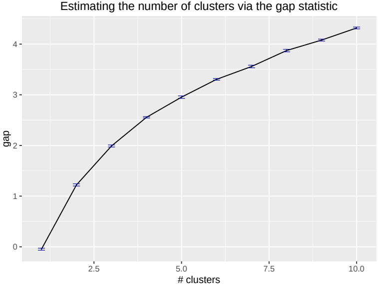


Data E Results

Table 6: Data E Simulation Result

	1	2	3	4	5	6	7	8	9	10
PS	8	52	28	7	3	2				
GAP										100





Overview

Data Description

Study of Parkinson Disease. Deleting obserbations with NAs there are 122 patients. The variable "group" is a label of whether the patient is classified as Parkinson patient or not. We want to use the following 9 variables to perform k-means clustering, prediction strength and gap statistic as described before, to compare the cluster result with the "roup" label and see if clustering analysis may provide a guideline of classifying patients.

Table 7: First 3 overvations of cluster variable

Group	WorkingMemory	Attention	ProcessingSpeed	Inhibition
0	1.56	1.51	-0.18	-0.65
0	0.57	0.94	0.89	1.45
0	1.66	0.58	0.09	0.65
Reasoning	**Language**	**Visual**	**Memory**	**Motor**
0	-0.27	-0.60	0.98	-0.65
0.59	1.03	0.54	0.79	-0.40
1.74	0.90	0.17	1.51	-0.30

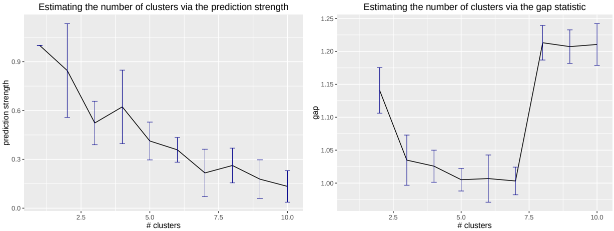
Data Description

The SAS System

The MEANS Procedure

Variable	N	N Miss	Minimum	Mean	Std Dev	Lower Quartile	Median	Upper Quartile	95th Pctl	99th Pctl	Maximum
Group	122	0	0	0.6147541	0.4886602	0	1.0000000	1.0000000	1.0000000	1.0000000	1.0000000
UPDRS_Pt3	118	4	0	12.3305085	11.0438559	3.0000000	9.5000000	21.0000000	32.0000000	44.0000000	46.0000000
DRS2_Total_z	118	4	-1.0000000	0.4152542	0.6653548	0	0.3300000	1.0000000	1.6700000	1.6700000	1.6700000
CRP	120	2	-0.6700000	1.4846665	1.9881994	0.5000000	1.0000000	1.8000000	4.0999900	8.1000000	17.6000000
Homocyst	119	3	0.4000000	10.7764703	4.7636105	8.4000000	10.4000000	12.4000000	17.3000000	22.2000000	45.5000000
WorkingMemory_z	121	1	-0.7833333	1.1427945	2.8156452	0.2133333	0.5733333	1.1400000	1.9166667	16.3999000	18.8999000
Attention_z	122	0	-1.3850000	0.3580464	0.7328176	-0.1200000	0.3800000	0.9350000	1.4350000	1.5750000	2.1150000
ProcessingSpeed_z	122	0	-2.7233333	-0.2226913	0.6932290	-0.6850000	-0.2000000	0.2566667	0.8333333	1.2666667	1.3333333
Inhibition_z	122	0	-2.1000000	0.0493989	0.7442270	-0.4000000	0.0500000	0.6000000	1.2500000	1.4500000	1.6000000
Reasoning_z	122	0	-1.4500000	0.6885246	0.7108863	0.1350000	0.7850000	1.1650000	1.7850000	2.0850000	2.2850000
Language_z	122	0	-1.9000000	0.3546585	0.7172599	-0.1000000	0.3091667	0.8333333	1.4666667	1.8000000	2.3000000
Visual_z	122	0	-2.5000000	0.2652596	0.7390993	-0.1350000	0.3050000	0.8000000	1.3400000	1.5900000	1.5900000
Memory_z	122	0	-2.2100000	0.2027186	0.7549846	-0.3100000	0.2883333	0.7666667	1.3333333	1.5100000	1.5100000
Motor_z	122	0	-3.0500000	-0.7075683	1.0345898	-1.5000000	-0.6500000	-0.0500000	1.0500000	1.8000000	1.8500000
id	122	0	1.0000000	61.5000000	35.3624094	31.0000000	61.5000000	92.0000000	116.0000000	121.0000000	122.0000000

Clustering Result



Both give us 2 clusters, which confirms with the "group" variable. But is it really good?

Clustering Result

Actually 122 patients are divided 155 vs. 6 patients into the two clusters. While the original "group" variable has 47(0) vs. 75(1) patients in each group. The result is not satisfactory.

Discussions

- Although in the original paper, the authors concludes that the prediction strength estimation did best among all other methods in simulation studies, we didn't get the same result.
- Actually Gap statistic outperformed prediction strength in most case, especially when the data is well separated.
- The case study gives us correct cluster numbers, but fail to classify each patients correctly.
- When dealing with data from the real world, we should use caution and get as detailed features (variables) of each observation (patient) as possible. T
- When classifying patients, the judgment and adjudication by doctors and experts is also very necessary.

Selected References

- Tibshirani, R., & Walther, G. (2005)
  "Cluster validation by prediction strength"
  Journal of Computational and Graphical Statistics 14(3), 511-528.
- Tibshirani, R., Walther, G., & Hastie, T. (2001)
  Estimating the number of clusters in a data set via the gap statistic
  Journal of the Royal Statistical Society: Series B (Statistical Methodology) 63(2), 411-423.

Thank you.