Clustering: kernel and hierarchical methods

P.J. Flynn

Meet the new guy

- CSE Prof @ ND since 2001
- Before: Ohio State EE
- Before^2: Washington State EECS
- Before^3: ND CSE year 1 (!)
- Before^4: Grad student @MSU
- Used clustering a lot over the years
- Clarinet player in 2 ND groups and a community band. Go nets!
- 2 grown-up kids (1 @Purdue, 1 teaching in TX)

Me at OSU haha

Today

- A few preachy slides about clustering in general
- Kernel k-means
- Hierarchical clustering

Yippy (formerly **Clusty**) is a metasearch engine developed by Vivísimo before Vivisimo was later acquired by IBM and renamed IBM Watson Explorer which offers clusters of results.

Clustering: confronting the literature

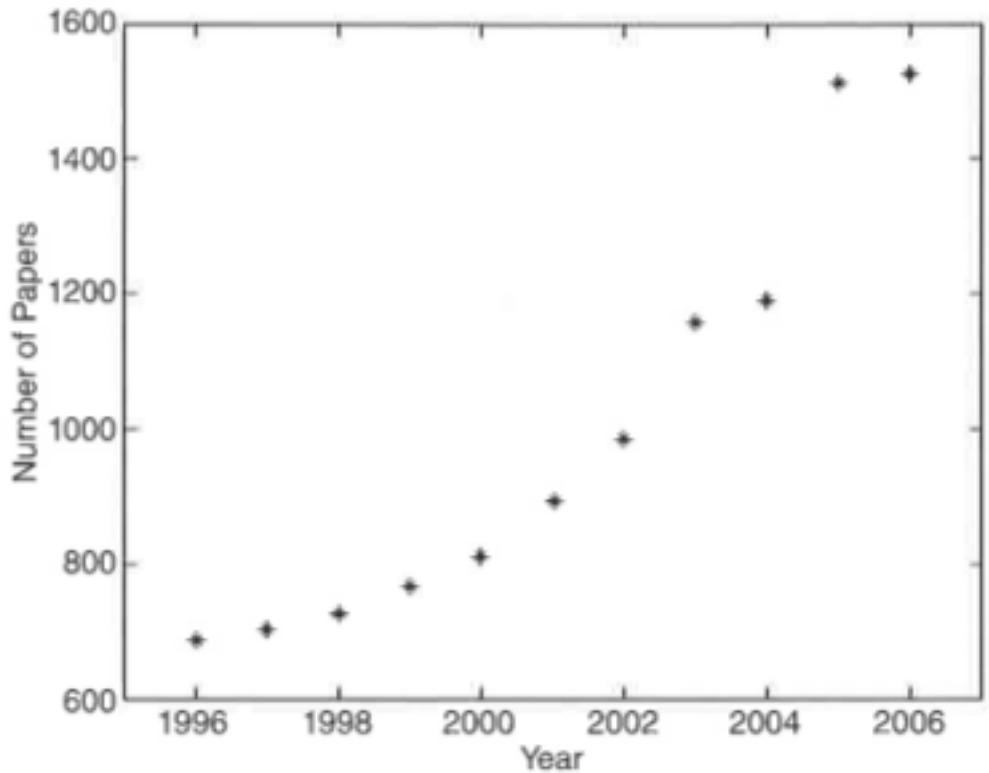


Fig. 1.3. Number of papers on cluster analysis from 1996 to 2006. The searches were performed using Web of Science ®, which includes three databases: the Science Citation Index Expanded™ (SCI_EXPANDED), the Social Sciences Citation Index ® (SSCI), and the Arts & Humanities Citation Index ® (A&HCI).[1]

Source: Xu and Wunsch, *Clustering*, IEEE/Wiley, 2009

Crisp Clustering: it's easy to define.

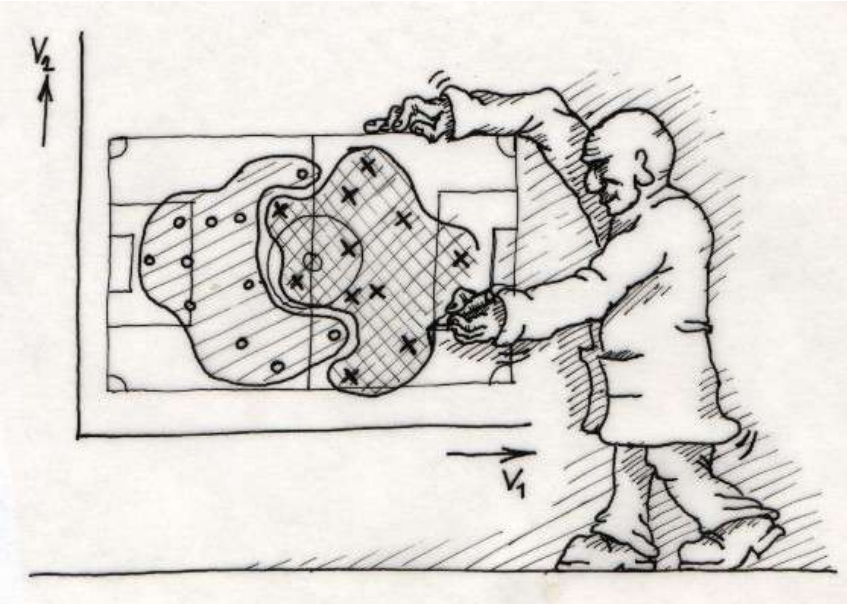
- The "best" clustering containing *k* clusters is that exclusive and exhaustive partitioning of items into *k* non-empty groups that minimizes the sum of distances between items within the group
- Directly leads to squared-error criteria

$$\min \; E_k^2 = \sum_{i=1}^{k} e_i^2 \quad \text{where} \quad e_i^2 = \sum_{j=1}^{n_i} \left\| \vec{x}_j^{(i)} - \vec{m}^{(i)} \right\|^2$$

Sum of squared distances between the points x_j in cluster i and the mean

And it's so simple to do

- You have N items
- You need to assign an integer from 1 to k to each of the N items.
- How hard can that be?
- Key problems
  - Proximity/similarity: how do you measure it?
  - Combinatorics
  - What's k?



http://www.vias.org/science_cartoons/cluster_analysis.html

Crisp Clustering and Stirling numbers

- N items, k clusters: each cluster nonempty; exhaustive & exclusive (every item gets a single label from 1 to k)
- Number of clusterings of size k containing N items is a *Stirling number of the second kind:*

$$S(N,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^{k-i}\binom{k}{i}i^N$$

Courtesy Doug Downey and Pryan Pardo, Northwestern, via CSU Pomona

Proximity: how to measure

- Sometimes it's baked into the data: geographical or other "physical" data
- Otherwise, you need a heuristic

Dilbert, June 13, 1997

How many clusters?

- Decades-old problem
- No theory, and no heuristic works well everywhere
- Some rules of thumb
  - Domain-specific criteria may imply/suggest k
  - Set a kMax, run clusterings up to kMax, choose k<=kMax that optimizes a validity criterion

HOW MANY CLUSTERS ARE BEST? – AN EXPERIMENT*

RICHARD C. DUBES
Department of Computer Science, Michigan State University, East Lansing, MI 48824, U.S.A.

Fave figure (Jason Grant's MS thesis)

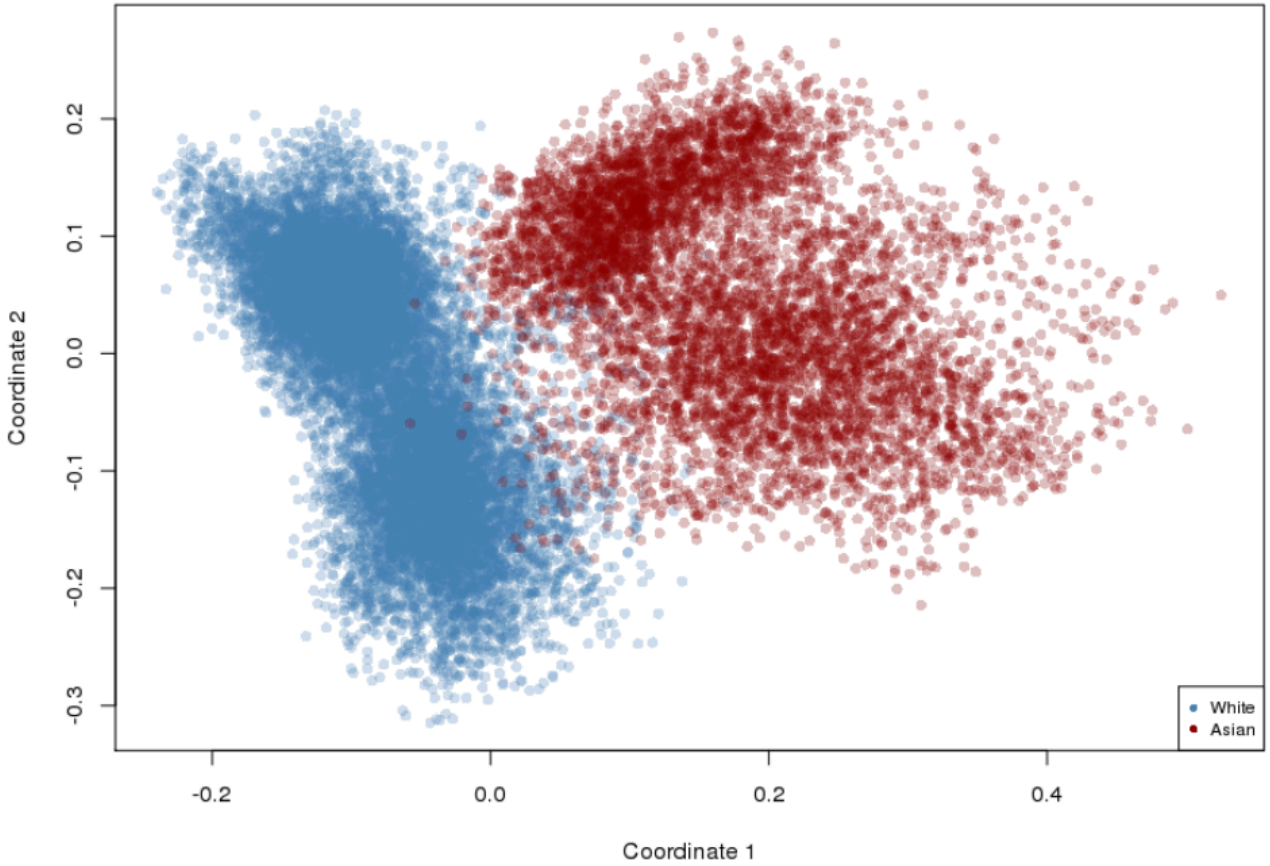
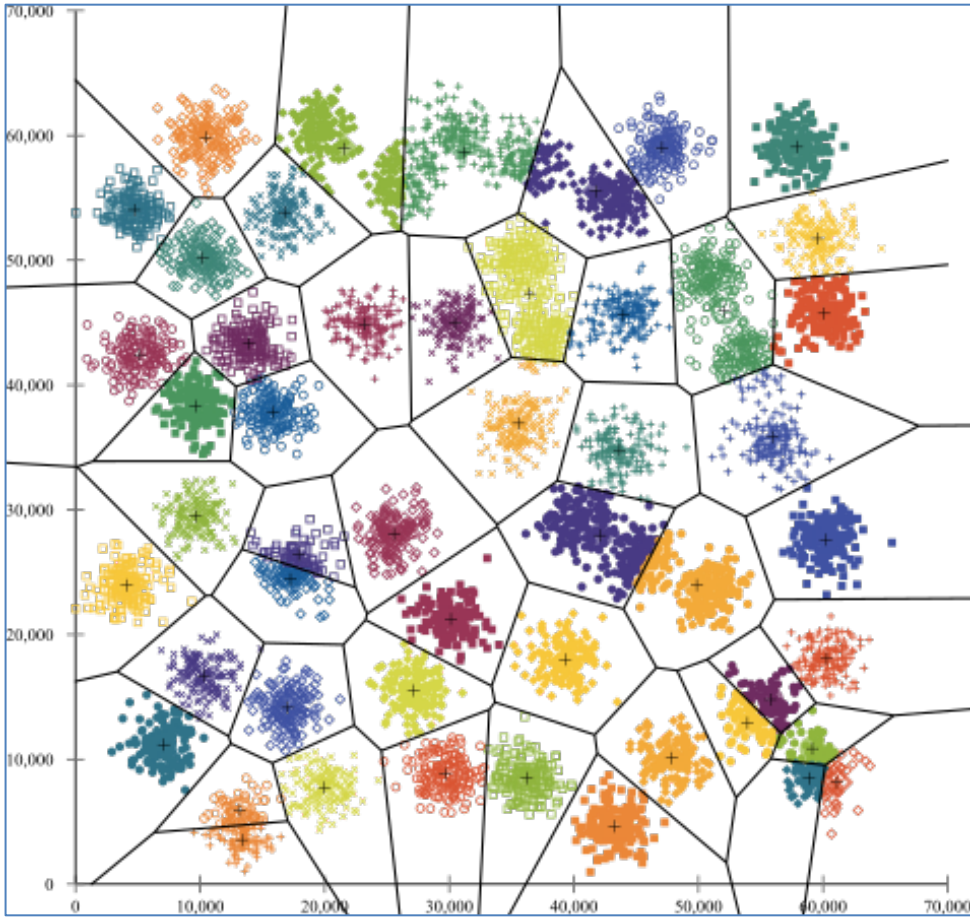


Figure 2.11: Two-dimensional representation of proximity data for face images of the FRGC ver2.0 dataset categorized by ethnicity.



END OF SERMON

Partitioning-Based Clustering Methods

- Basic Concepts of Partitioning Algorithms
- The K-Means Clustering Method
- Initialization of K-Means Clustering
- The K-Medoids Clustering Method
- The K-Medians and K-Modes Clustering Methods
- **The Kernel K-Means Clustering Method**

Data clustering: 50 years beyond K-means ☆

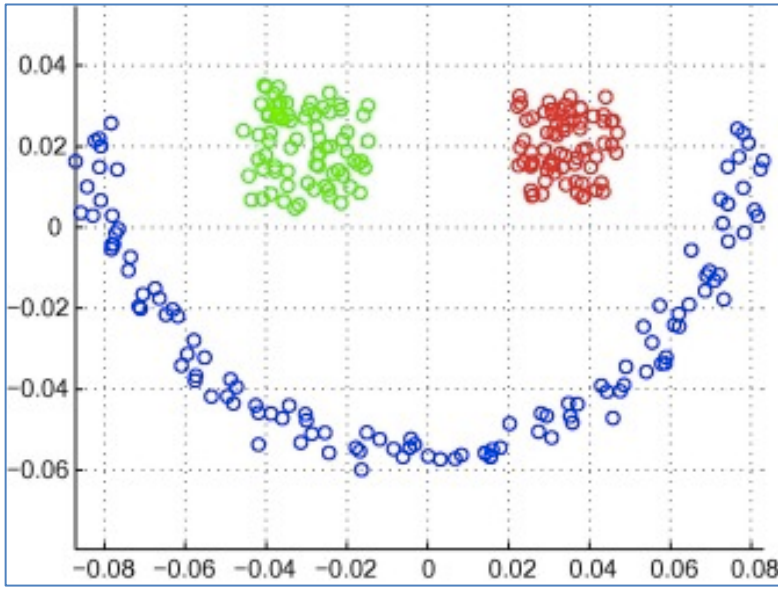
Anil K. Jain *

Department of Computer Science and Engineering, Michigan State University, East Lansing, Michigan 48824, USA
Department of Brain and Cognitive Engineering, Korea University, Anam-dong, Seoul, 136-713, Korea

Patt Recog. Letters 31, 2010

Kernel K-Means Clustering: o'view

- Perform k-means, but in a different feature space
- Conceptually, use $\varphi(x_i)$, instead of x_i, as the points you are clustering
- $\varphi(x_i)$ is a vector-valued function of x_i
- BUT you never need to compute $\varphi(x_i)$
- You do need to compute and store n x n kernel matrix generated from the kernel function on the original data
- Computational complexity is higher than K-Means



Kernel trick (recap)

Pick yourself a kernel. Common choices are

- Gaussian kernel: $K(x_i,x_j) = x_i^T x_j$
- Polynomial kernel: $K(x_i,x_j) = (x_i^T x_j + c)^d$
  - c is a free parameter (can tune for performance)
  - d is the desired degree of the polynomial
- RBF kernel: $K(x_i,x_j) = \exp(-\gamma|x_i-x_j|^2)$
  - $\gamma = (-1/2\sigma^2)$
  - Σ is a tunable free parameter ("influence")

Recall that $K(x_i,x_j) = \varphi(x_i)^T\varphi(x_j)$
The kernel trick allows you to avoid computing $\varphi(x_i)$

Kernel k-means: setup

Given all of those points x_i, i=1…N,

- Compute Gram matrix $K = [k_{ij}]$
  - $k_{ij} = K(x_i,x_j)$ for a prespecified kernel $K(x_i,x_j)$
  - Positive semidefinite matrix
- Choose initial centers m_i, i = 1..k at random

Kernel k-means

- Standard algorithm, except use this to calculate transformed-point-to-centroid distances

$$\|\phi(x_i) - m_k\| = K_{ii} - \frac{2}{n_k} \sum_{j|L(x_j)=k} K_{ij} + \frac{1}{n_k^2} \sum_{\substack{j|L(x_j)=k \\ m|L(x_m)=k}} K_{jm}$$

Example: Kernel Functions and Kernel K-Means Clustering

- Gaussian radial basis function (RBF) kernel: $K(\boldsymbol{X}_i, \boldsymbol{X}_j) = e^{-\|X_i - X_j\|^2/2\sigma^2}$
- Suppose there are 5 original 2-dimensional points: $K_{x_i x_j} = \phi(x_i) \bullet \phi(x_j)$
  - $x_1(0, 0)$, $x_2(4, 4)$, $x_3(-4, 4)$, $x_4(-4, -4)$, $x_5(4, -4)$
- If we set σ to 4, we will have the following points in the kernel space
  - E.g., $\left|\left|x_1 - x_2\right|\right|^2 = (0-4)^2 + (0-4)^2 = 32$, therefore, $K(x_1, x_2) = e^{-\frac{32}{2 \cdot 4^2}} = e^{-1}$

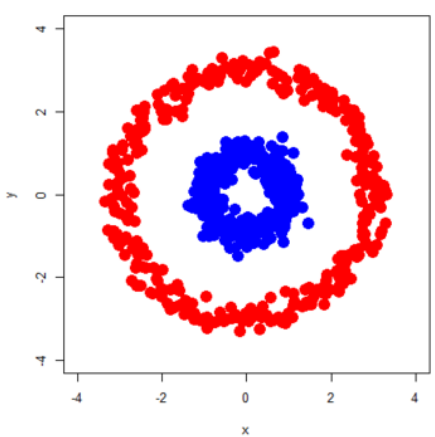
Original Space

RBF Kernel Space ($\sigma = 4$)

	x	y	$K(x_i, x_1)$	$K(x_i, x_2)$	$K(x_i, x_3)$	$K(x_i, x_4)$	$K(x_i, x_5)$
x_1	0	0	1	$e^{-\frac{4^2+4^2}{2 \cdot 4^2}} = e^{-1}$	e^{-1}	e^{-1}	e^{-1}
x_2	4	4	e^{-1}	1	e^{-2}	e^{-4}	e^{-2}
x_3	-4	4	e^{-1}	e^{-2}	1	e^{-2}	e^{-4}
x_4	-4	-4	e^{-1}	e^{-4}	e^{-2}	1	e^{-2}
x_5	4	-4	e^{-1}	e^{-2}	e^{-4}	e^{-2}	1

Example: Kernel Functions and Kernel K-Means Clustering

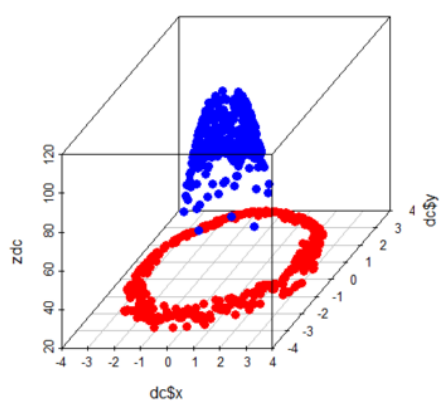
- Gaussian radial basis function (RBF) kernel: $K(\boldsymbol{X}_i, \boldsymbol{X}_j) = e^{-\|X_i - X_j\|^2/2\sigma^2}$

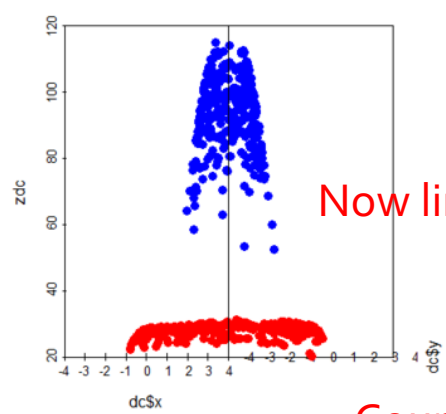


$$\underset{\mathcal{J}_1,\ldots,\mathcal{J}_k}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{j \in \mathcal{J}_i} \left\| \mathbf{a}_j - \frac{1}{|\mathcal{J}_i|} \sum_{l \in \mathcal{J}_i} \mathbf{a}_l \right\|_2^2.$$

$$\underset{\mathcal{J}_1,\ldots,\mathcal{J}_k}{\operatorname{argmin}} \sum_{i=1}^{k} \sum_{j \in \mathcal{J}_i} \left\| \phi(\mathbf{a}_j) - \frac{1}{|\mathcal{J}_i|} \sum_{l \in \mathcal{J}_i} \phi(\mathbf{a}_l) \right\|_2^2.$$

$$\kappa(\mathbf{a}_i, \mathbf{a}_j) = \langle \phi(\mathbf{a}_i), \phi(\mathbf{a}_j) \rangle.$$



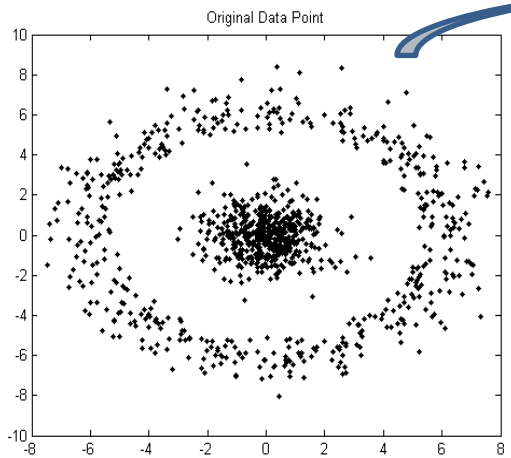


Now linearly separable!!!

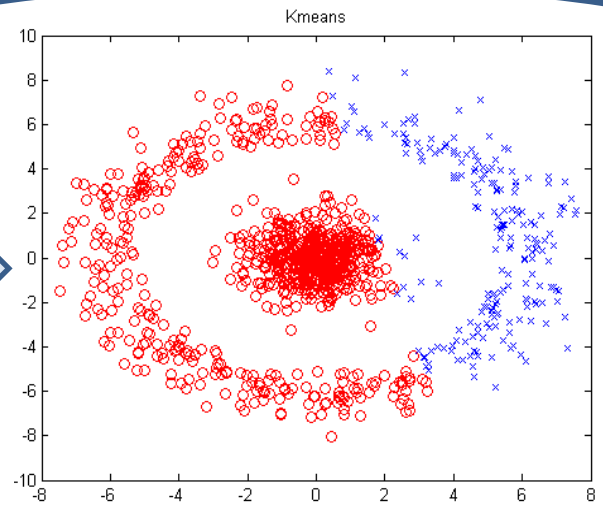
$$\exp(z) = \sum_{k=0}^{\infty} \frac{z^k}{k!} = 1 + z + \frac{z^2}{2} + \frac{z^3}{6} + \frac{z^4}{24} + \cdots$$

Countless new features in RBF kernel space…

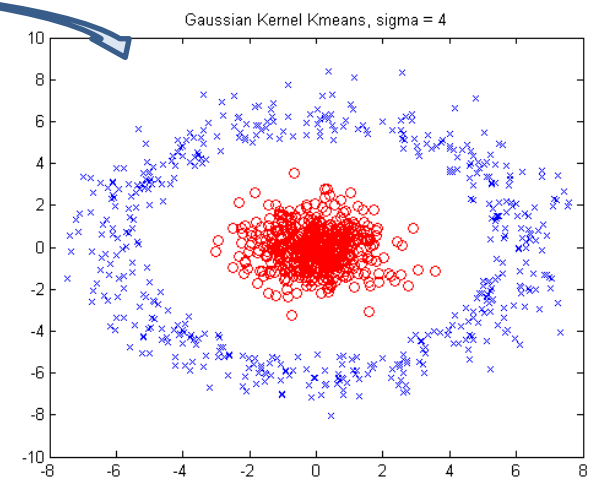
Example: Kernel K-Means Clustering



The original data set



The result of K-Means clustering



The result of Gaussian Kernel K-Means clustering

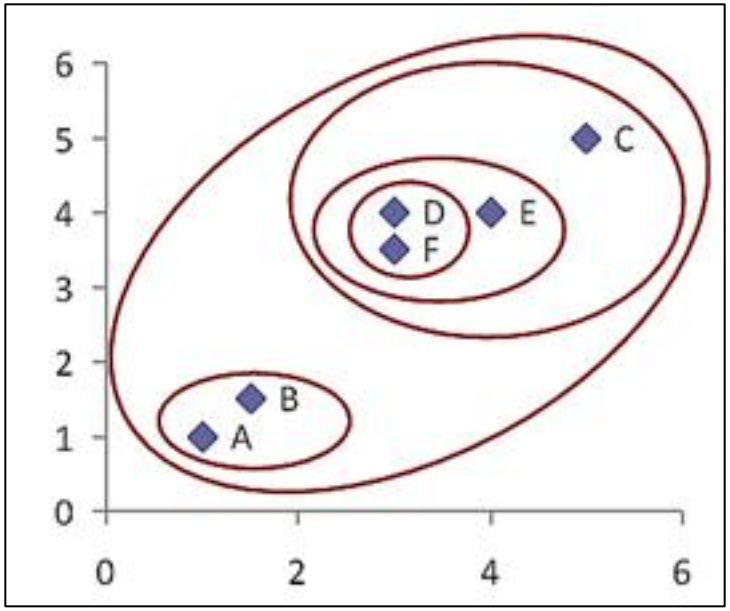
- The above data set cannot generate quality clusters by K-Means since it contains non-convex clusters
- Gaussian RBF Kernel transformation maps data to a kernel matrix K for any two points x_i, x_j: $K_{x_i x_j} = \phi(x_i) \bullet \phi(x_j)$ and Gaussian kernel: $K(\boldsymbol{X_i}, \boldsymbol{X_j}) = e^{-\|X_i - X_j\|^2 / 2\sigma^2}$
- K-Means clustering is conducted on the mapped data, generating quality clusters

References: (II) Partitioning Methods

- J. MacQueen. Some Methods for Classification and Analysis of Multivariate Observations. In Proc. of the 5th Berkeley Symp. on Mathematical Statistics and Probability, 1967
- S. Lloyd. Least Squares Quantization in PCM. IEEE Trans. on Information Theory, 28(2), 1982
- A. K. Jain and R. C. Dubes. Algorithms for Clustering Data. Prentice Hall, 1988
- R. Ng and J. Han. Efficient and Effective Clustering Method for Spatial Data Mining. VLDB'94
- B. Schölkopf, A. Smola, and K. R. Müller. Nonlinear Component Analysis as a Kernel Eigenvalue Problem. Neural computation, 10(5):1299–1319, 1998
- I. S. Dhillon, Y. Guan, and B. Kulis. Kernel K-Means: Spectral Clustering and Normalized Cuts. KDD'04
- D. Arthur and S. Vassilvitskii. K-means++: The Advantages of Careful Seeding. SODA'07
- C. K. Reddy and B. Vinzamuri. A Survey of Partitional and Hierarchical Clustering Algorithms, in (Chap. 4) Aggarwal and Reddy (eds.), Data Clustering: Algorithms and Applications. CRC Press, 2014

Hierarchical clustering

- Suppose you want clusters-within-clusters.
- Why? You might suspect that the data reflects a hierarchical process and want to recover the hierarchy (it might matter more than the data)

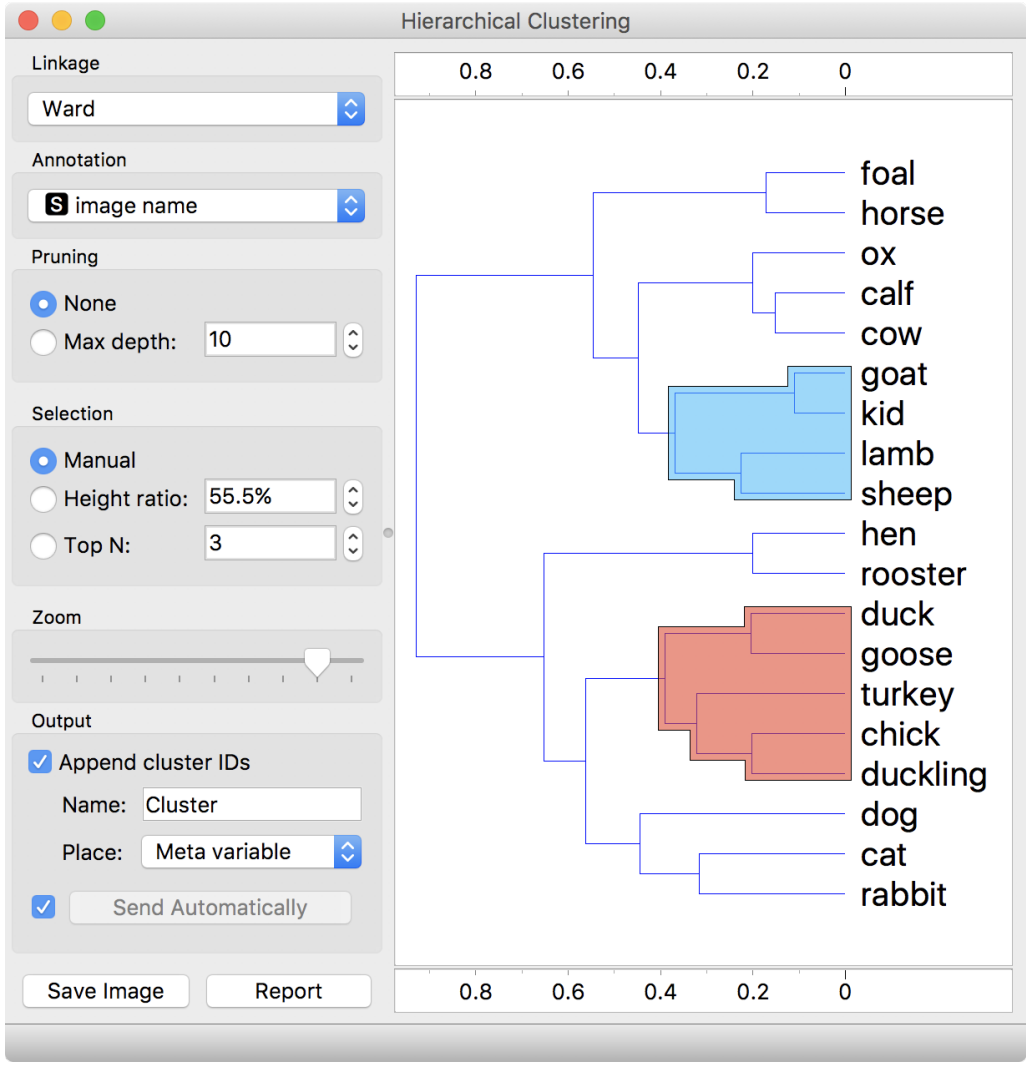


Two basic approaches

- Agglomerative (bottom-up)
  - Start with each item in its own cluster, then merge the clusters according to some criterion, until only one cluster is present.
- Divisive (top-down)
  - Start with one, divide, end with each in its own cluster
- In either case, can "stop early" with an intermediate number of clusters
- In both cases, some notion of "similarity" or "dissimilarity" drives the merges/splits. This is based exclusively on a similarity or dissimilarity measure.

Representing clustering hierarchies: dendrogram

- Membership versus similarity
- See merging happening at various levels
- Figure: Wikipedia (Orange software)



Generic agglomerative clustering

- Start with each item in its own cluster.
- WHILE more than one cluster
  - Merge the two **"closest"** clusters, as measured by a specific inter-cluster distance measure (perhaps keep track of the similarity where the merge occurred)

Inter-cluster distances used for agglomerative clustering

- Clusters C_i, C_j

- Min(C_i,C_j): min {|p-p'|, forall p in C_i and p' in C_j}
- Max(C_i,C_j): max {|p-p'|, forall p in C_i and p' in C_j}
- Mean(C_i,C_j): $|m_i - m_j|$, m_i is mean of C_i, etc.
- Average(Ci,Cj): average of all distances between a point in Ci and a point in Cj

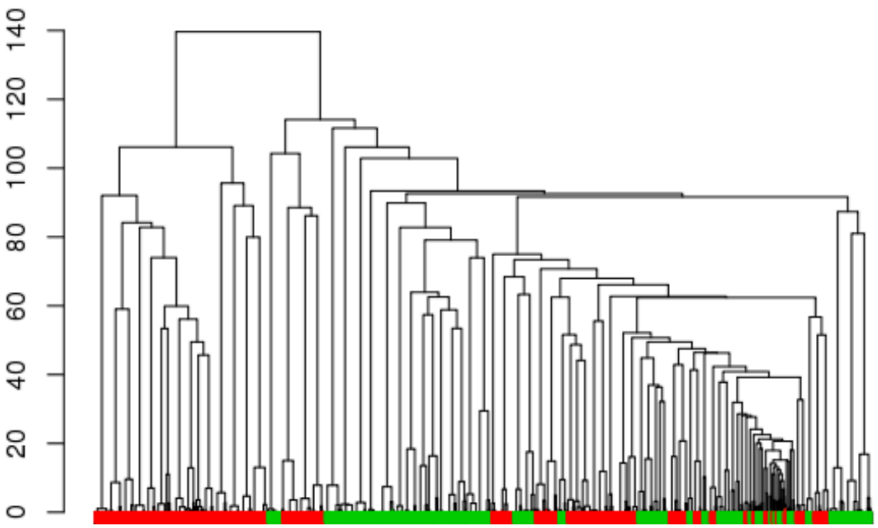
X-link

- Single-Link uses min (also called nearest-neighbor clustering, MST clustering)
- Complete-link uses max (also called farsthest-neighbor clustering)

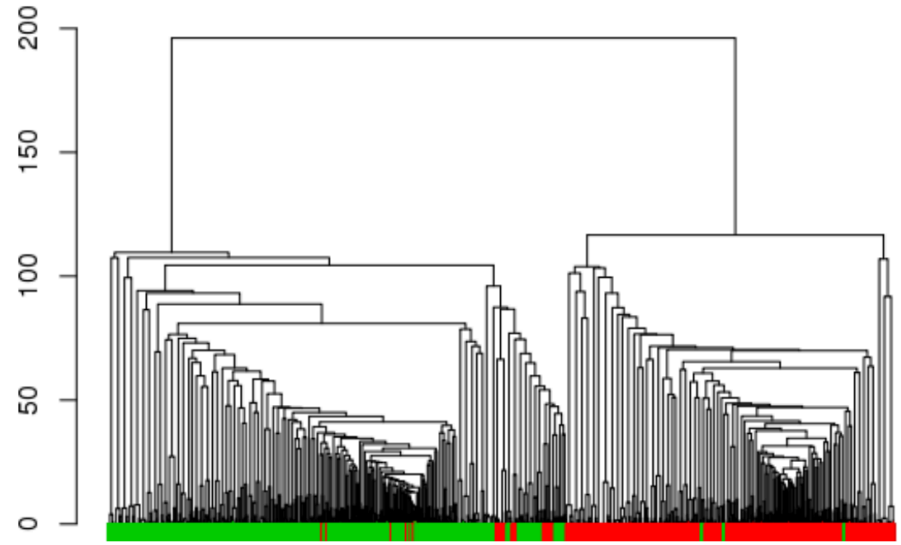
Graph-theoretic view

- Consider the items as nodes in a graph and the distances between them as weighted edges
- Specify a distance threshold t and throw away edges longer than the threshold
- Consider the resulting graph
- Single-link tends to produce "straggly" clusters (chaining/connected components); complete-link produces more compact clusters and enforces maximal cliquing

Dendrograms from Jason Grant's MS thesis



(a) Clustering of images from Asian subjects



(b) Clustering of images from white subjects

Figure 2.9: Shown are the clusterings of Asian and white subjects using Ward's method. Images female subject are shown in red and male subjects in green. While there appears to be a natural grouping between white male and female subjects, these observations are not as evident in with Asian subjects.