Modern Methods of Statistical Learning sf2935 Unsupervised Learning Part 3: Hierarchic Clustering

Timo Koski

TK

05.10.2018

KTH Matematik

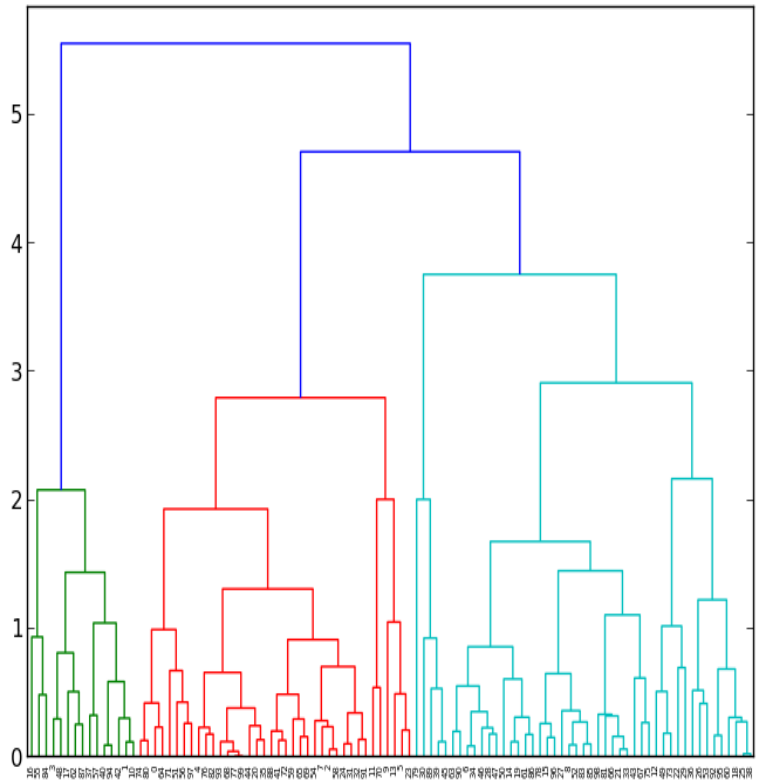
Clustering :
OUTLINE OF CONTENTS:

- Hierarchic clustering: dendrograms
- Woodger-Gregg axioms, ultrametricity
- Hierarchical Agglomerative methods of Clustering SLINK, CLINK, UPGMA WARD
- Fisher's Iris Data
- Chan-Darwiche Distance, Predictive Distributions, and Statistical Clusters
- Bayesian Hierarchic Sample Clustering

KTH VETENSKAP OCH KONST
KTH Matematik

Dendrograms

Dendrograms are trees that visualize hierarchic clusterings of a



finite set of data.

Dendrograms

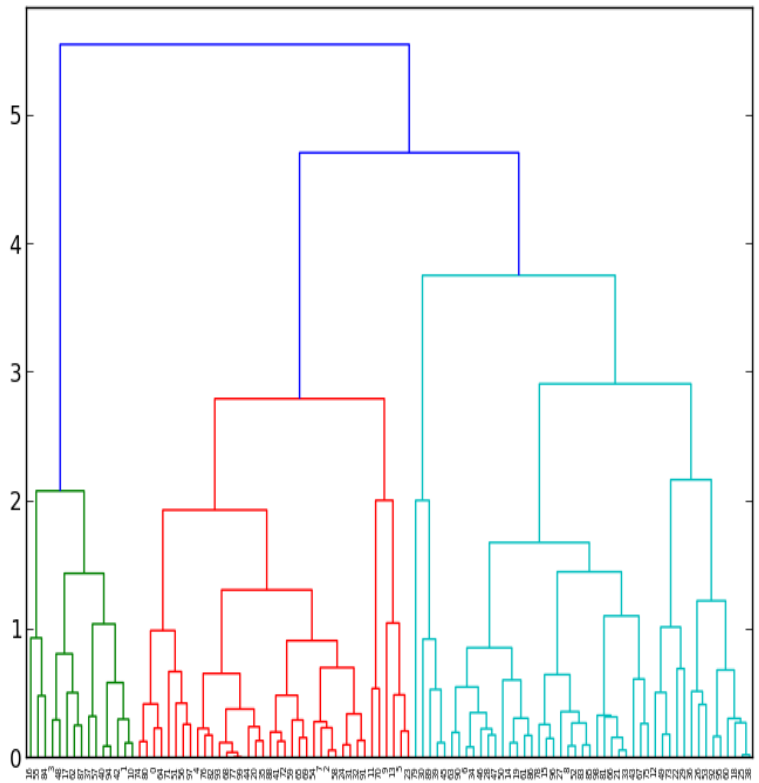
Dendrograms are extensively used in data analysis, where the goal is to extract summaries, with a high information content, of the original data and to construct hypotheses within the subject involved.

KTH VETENSKAP OCH KONST

KTH Matematik

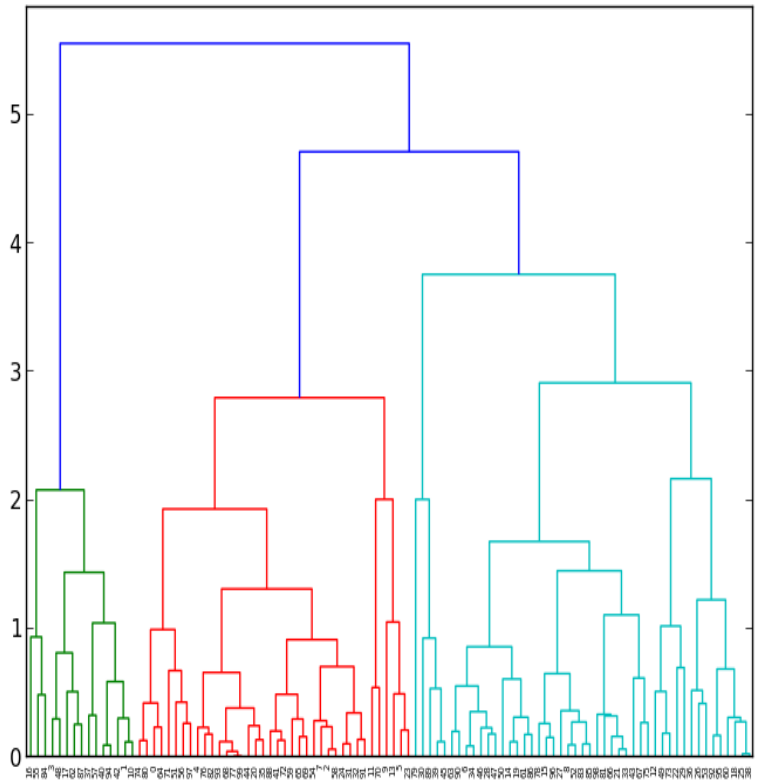
Dendrograms

The dendrogram has branching levels or ultrametric ranks, understood as levels of proximity (or cophenetic value). Each level in the tree corresponds to a partition of the set of data and the partitions are ordered by inclusion of partition sets.



KTH Matematik

Actual clustering is done by cutting the dendrogram at some threshold value indicating the level of dissimilarity in the clusters joined at that value.



KTH Matematik

Dendrograms

Dendrograms were popularized and advocated for numerical biological taxonomy and are used widely in all kinds of practical data analysis, as they are intuitive and easily visualized (as trees) to give a wealth of information. A technical definition is also intuitive, and is recapitulated below.

KTH VETENSKAP OCH KONST

KTH Matematik

Dendrograms

A feature of data analysis by means of dendrograms is that statistical models of clustering are frequently absent. Of course, probabilistic models for random dendrograms have been developed. The difficulty for statistical modelling of clustering hierarchies is that a dendrogram is technically speaking a *set tree*, as the tree corresponds to nested partitions of one single set of data.

KTH VETENSKAP OCH KONST

KTH Matematik

Dendrograms

P.H.A. Sneath analysed the issues involved as follows:

> *. . . (hierarchies) can be legitimately regarded as a system for generating hypotheses. Probabilistic models are then unimportant, because the hypotheses would be tested by further work on new observations.*

We shall take a look at a method of Bayesian learning of dendrograms invoking predictive distributions. This constitutes also a method for probabilistic learning in hierarchic clustering.

Dendrograms

A hierarchic clustering is an algorithm that has as in-data a matrix of dissimilarities or proximities matrix calculated on some set of basic data. The **algorithm transforms the dissimilarities to an (ultrametric) tree**. We are going to think of the dissimilarity data as statistical samples of a distribution that depends on the true ultrametric tree, or a hypothesis about it. We shall consider a primary hierarchic clustering as an estimate of the true tree .

KTH VETENSKAP OCH KONST
KTH Matematik

Dendrograms

Dendrograms were popularized and advocated for numerical biological taxonomy and are used widely in all kinds of practical data analysis, as they are intuitive and easily visualized (as trees) to give a wealth of information. A technical definition is also intuitive, but requires notions and terminology to be recapitulated now.

KTH VETENSKAP OCH KONST
KTH Matematik

Dendrograms: technical terminology

We start with sequences of partitions of a finite set $\mathcal{P}$.

1. $\mathcal{P}$ is a finite set. The cardinality $\mid \mathcal{P} \mid$ of $\mathcal{P}$ is ≥ 2. The elements of $\mathcal{P}$ will denoted by $a, b, c, \ldots, p, \ldots, x, y$. The elements are abstract at this level of discussion, but could for the sake of definiteness be called operative taxonomic units (OTUs).
2. N is a finite positive integer.
3. $M(n)$ is a function defined on integers n in $0 \ldots N$ and taking its values in partitions of $\mathcal{P}$, or, in the equivalence relations on $\mathcal{P}$ such that the 'Woodger-Gregg' axioms(a)-(c) hold:
   (a) $M(0) = \{(p, p) \mid p \in \mathcal{P}\}$.
   (b) $M(N) = \mathcal{P} \times \mathcal{P}$.
   (c) $0 \leq n \leq m \leq N \Rightarrow M(n) \subseteq M(m)$.

KTH VETENSKAP OCH KONST
KTH Matematik

Definition

A *Clustering Hierarchy* H is a triple $H = (\mathcal{P}, N, M)$ with $\mathcal{P}, N, M$ as given above.

□

We shall denote the set of equivalence classes in the image of $M(n)$ by

$$\mathbf{P}(n) = \mathcal{P} / M(n). \tag{1}$$

Each $\mathbf{P}(n)$ is a partition of $\mathcal{P}$.

KTH VETENSKAP OCH KONST
KTH Matematik

From this and from the properties (a)-(c) it follows that for $n \leq m$

$$\mathbf{P}(0) \leq \mathbf{P}(n) \leq \mathbf{P}(m) \leq \mathbf{P}(N),$$

where $\leq$ is the usual order for partitions, i.e., $\mathbf{P}(n) \leq \mathbf{P}(m)$ means that $\mathbf{P}(m)$ is a coarser partition of $\mathcal{P}$ than $\mathbf{P}(n)$. $\mathbf{P}(N)$ is the maximal element and $\mathbf{P}(0)$ is the minimal element, and $\mathbf{P}(N)$ is the coarsest, where all OTU's are in a single set. $\mathbf{P}(0)$ is the the most refined partition consisting of singletons. Clearly $\leq$ is a partial order and $(\mathbf{P}(n))_{n=0}^{N}$ is a lattice. Now we introduce a technical conceptual aid, which is, called taxon in a hierarchy.

KTH VETENSKAP OCH KONST
KTH Matematik

Definition

Y is a taxon in the hierarchy $(\mathcal{P}, N, M)$, if there is an integer n in $0, \ldots, N$ such that

$$Y \in \{(A, n) \mid A \in \mathbf{P}(n)\}.$$

□

Here (A, n) is an ordered pair. If Y is a taxon in a hierarchy, we say that

$$\mathrm{Ext}(Y) = A, \mathrm{Rank}(Y) = n.$$

On occasion we shall also talk about the sets $\mathrm{Ext}(Y)$ as extensions. In this terminology, $\mathbf{P}(n) \leq \mathbf{P}(m)$ means that every extension of a taxon in $\mathbf{P}(m)$ is a union of extensions in $\mathbf{P}(n)$.

KTH VETENSKAP OCH KONST
KTH Matematik

Next we can introduce a map $J(x, y)$ from pairs of OTU's to $\{0 \dots N\}$ initially named as the *cophenetic value* .

Definition

Let $x, y \in \mathcal{P}$ and let Y be the taxon of lowest rank, $\mathrm{Rank}(Y)$, in the hierarchy $(\mathcal{P}, N, M)$ such that $x, y \in \mathrm{Ext}(Y)$. Then we set

$$J(x, y) = \mathrm{Rank}(Y). \tag{2}$$

□

The definition presupposes tacitly that Y is unique, which holds by the properties of partitions.

KTH VETENSKAP OCH KONST
KTH Matematik

Dendrograms

Due to this we can introduce $\mathrm{lca}(x, y)$, the *least common ancestor* of x and y by

$$\mathrm{lca}(x, y) = \mathrm{Ext}(Y), \quad \text{such that} \quad \mathrm{Rank}(Y) = J(x, y). \tag{3}$$

Let now H be a clustering hierarchy and let $J(x, y)$ be the cophenetic value defined in (2).

Proposition

$$J(x, z) \leq \max\{J(x, y), J(y, z)\}. \tag{4}$$

$$J(x, y) = J(y, x) \tag{5}$$

$$J(x, y) = 0 \Leftrightarrow x = y. \tag{6}$$

Ultrametrics

$$J(x,z) \leq \max\{J(x,y), J(y,z)\}. \tag{7}$$

An interpretation of (7): If the difference between x and z is big, and the difference between x and y is small, then the difference between y and z has to be big.

KTH VETENSKAP OCH KONST

KTH Matematik

$$J(x,z) \leq \max\{J(x,y), J(y,z)\}. \tag{8}$$

The inequality (8) is the *ultrametric inequality* and we say that $J(x,y)$ is an *ultrametric* on $\mathcal{P} \times \mathcal{P}$.

Rammal, Rammal and Toulouse, Gérard and Virasoro, Miguel Angel: Ultrametricity for physicists, *Reviews of Modern Physics*, 58, 3, 765−, 1986.

Proof: Since $J(x, y) = Rank(Y)$ given $lca(x, y) = Ext(Y)$ and Rank(Y) $\in \{0, 1, .., N\}$ it follows that $0 \leq J(x, y)$. If $J(x, y) = 0$ then Ext(Y) $\in P(0)$ which implies that Ext(Y) is a singleton set, thus $x = y$. The symmetry $J(x, y) = J(y, x)$ follows from the fact that if Y is the taxon with the lowest rank such that $x, y \in Ext(Y)$ then Y is the taxon with the lowest rank such that $y, x \in Ext(Y)$, thus $J(x, y) = Rank(Y) = J(y, x)$.

To prove the ultrametric inequality $J(x,y) \leq max(J(x,z), J(z,y))$, we assume without loss of generality that $k = J(x,z) < J(x,y) = m$. If this is the case then $P(k) \leq P(m)$ which implies that $lca(x,z) \subset lca(x,y)$. But if $lca(x,z) \subset lca(x,y)$ then x and z are in the same set in every partition $P(l)$ where $k \leq l \leq m$, which implies $lca(x,y) = lca(z,y)$ which implies that $J(x,y) = J(z,y)$.

□

Dendrograms

Suppose now that $d(x, y)$ is any ultrametric (i.e., $d(x, y)$ satisfies (4), (5) and (6)) on $\mathcal{P} \times \mathcal{P}$ and let $d(\mathcal{P}, \mathcal{P})$ be the set of all values of the ultrametric. Let next λ be a real valued function defined on $d(\mathcal{P}, \mathcal{P})$ to be called the valuation function. We require that λ satisfies the following conditions

i) $\lambda(0) = 0$.

ii) $r \in d(\mathcal{P}, \mathcal{P})$, and $r > 0 \Rightarrow 0 < \lambda(r)$.

iii) $r < r'$ and $r, r' \in d(\mathcal{P}, \mathcal{P}), \quad \lambda(r) \leq \lambda(r')$.

KTH Matematik

Proposition

Let $d(x, y)$ be an ultrametric on $\mathcal{P} \times \mathcal{P}$. Let λ be an evaluation function. If N is given by

$$N = \lceil \max_{x,y \in \mathcal{P} \times \mathcal{P}} \lambda(d(x, y)) \rceil, \tag{9}$$

and $M_\lambda(n)$ is for $0 \leq n \leq N$ defined as

$$(x, y) \in M_\lambda(n) \Leftrightarrow \lambda\left(d(x, y)\right) \leq n, \tag{10}$$

then $(\mathcal{P}, N, M_\lambda)$ is a clustering hierarchy.

Proof: This is a straightforward exercise and is thus omitted. □

KTH Matematik

Dendrograms

Let a clustering hierarchy be given and let $J(x, y)$ be the ultrametric on $\mathcal{P}$ defined in (2) and let λ satisfy **i**) $-$ **iii**) above. We set

$$\tau(x, y) = \lambda\left(J(x, y)\right), \quad x, y \in \mathcal{P}. \tag{11}$$

$\tau(x, y)$ is called a continuous ultrametric distance. When

$$(x, y) \in M_\tau(n) \Leftrightarrow \tau\left(x, y\right) \leq n, \tag{12}$$

we express in the mathematical model the fact that clustering algorithms transform data to a hierarchy with numerical (not necessarily integer valued) levels of proximity. We set as in (1)

$$\mathbf{P}_\tau(n) = \mathcal{P} / M_\tau(n). \tag{13}$$

In (12) may happen with that there may be one or several n such that $\mathbf{P}_\tau(n) = \mathbf{P}_\tau(n+1)$. Using (13) we define an indexed dendrogram.

KTH VETENSKAP OCH KONST

KTH Matematik

Indexed Dendrograms

An *Indexed Dendrogram* (ID) is

$$\mathcal{T} = \left(\mathcal{P}, \{\mathbf{P}_\tau(n)\}_{n=0}^{N} \right),$$

where

$$N = \lceil \max_{x,y \in \mathcal{P} \times \mathcal{P}} \tau(x,y) \rceil.$$

We assume from now on that the indexed dendrogram is binary in the following sense. The indexed dendrogram $\mathcal{T}$ is a *binary indexed dendrogram*, if it holds for every $n \in \{1, \ldots, N\}$

$$\mathbf{P}_\tau(n) = (\mathbf{P}_\tau(n-1) \setminus \{A_{n-1}, B_{n-1}\}) \cup (A_{n-1} \cup B_{n-1}), \quad (14)$$

where A_{n-1} and B_{n-1} are extensions in $\mathbf{P}_\tau(n-1)$. In words, the partition $\mathbf{P}_\tau(n)$ is obtained by merging of exactly two extensions in $\mathbf{P}_\tau(n-1)$ (or of none, if (12) gives this outcome).

KTH Matematik

Indexed Dendrograms

The the intuitively useful image of a binary dendrogram coincides at first sight with that of a (rooted) binary tree.
Let us recall that a (rooted/unrooted) binary tree is a connected graph with no cycles, where each node has degree one or three. The nodes with degree three are called inner nodes and the nodes with degree one are terminal nodes. Any terminal node can in fact be taken as a root.

Indexed Dendrograms

Since the tree is supposed to represent a binary indexed dendrogram $\mathcal{T}$, this imposes certain additional properties on the tree. The inner nodes will be *ranked* by $J(x, y)$, or, on the basis of relative distance to the root. Note that an ultrametric also gives a metric on $\mathcal{P}$ by path lenghts of the tree as shown by Anette Dobson.

KTH VETENSKAP OCH KONST
KTH Matematik

Indexed Dendrograms

The terminal nodes will be *labelled* by the OTU's in $\mathcal{P}$, whereas the inner nodes are not labelled. We talk about *ranked labelled binary trees* in this sense. A more precise wording would be ranked terminally labelled binary trees. We note that a binary ranked (labelled or unlabelled) tree is such that the distance (in path length) to the root is the same for every terminal node. This property need not hold in a general binary tree.

Hierarchical Agglomerative methods of Clustering

The hierarchical agglomerative clustering methods are most commonly used. The construction of an hierarchical agglomerative classification can be achieved by the following general algorithm.

1. Find the two closest objects and merge them into a cluster
2. Find and merge the next two closest points, where a point is either an individual object or a cluster of objects.
3. If more than one cluster remains, return to step 2

KTH Matematik

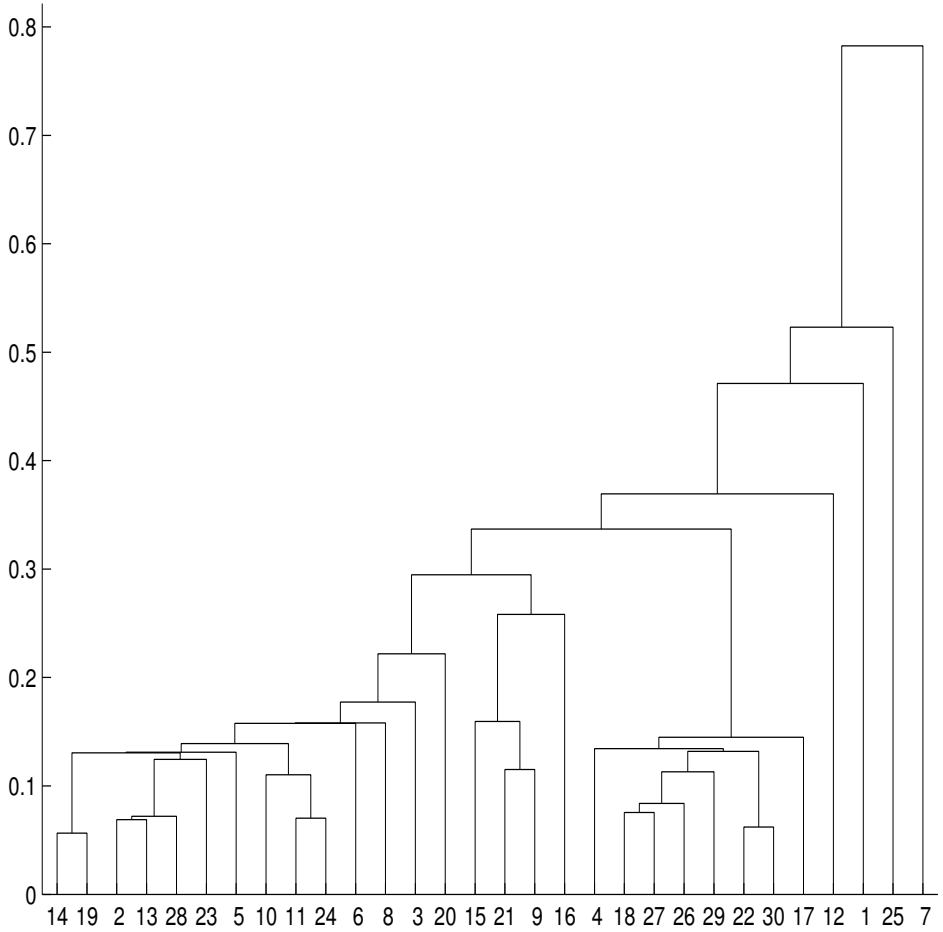
Individual hierarchic methods are characterized by the definition used for identification of the closest pair of points, and by the means used to describe the new cluster when two clusters are merged. We shall next describe some of the most well known of these methods:

- SLINK
- CLINK
- UPGMA
- WARD

KTH VETENSKAP OCH KONST

KTH Matematik

Actual clustering is done by cutting the dendrogram at some threshold value indicating the level of dissimilarity in the clusters joined at that value.



KTH Matematik

- The Single Linkage (SLINK, Nearest Neighbor)
  The single link method is probably the best known of the hierarchical methods and operates by joining, at each step, the two most similar objects, which are not yet in the same cluster. The name single link thus refers to the joining of pairs of clusters by the single shortest link between them.

KTH Matematik

Single Linkage: Formally

Dissimilarity Matrix

$$\mathcal{X}^t = \left\{ \mathbf{x}^{(l)} \right\}_{l=1}^{t}, \; d_{i,j} = D\left(\mathbf{x}^{(i)}, \mathbf{y}^{(j)} \right)$$

SLINK

1. Find
$$(i^*, j^*) = \arg\min_{i,j} d_{i,j}$$
2. Join i^*, j^*
3. $d_{i,i^* \cup j^*} = \min\left(d_{i,i^*}, d_{i,j^*}\right)$.

KTH VETENSKAP OCH KONST

KTH Matematik

- The Complete Linkage (CLINK, Farthest Neighbor)
  The complete link method is similar to the single link method except that it uses the least similar pair between two clusters to determine the inter-cluster similarity (so that every cluster member is more like the furthest member of its own cluster than the furthest item in any other cluster). This method is characterized by small, tightly bound clusters.

1. Find
$$(i^*, j^*) = \arg\max_{i,j} d_{i,j}$$
2. Join i^*, j^*
3. $d_{i,i^* \cup j^*} = \max\left(d_{i,i^*}, d_{i,j^*}\right)$.

UPGMA

- The Group Average Linkage (UPGMA= unweighted pair-group arithmetic average clustering)
  The group average method relies on the average value of the pairwise within a cluster, rather than the maximum or minimum similarity as with the single link or the complete link methods. Since all objects in a cluster contribute to the inter-cluster similarity, each object is, on average more like every other member of its own cluster then the objects in any other cluster.

KTH Matematik

WARD

Ward's minimum variance criterion minimizes the total within-cluster variance. At each step the pair of clusters with minimum between-cluster distance are merged. To implement this method, at each step find the pair of clusters that leads to minimum increase in total within-cluster variance after merging. This increase is a weighted squared distance between cluster centers. This is appropriate for Euclidean distances only. For disjoint clusters C_i, C_j, and C_k with sizes n_i, n_j, and n_k respectively:

$$D(C_i \cup C_j, C_k) = \frac{n_i + n_k}{n_i + n_j + n_k} D(C_i, C_k)$$

$$+ \frac{n_j + n_k}{n_i + n_j + n_k} D(C_j, C_k) - \frac{n_k}{n_i + n_j + n_k} D(C_i, C_j).$$

is the up-date of the distance.

To perform cluster analysis on a data set, follow this procedure:

1. Find the similarity or dissimilarity between every pair of objects in the data set. In this step, you calculate the distance between objects using the pdist function (in Matlab).
2. Group the objects into a binary, hierarchical cluster tree. In this step, you link together pairs of objects that are in close proximity using the linkage function. The linkage function uses the distance information generated in step 1 to determine the proximity of objects to each other. As objects are paired into binary clusters, the newly formed clusters are grouped into larger clusters until a hierarchical tree is formed.
3. Determine where to divide the hierarchical tree into clusters. In this step, you divide the objects in the hierarchical tree into clusters using the cluster function (in Matlab). The cluster function can create clusters by detecting natural groupings in the hierarchical tree or by cutting off the hierarchical tree at an arbitrary point.

KTH VETENSKAP OCH KONST
KTH Matematik

Fisher's Iris Data

Iris setosa

Iris versicolor

Iris virginica

KTH VETENSKAP OCH KONST

KTH Matematik

Fisher's Iris Data

Fisher's Iris data set is a multivariate data set introduced by Sir Ronald Fisher (1936) as an example of discriminant analysis. Edgar Anderson collected the data to quantify the morphologic variation of Iris flowers of three related species. Two of the three species were collected in " all from the same pasture, and picked on the same day and measured at the same time by the same person with the same apparatus".

KTH Matematik

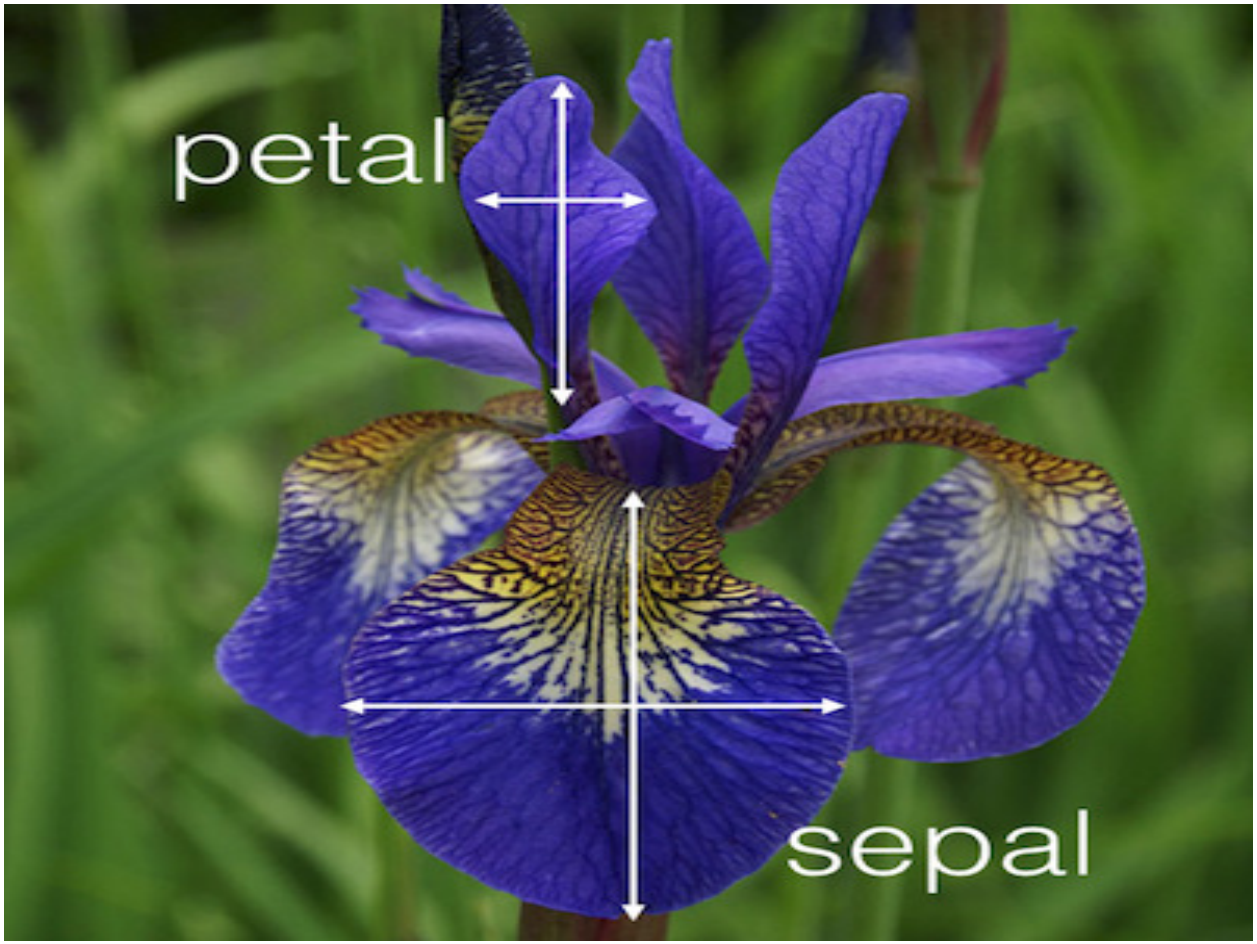
Fisher's Iris Data

The data set consists of 150 samples from each of three species of Iris (*Iris setosa*, *Iris virginica* and *Iris versicolor*). Four features were measured from each sample: the length and the width of the sepals and petals, in centimetres.

KTH VETENSKAP OCH KONST

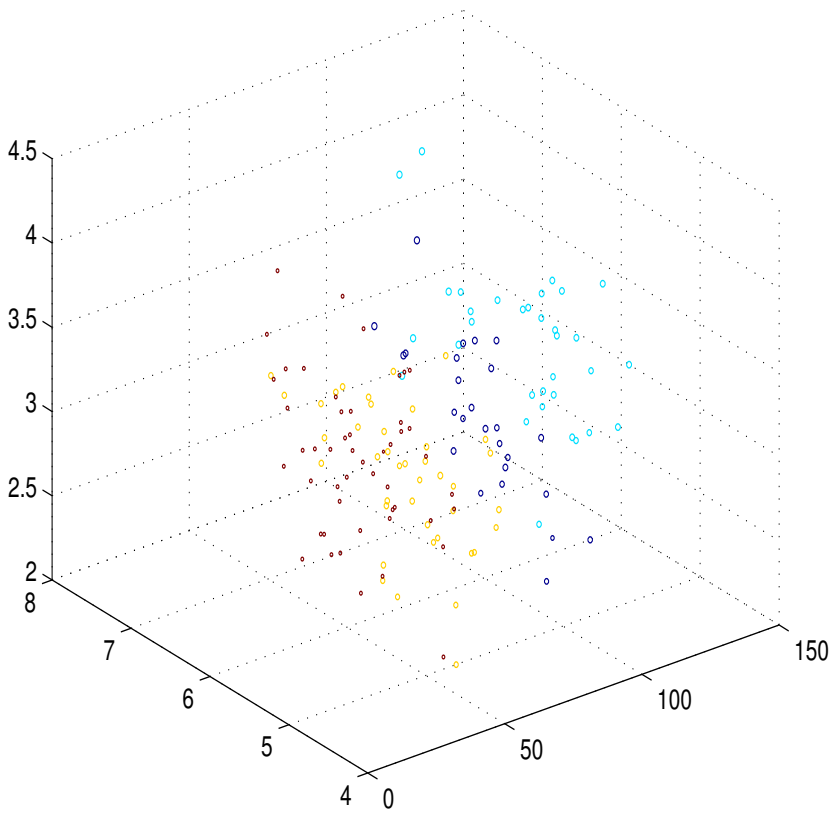
KTH Matematik

Four features were measured from each sample: the length and the width of the sepals and petals, in centimetres



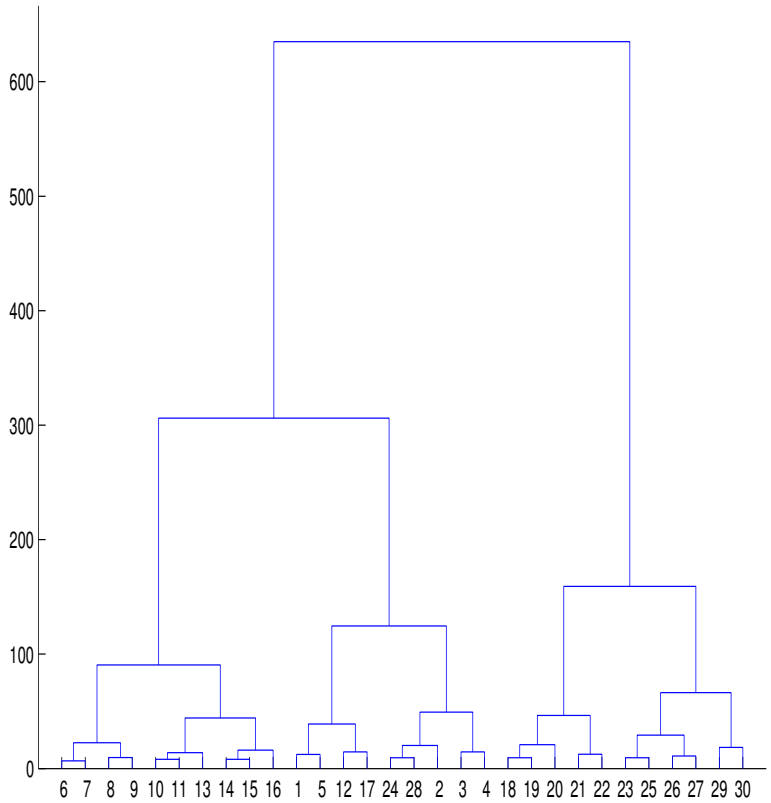
KTH Matematik

Fisher's Iris Data: 3Dscatterplot



KTH Matematik

Fisher's Iris Data: Euclidean distance, Ward Method



KTH VETENSKAP OCH KONST

KTH Matematik

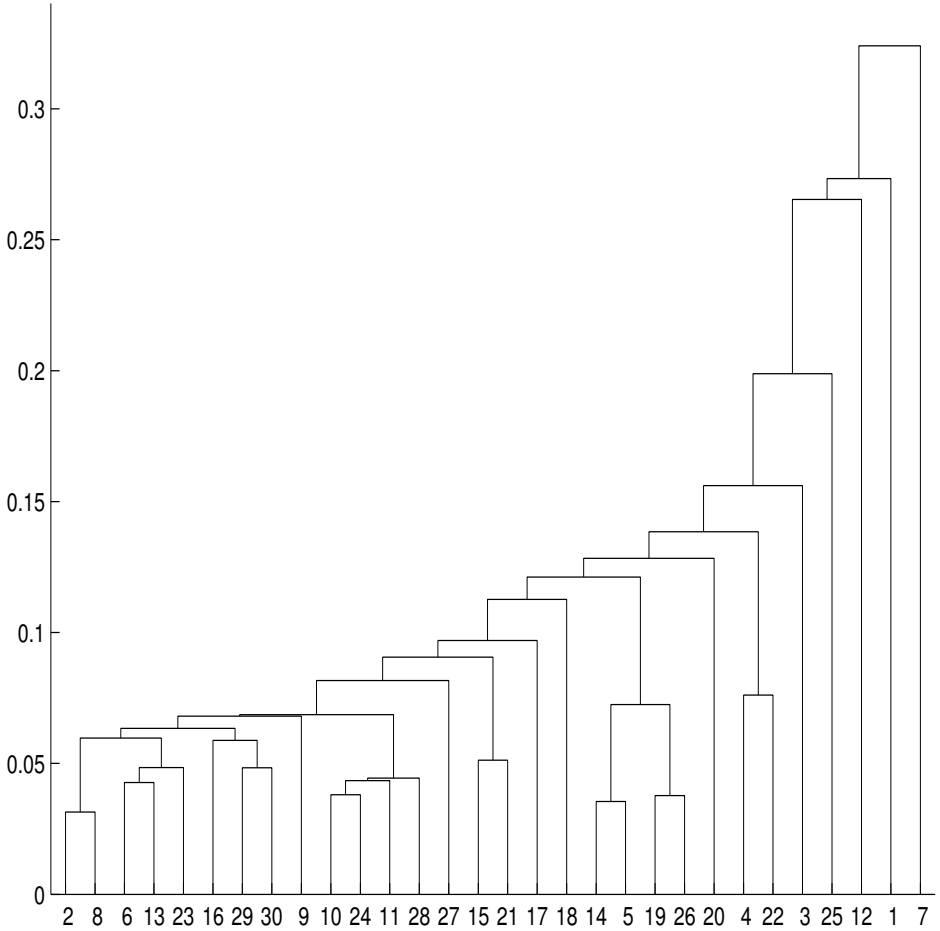
In the MATLAB^{TM} **Statistics Toolbox** these methods are used as options in
Z=linkage(Y, method)
Here Y is a distance matrix $\{d_{i,j}\}$ and 'method' is one of

- single, complete, average, centroid, ward

In the MATLAB^{TM} **Statistics Toolbox** the distances above (and some others) are computed using the function *pdist*.

KTH VETENSKAP OCH KONST
KTH Matematik

For example, are there clusters in the dendrogram in the figure below?



The **cophenet** function in MATLAB^{TM} returns the cophenetic correlation coefficient. The closer the value of the cophenetic correlation coefficient is to 1, the better the clustering solution. In the figure the cophenetic correlation coefficient equals 0.5.

Carlsson, Gunnar and Mémoli, Facundo: Characterization, stability and convergence of hierarchical clustering methods, *Journal of machine learning research*, 11, Apr, 1425–1470, 2010

Represents dendrograms as ultrametric spaces and uses tools from metric geometry, namely the Gromov-Hausdorff distance, to quantify the degree to which perturbations in the input metric space affect the result of hierarchical methods.

KTH VETENSKAP OCH KONST

KTH Matematik

Predictive Distribution for a Taxon

Consider a primary data set $X^t = \left\{x^{(l)}\right\}_{l=1}^{t}$ of t vectors, that is, each $x^{(l)}$ is a d-dimensional vector with components $x_i^{(l)}$ in some **finite discrete set** $\mathcal{X}_i$. Thus $x^{(l)} \in \mathcal{X} = \times_{i=1}^{d} \mathcal{X}_i$. Assume that we have for the vectors in X^t a partition $\{c_j\}_{j=1}^{k}$ into k classes $c_j = c_j\left(X^t\right)$. If two items in X^t are identical, then they belong to the same class c_j.

Predictive Distribution for a Taxon

Let us now assume that we have on $\mathcal{X}$ a probability distribution, or a statistical model $p(x \mid \theta_j)$, where $\theta_j \in \Theta_j$ is a real or vector valued parameter of the distribution assigned to c_j. Then we can write a joint probability for the data items in c_j as

$$\prod_{x^{(l)} \in c_j} p\left(x^{(l)} \mid \theta_j\right).$$

Predictive Distribution for a Taxon

Let us also assume that we a prior density $\phi(\theta_j)$ on Θ_j. Then we obtain by Bayes formula the posterior density

$$\Phi\left(\theta_j \mid c_j\left(X^t\right)\right) = \frac{\prod_{x^{(l)} \in c_j} p\left(x^{(l)} \mid \theta_j\right) \phi(\theta_j)}{\int_{\Theta_j} \prod_{x^{(l)} \in c_j} p\left(x^{(l)} \mid \theta_j\right) \phi(\theta_j)\, d\theta_j} \quad \text{on } \Theta_j.$$

A *predictive distribution* on $\mathcal{X}$, given c_j, is now denoted by

$$Q\left(x \mid c_j\left(X^t\right)\right) \tag{15}$$

and computed as the continuous mixture

$$Q\left(x \mid c_j\left(X^t\right)\right) = \int_{\Theta} p\left(x \mid \theta_j\right) \cdot \Phi\left(\theta_j \mid c_j\left(X^t\right)\right) d\theta_j. \tag{16}$$

Chan-Darwiche Distance

$$Q\left(x \mid c_j\left(X^t\right)\right) > 0, \qquad \text{for all } x \in \mathcal{X}. \tag{17}$$

Definition

Let $P(x)$ and $Q(x)$ be two probability distributions on a finite discrete set $\mathcal{X}$. Define

$$D_{CD}(P, Q) = \ln \max_{x \in \mathcal{X}} \frac{P(x)}{Q(x)} - \ln \min_{x \in \mathcal{X}} \frac{P(x)}{Q(x)}. \tag{18}$$

□

It can be shown that $D_{CD}(P, Q)$ is in fact a metric. It holds that $D_{CD}(P, Q) = \infty$, if P and Q have disjoint supports.

KTH VETENSKAP OCH KONST
KTH Matematik

Chan-Darwiche Distance and Statistical Clusters

We give first an assumption that yields a simple expression for $D_{CD}(P, Q)$. Assume $P(x)$ and $Q(x)$ are such that there exist $x^a \in \mathcal{X}$ and $x^b \in \mathcal{X}$ with

$$\begin{aligned} 0 < P\left(x^b\right) &\leq P(x) \leq P\left(x^a\right) \\ 0 < Q\left(x^a\right) &\leq Q(x) \leq Q\left(x^b\right) \end{aligned} \tag{19}$$

The inequalities (19) imply

$$\frac{P\left(x^b\right)}{Q\left(x^b\right)} \leq \frac{P(x)}{Q(x)} \leq \frac{P\left(x^a\right)}{Q\left(x^a\right)}.$$

From (18) we obtain

$$D_{CD}\left(P,Q\right)=\ln\frac{P\left(x^{a}\right)}{P\left(x^{b}\right)}+\ln\frac{Q\left(x^{b}\right)}{Q\left(x^{a}\right)}. \tag{20}$$

Let us consider a case, where (19) holds in natural way for two predictive distributions.

Chan-Darwiche Distance and Statistical Clusters

Suppose that we have two different empirical probability distributions on $\mathcal{X}$, or

$$\left\{ \frac{n_x^{(1)}}{n^{(1)}} \right\}_{x \in \mathcal{X}}, \quad \left\{ \frac{n_x^{(2)}}{n^{(2)}} \right\}_{x \in \mathcal{X}}, \tag{21}$$

where $n_x^{(i)}$, $i = 1, 2$, is the observed number of values x corresponding to two (presumed) clusters, respectively, and $n^{(i)}$, $i = 1, 2$, are the total number of cases in the two clusters. We take the corresponding predictive distributions as given by

$$P(x) = \frac{n_x^{(1)} + \alpha q_i}{n^{(1)} + \alpha}, x \in \mathcal{X} \quad Q(x) = \frac{n_x^{(2)} + \alpha q_i}{n^{(2)} + \alpha}, x \in \mathcal{X}, \tag{22}$$

where the *hyperparameters* are $\alpha > 0$, $q_x \geq 0$, $\sum_{i=1} q_i = 1$. We obtain (22) as a predictive distribution using (16) via a Dirichlet prior on the probability distributions modeling the empirical distributions in (21).

Chan-Darwiche Distance and Statistical Clusters

Let us now assume that there exist $x^a \in \mathcal{X}$ and $x^b \in \mathcal{X}$ such that

$$n_{x^a}^{(1)} \geq n_x^{(1)}, \qquad \text{for all } x \in \mathcal{X},$$

and

$$n_{x^b}^{(2)} \geq n_x^{(2)}, \qquad \text{for all } x \in \mathcal{X},$$

and that there is a connected domain $\mathcal{X}_1$ containing x^a such that

$$n_x^{(2)} = 0, \qquad \text{for all } x \in \mathcal{X}_1$$

and that there is a connected domain $\mathcal{X}_2$ containing x^b such that

$$n_x^{(1)} = 0, \qquad \text{for all } x \in \mathcal{X}_2$$

KTH VETENSKAP OCH KONST

KTH Matematik

Chan-Darwiche Distance and Statistical Clusters

This reflects a statistical interpretation of the notion of a cluster, which has been described as a connected region of high density in $\mathcal{X}$. The high density regions in the empirical clusters are well separated, too.

KTH VETENSKAP OCH KONST
KTH Matematik

Let us also assume that, for simplicity, that $q_i = 1/M$, where $M = |\mathcal{X}| =$(carinality of $\mathcal{X}$. Then we clearly get from (22) that

$$
\begin{aligned}
0 < \frac{\alpha/M}{n^{(1)}+\alpha} \quad &\leq P(x) \leq \quad \frac{n^{(1)}_{x^a}+\alpha/M}{n^{(1)}+\alpha} \\
0 < \frac{\alpha/M}{n^{(2)}+\alpha} \quad &\leq Q(x) \leq \quad \frac{n^{(2)}_{x^b}+\alpha/M}{n^{(2)}+\alpha}
\end{aligned}
\tag{23}
$$

The two distributions $P(x)$ and $Q(x)$ represent distinct or well separated clusters, yet not permitting statistical discrimination without error.

Chan-Darwiche Distance

$$D_{CD}\left(P,Q\right)=\ln\frac{n_{x^a}^{(1)}+\alpha/M}{\alpha/M}+\ln\frac{n_{x^b}^{(2)}+\alpha/M}{\alpha/M}. \tag{24}$$

or

$$D_{CD}\left(P,Q\right)=\ln\left(n_{x^a}^{(1)}+\alpha/M\right)+\ln\left(n_{x^b}^{(2)}+\alpha/M\right)+2\ln M-2\ln\alpha. \tag{25}$$

Clearly, we can expect this to be a large number for large sets of data, or, the CD -distance assumes a large value for well separated clusters represented by the respective predictive distributions.

Chan-Darwiche Distance

Assume that we have third sample of $n^{(3)}$ data points with frequencies $n_x^{(3)}$ on $x \in \mathcal{X}$, well separated from the two other samples. Let us thus assume that

$$n_{x^c}^{(3)} \geq n_x^{(3)}, \quad \text{for all } x \in \mathcal{X},$$

and that there is a connected domain $\mathcal{X}_3$ containing x^c such that

$$n_x^{(2)} = n_x^{(1)} = 0, \quad \text{for all } x \in \mathcal{X}_3$$

and that

$$n_x^{(3)} = 0, \quad \text{for all } x \in \mathcal{X}_1 \cup \mathcal{X}_2.$$

Then, if

$$R(x) = \frac{n_x^{(3)} + \alpha q_x}{n^{(3)} + \alpha}, x \in \mathcal{X},$$

we can compute $D_{CD}(P, R)$ and $D_{CD}(Q, R)$ as in (25) with obvious modifications. Let us suppose that

$$n_{x^c}^{(3)} = n_{x^b}^{(2)} < n_{x^a}^{(1)}.$$

Dendrogram & Chan-Darwiche Distance

Then it follows that

$$D_{CD}(R, Q) < \max\left(D_{CD}(P, Q), D_{CD}(P, R)\right). \qquad (26)$$

Thus, if there are lot of triplets of predictive distributions P, R, Q for a set of data with (26), then the data underlying the empirical distributions is well classifiable in the sense of being representable by indexed dendrograms based on D_{CD}.

Dendrogram Chan-Darwiche Distance

Or, separatedness of clusters, which is supported by spatial sparsity of data, is the way predictive empirical distributions on a discrete space $\mathcal{X}$ tend to be ultrametric.
Compute for all $k(k-1)/2$ distinct pairs c_j, c_i the corresponding metrics

$$D_{ij} = D\left(c_i\left(X^t\right), c_j\left(X^t\right)\right) = D_{CD}\left(Q_{c_i(X^t)}, Q_{c_j(X^t)}\right). \qquad (27)$$

In view of (17) it holds that $0 \leq D_{ij} < \infty$.

Next we can define a hierarchic clustering method that is a modified version of the single linkage method. We write $\{c_j(X^t)\}_{j=1}^k$ as $\left\{c_j^{\gamma_k}(X^t)\right\}_{j=1}^k$. In this notation $\left\{c_j^{\gamma_t}(X^t)\right\}_{j=1}^t$ is the partitioning of X^t, where each individual item $x^{(l)}$ is a cluster and $\left\{c_1^{\gamma_1}(X^t)\right\}_{j=1}^1$ means that X^t is the one and only cluster.

1) For $k = 1, \ldots, t-1$
2) Compute

$$Q_{c_j^{\gamma_{t-k}}(X^t)} = Q\left(x \mid c_j^{\gamma_{t-k}}(X^t)\right)$$

for $j = 1, \ldots, k$ using (16). Compute

$$D_{ij} = D_{CD}\left(Q_{c_i^{\gamma_{t-k}}(X^t)}, Q_{c_j^{\gamma_{t-k}}(X^t)}\right).$$

KTH Matematik

3) Merge those two classes $c_j^{\gamma_{t-k}}\left(X^t\right)$ and $c_i^{\gamma_{t-k}}\left(X^t\right)$ with the smallest D_{ij}, ties being resolved arbitrarily. With $(r,s) = \arg\max_{(i,j)} D_{ij}$ the cluster in the next level of hierarchy will be

$$c^{\gamma_{t-k+1}}\left(X^t\right) = c_r^{\gamma_{t-k}}\left(X^t\right) \cup c_s^{\gamma_{t-k}}\left(X^t\right).$$

The new partition is $\left\{c_j^{\gamma_{t-k+1}}\left(X^t\right)\right\}_{j=1}^{t-k+1}$

4) Set $k+1 \mapsto k$. Go to **2)**.

Bayesian Hierarchic Sample Clustering

We set $N = t - 1$, $0 \leq n \leq N$, and then we have

- $\mathbf{P}(0) = \{\{x^{(1)}\}, \ldots, \{x^{(t)}\}\}$.
- $\mathbf{P}(n) = \{c_1^{\gamma_{t-n}}(X^t), \ldots, c_{N-n}^{\gamma_{t-n}}(X^t)\}$.
- $\mathbf{P}(N) = \{c_1^{\gamma_1}(X^t)\} = \{X^t\}$

A taxon Y in the hierarchy or the dendrogram is thus of the form $Y = \left(c_j^{\gamma_{t-n}}(X^t), n\right)$.

Bayesian Hierarchic Sample Clustering

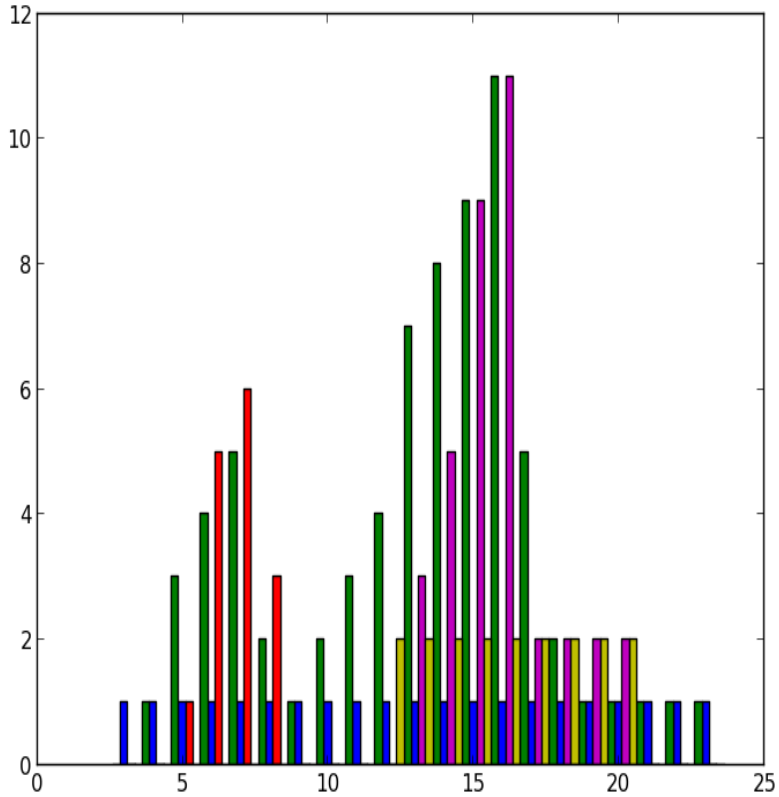
Thus we have defined a predictive distribution

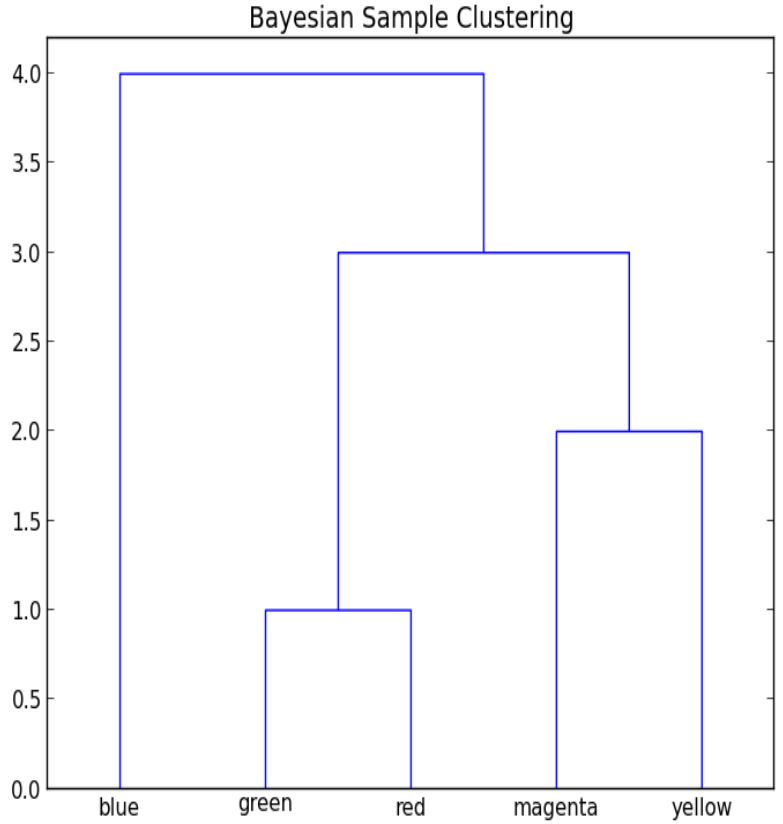
$$Q(x \mid \text{taxon}X) = Q(x \mid \text{Ext}(X), \text{Rank}(X)), \quad x \in \mathcal{X}$$

for any taxon X in the taxonomic hierarchy defined above.

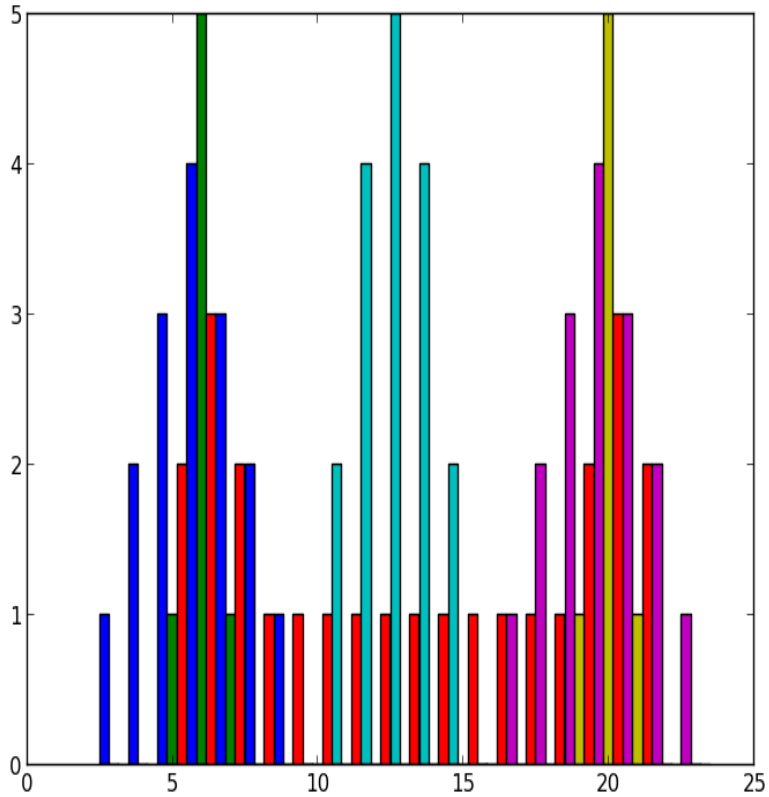
KTH Matematik

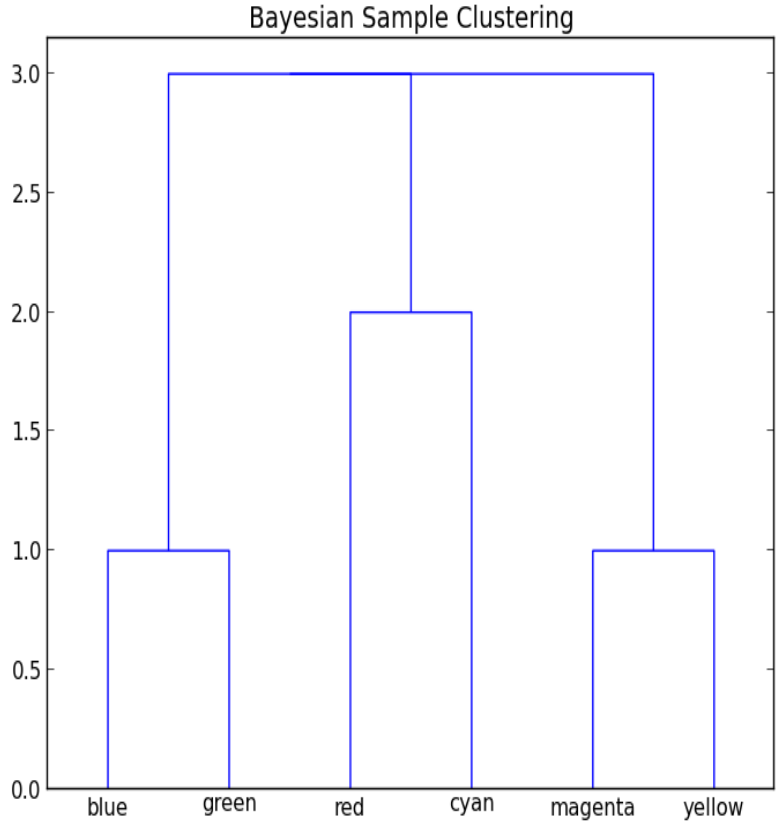
Simulations (Gabriel Isheden)





Simulations (Gabriel Isheden)

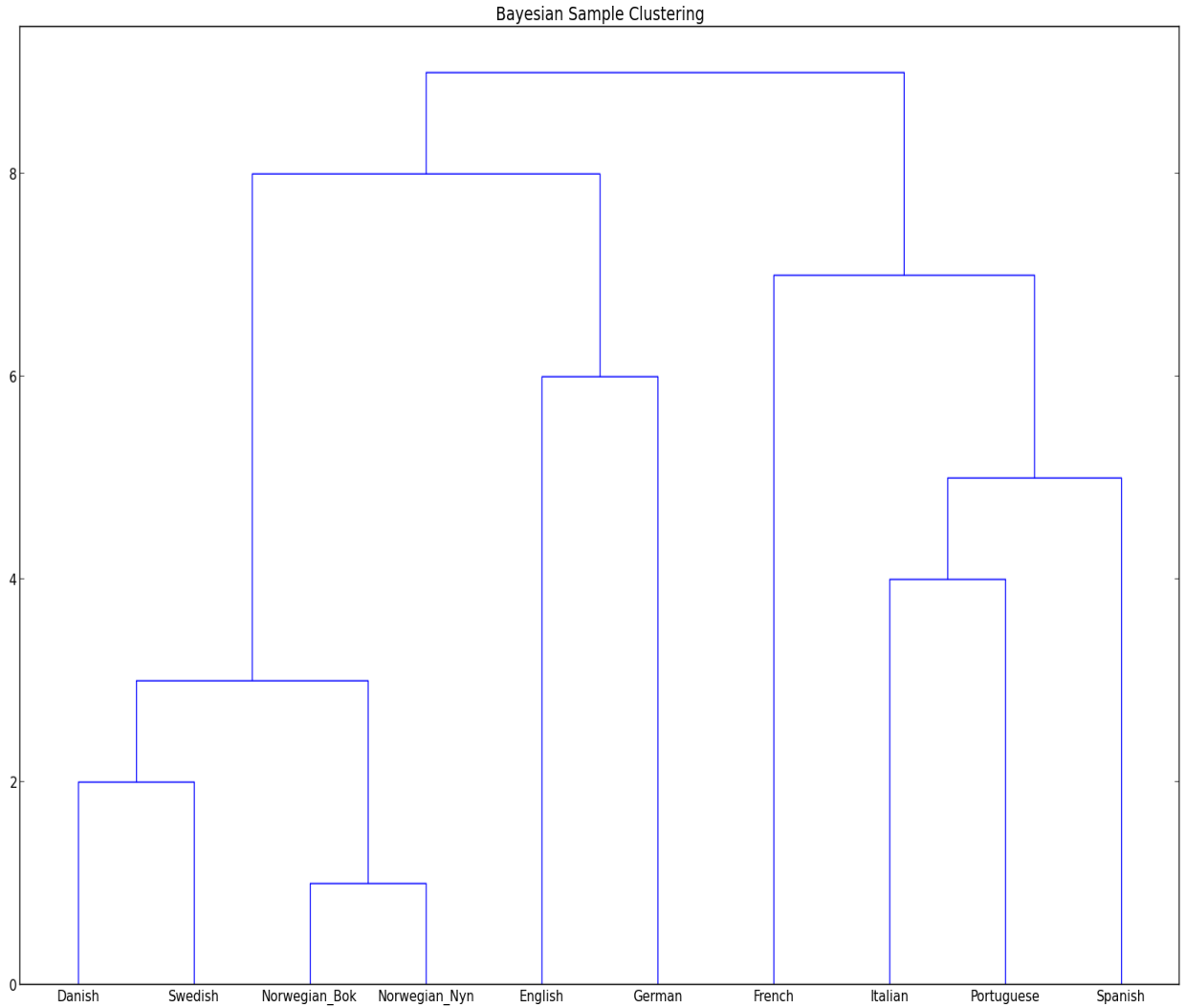




KTH VETENSKAP OCH KONST

KTH Matematik

Dendrogram over Some European Languages (Isheden)



KTH Matematik

Software (Gabriel Isheden)

https://github.com/Skjulet/Bayesian_Sample_Clustering/blob/master/Bayesian_Sample_Clustering.py

KTH Matematik