Statistics 202: Data Mining

K-means clustering

Based in part on slides from textbook, slides of Susan Holmes

©Jonathan Taylor

December 2, 2012

K-means

Outline

- K-means, K-medoids
- Choosing the number of clusters: Gap test, silhouette plot.
- Mixture modelling, EM algorithm.

K-means

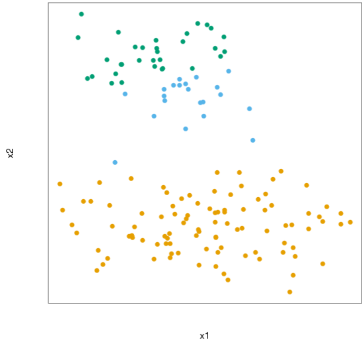


Figure : Simulated data in the plane, clustered into three classes (represented by red, blue and green) by the K-means clustering algorithm. From *ESL*.

K-means

Algorithm (Euclidean)

1. For each data point, the closest cluster center (in Euclidean distance) is identified;
2. Each cluster center is replaced by the coordinatewise average of all data points that are closest to it.
3. Steps 1. and 2. are alternated until convergence. Algorithm converges to a local minimum of the within-cluster sum of squares.

Typically one uses multiple runs from random starting guesses, and chooses the solution with lowest within cluster sum of squares.

K-means

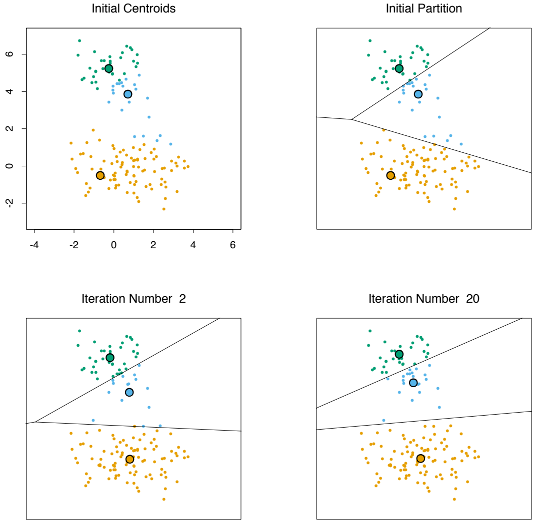
Non-Euclidean

1. We can replace the Euclidean distance squared with some other dissimilarity measure d, this changes the assignment rule to minimizing d.. is identified;
2. Each cluster center is replaced by the point that minimizes the sum of all pairwise d's.
3. Steps 1. and 2. are alternated until convergence. Algorithm converges to a local minimum of the within-cluster sum of d's.

K-means

Statistics 202: Data Mining

©Jonathan Taylor



K-means

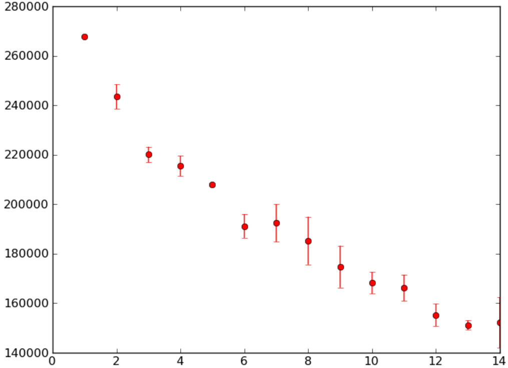


Figure : Decrease in $W(C)$, the within cluster sum of squares.

K-means

Statistics 202: Data Mining

© Jonathan Taylor

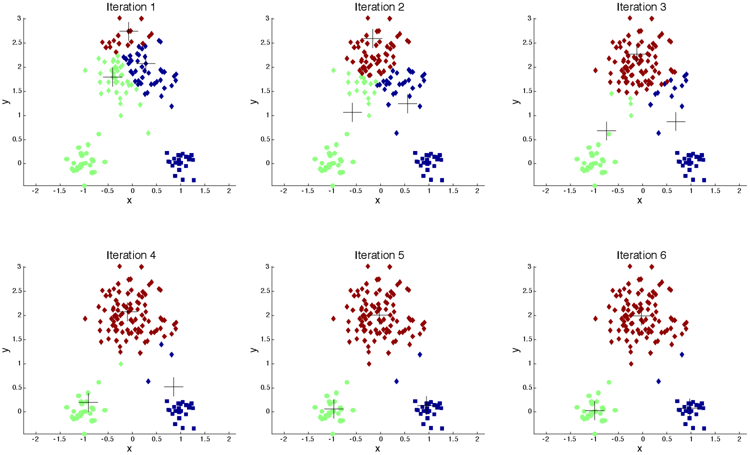
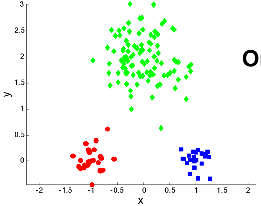
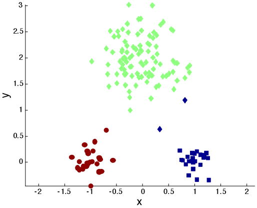


Figure : Another example of the iterations of K-means

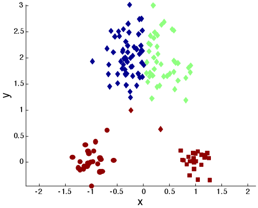
K-means



Original Points

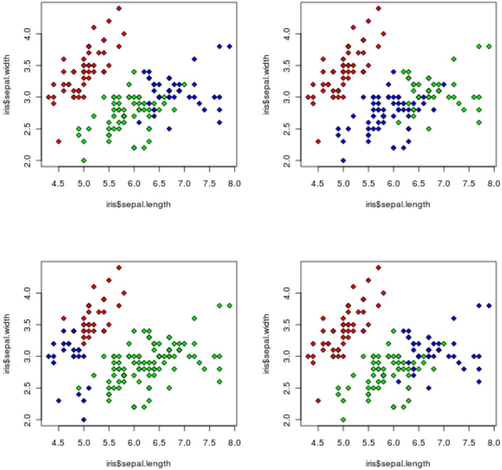


Optimal Clustering



Sub-optimal Clustering

The Iris data (K-means)



K-means

Issues to consider

- Non-quantitative features, e.g. categorical variables, are typically coded by dummy variables, and then treated as quantitative.
- How many centroids k do we use? As k increases, both training and test error decrease!
- By test error, we mean the within-cluster sum of squares for data held-out when fitting the clusters . . .
- Possible to get empty clusters . . .

K-means

Statistics 202: Data Mining

© Jonathan Taylor

Choosing K

- Ideally, the within cluster sum of squares flattens out quickly and we might choose the value of K at this "elbow".
- We might also compare the observed within cluster sum of squares to a *null* model, like uniform on a box containing the data.
- This is the basis of the gap statistic.

K-means

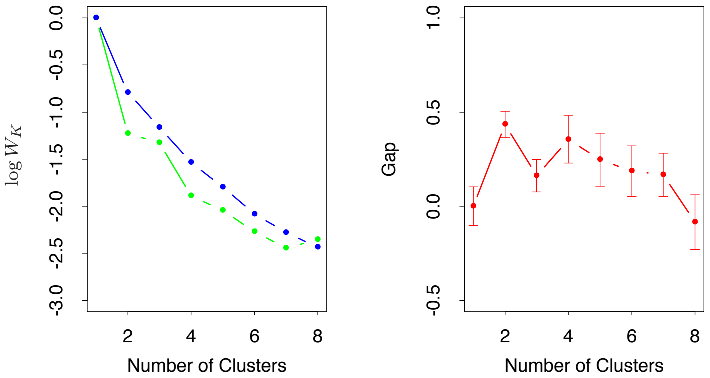


Figure : Blue curve is the W_K for uniform, green curve is for data.

K-means

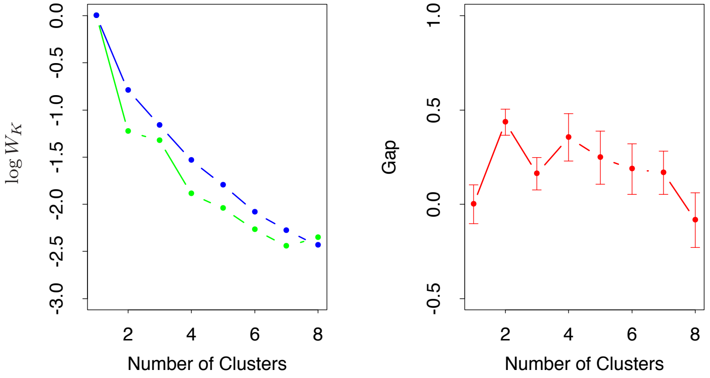


Figure : Largest gap is at 2, and the formal rule also takes into account the variability of estimating the gap.

K-medoid

Algorithm

- Same as K-means, except that centroid is estimated not by the average, but by the observation having minimum pairwise distance with the other cluster members.
- Advantage: centroid is one of the observations— useful, eg when features are 0 or 1. Also, one only needs pairwise distances for K-medoids rather than the raw observations.
- In `R`, the function `pam` implements this using Euclidean distance (not distance squared).

K-medoid

Statistics 202: Data Mining

©Jonathan Taylor

Example: Country Dissimilarities

This example comes from a study in which political science students were asked to provide pairwise dissimilarity measures for 12 countries.

	BEL	BRA	CHI	CUB	EGY	FRA	IND	ISR	USA	USS	YUG
BRA	5.58										
CHI	7.00	6.50									
CUB	7.08	7.00	3.83								
EGY	4.83	5.08	8.17	5.83							
FRA	2.17	5.75	6.67	6.92	4.92						
IND	6.42	5.00	5.58	6.00	4.67	6.42					
ISR	3.42	5.50	6.42	6.42	5.00	3.92	6.17				
USA	2.50	4.92	6.25	7.33	4.50	2.25	6.33	2.75			
USS	6.08	6.67	4.25	2.67	6.00	6.17	6.17	6.92	6.17		
YUG	5.25	6.83	4.50	3.75	5.75	5.42	6.08	5.83	6.67	3.67	
ZAI	4.75	3.00	6.08	6.67	5.00	5.58	4.83	6.17	5.67	6.50	6.92

K-medoid

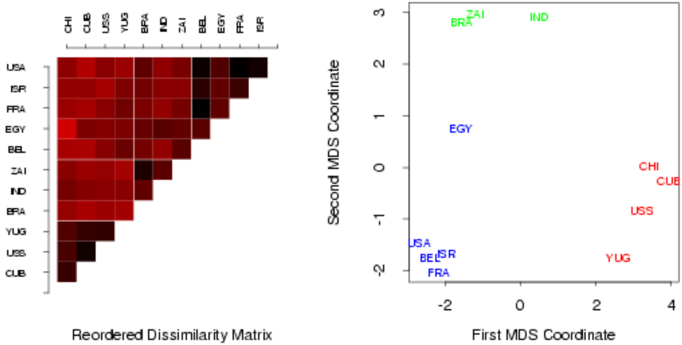
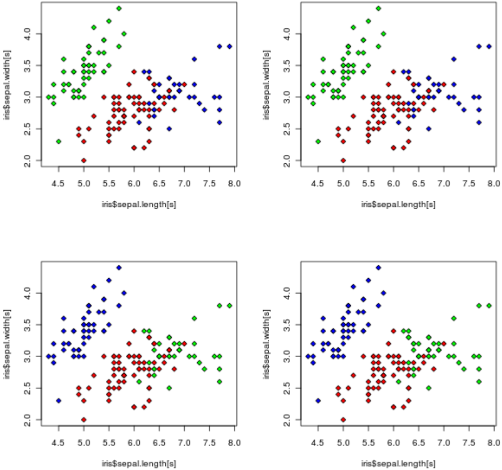


Figure : Left panel: dissimilarities reordered and blocked according to 3-medoid clustering. Heat map is coded from most similar (dark red) to least similar (bright red). Right panel: two-dimensional multidimensional scaling plot, with 3-medoid clusters indicated by different colors.

The Iris data: K-medoid (PAM)



K-medoid

Silhouette

- For each case $1 \leq i \leq n$, and set of cases C and dissimilarity d define

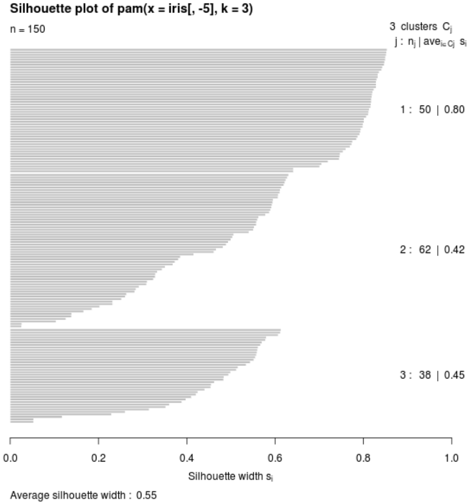
$$\bar{d}(i, C) = \frac{1}{\#C} \sum_{j \in C} d(i, j).$$

- Each case $1 \leq i \leq n$ is assigned to a cluster $C_{l(i)}$. The silhouette width is defined for each case as

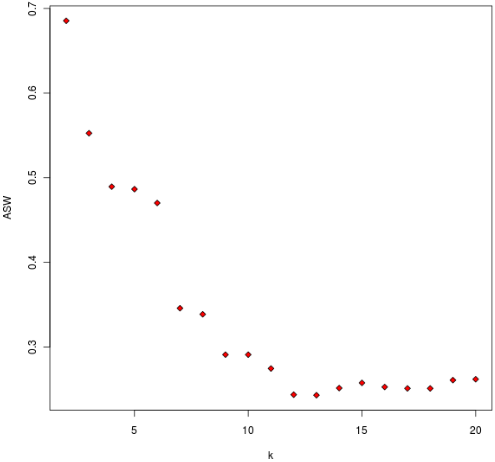
$$\text{silhouette}(i) = \frac{\min_{j \neq l(i)} \bar{d}(i, C_j) - \bar{d}(i, C_{l(i)})}{\max(\bar{d}(i, C_{l(i)}), \min_{j \neq l(i)} \bar{d}(i, C_j))}.$$

- High values of silhouette indicate good clusterings.
- In R this is computable for `pam` objects.

The Iris data: silhouette plot for K-medoid



The Iris data: average silhouette width



Mixture modelling

A soft clustering algorithm

- Imagine we actually had labels $\boldsymbol{Y}$ for the cases, then this would be a classification problem.
- For this classification problem, we might consider using a Gaussian discriminant model like LDA or QDA.
- We would then have to estimate (μ_j, Σ_j) within each "cluster." This would be easy ...
- The next model is based on this realization ...

Mixture modelling

EM algorithm

- The abbreviation: *E=expectation*, *M=maximization*.
- A special case of an *majorization-minimization* algorithm and widely used throughout statistics.
- Particularly useful for situations in which there might be some hidden data that would make the problem easy . . .

Mixture modelling

EM algorithm

- In this mixture model framework, we assume that the data were drawn from the same model as in QDA (or LDA).

$$\begin{aligned} Y &\sim \text{Multinomial}(1, \pi) \qquad \text{(choose a label)} \\ X|Y = \ell &\sim N(\mu_\ell, \Sigma_\ell) \end{aligned}$$

- Only, we have lost our labels and only observe $\boldsymbol{X}_{n \times p}$.
- The goal is still the same, to estimate $\pi, (\mu_\ell, \Sigma_\ell)_{1 \le \ell \le k}$.

Mixture modelling

EM algorithm

- The algorithm keeps track of $(\mu_\ell, \Sigma_\ell)_{1 \le l \le k}$
- It also trackes "guesses" at $\boldsymbol{Y}$ in the form of $\Gamma_{n \times k}$.
- Alternates between "guessing" $\boldsymbol{Y}$ and estimating $\pi, (\mu_\ell, \Sigma_\ell)_{1 \le \ell \le k}$.

Mixture modelling

EM algorithm

Initialize Γ, μ, Σ, π.

Repeat For $1 \leq t \leq T$,

Estimate Γ These are called the *responsibilities*

$$\widehat{\gamma}_{i\ell}^{(t+1)} = \frac{\widehat{\pi}_\ell^{(t)} \phi_{\widehat{\mu}_\ell^{(t)}, \widehat{\Sigma}_\ell^{(t)}}(X_i)}{\sum_{l=1}^K \widehat{\pi}_\ell^{(t)} \phi_{\widehat{\mu}_\ell^{(t)}, \widehat{\Sigma}_\ell^{(t)}}(X_i)}$$

Estimate $\mu_\ell, 1 \leq k$

$$\widehat{\mu}_\ell^{(t+1)} = \frac{\sum_{i=1}^n \widehat{\gamma}_{i\ell}^{(t+1)} X_i}{\sum_{i=1}^n \widehat{\gamma}_{i\ell}^{(t+1)}}$$

This is just weighted average with weights $\widehat{\gamma}_{\cdot\ell}^{(t+1)}$.

Mixture modelling

EM algorithm

Estimate $\Sigma_\ell, 1 \leq k$

$$\widehat{\Sigma}_\ell^{(t+1)} = \frac{\sum_{i=1}^n \widehat{\gamma}_{i\ell}^{(t+1)}(X_i - \widehat{\mu}_\ell^{(t+1)})(X_i - \widehat{\mu}_\ell^{(t+1)})^T}{\sum_{i=1}^n \widehat{\gamma}_{i\ell}^{(t+1)}}$$

This is just a weighted estimate of the covariance matrix with weights $\widehat{\gamma}_{\cdot l}^{(t+1)}$.

Estimate π_ℓ

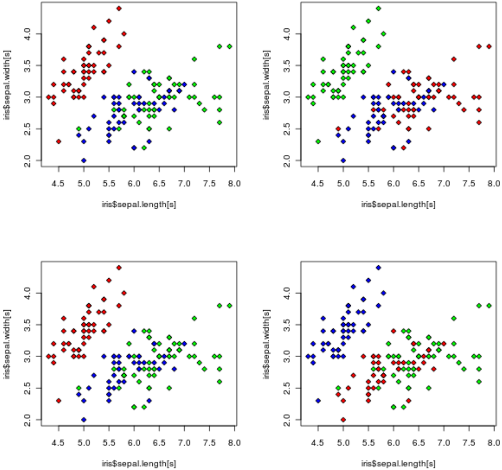
$$\widehat{\pi}_\ell^{(t+1)} = \frac{1}{n}\sum_{i=1}^n \widehat{\gamma}_{i\ell}^{(t+1)}$$

Mixture modelling

EM algorithm

- The quantities Γ are not really parameters, they are "estimates" of the random labels $\boldsymbol{Y}$ which were unobserved.
- If we had observed $\boldsymbol{Y}$ then the rows of Γ would be all zero except one entry, which would be 1.
- In this case, estimation of $\pi_\ell, \mu_\ell, \Sigma_\ell$ is just as it would have been in QDA ...
- The EM simply replaces the unobserved $\boldsymbol{Y}$ with a guess ...

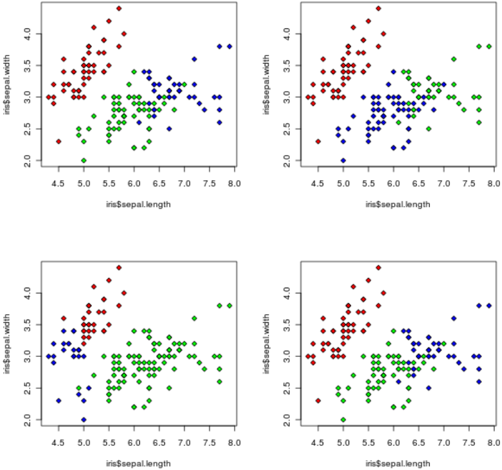
The Iris data: Gaussian mixture modelling



The Iris data (K-means)

Statistics 202: Data Mining

©Jonathan Taylor



The Iris data: silhouette plot for K-medoid

Statistics 202: Data Mining

©Jonathan Taylor