Modern Methods of Statistical Learning sf2935
Lecture 13: Unsupervised Learning 1.
Timo Koski

TK

28.11.2017

KTH Matematik

Clustering

This part of the course deals mainly with tools in *clustering* and *model based clustering*. Modelling tools for classification are treated in two separate lectures: The first of lectures deal with *summarizing & descriptive methods* and the second part treats *hierarchic* andpredictive methods.

KTH VETENSKAP OCH KONST

KTH Matematik

Clustering 1: Descriptive methods
OUTLINE OF CONTENTS:

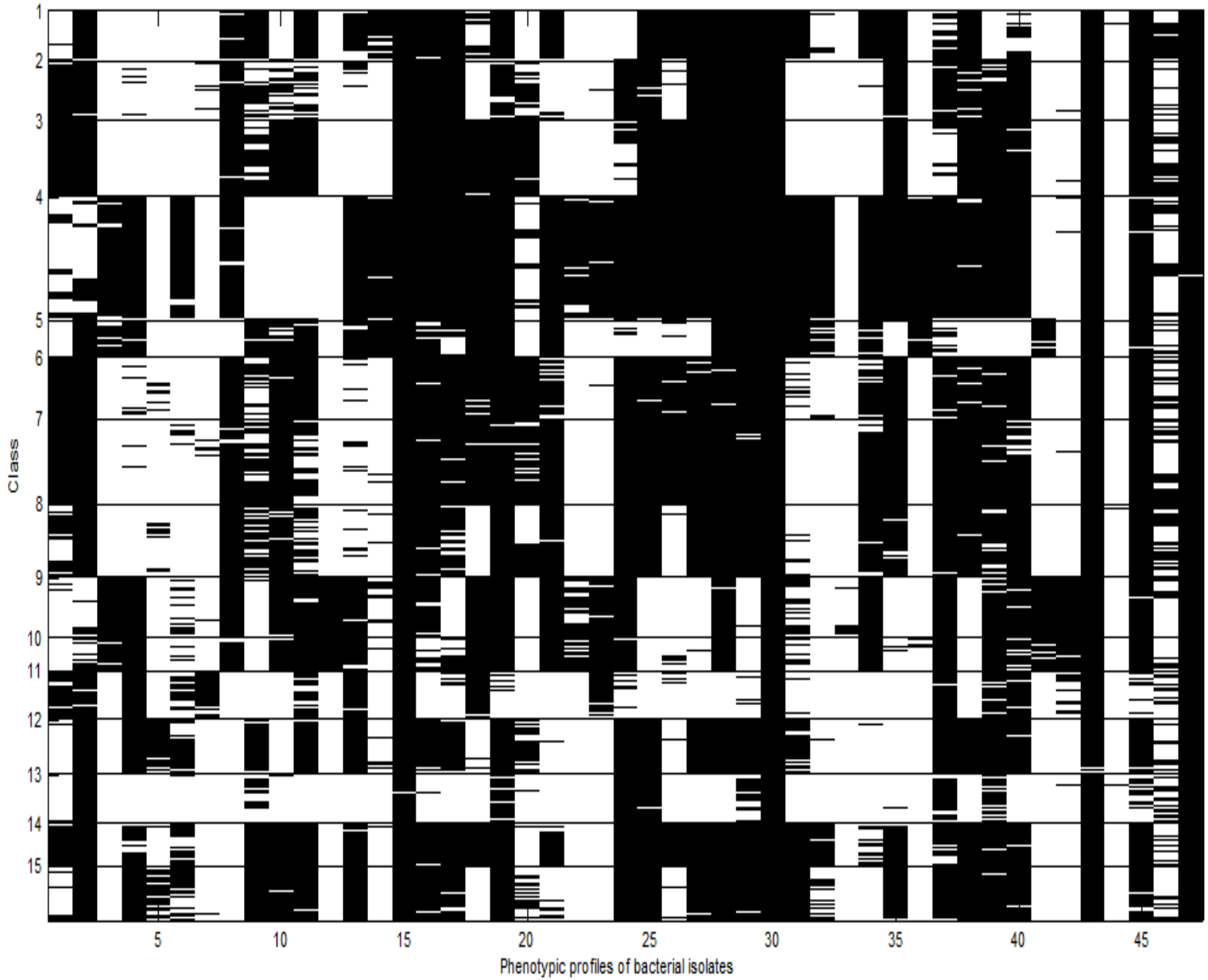
- Clustering techniques: an introduction
- The principles of k-means clustering
- Distances in clustering
- Gaussian mixtures
- The EM-algorithm
- The CEM -algorithm
- Learning vector quantization
- Validity of Clustering

KTH VETENSKAP OCH KONST
KTH Matematik

Clustering 1: Learning outcomes

- k-means clustering, Voronoi regions
- Distances in clustering
- Gaussian mixtures modelling for clustering
- The EM-algorithm
- The CEM -algorithm
- Learning vector quantization
- Validity of Clustering

KTH VETENSKAP OCH KONST
KTH Matematik


Class
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
5
10
15
20
25
30
35
40
45
Phenotypic profiles of bacterial isolates

Clustering Techniques

'Clusterings and classifications are important ways of representing information about the similarities and dissimilarities of observations, and as such they are a useful form of background knowledge. It is generally believed that taking such background information into account is necessary for useful statistics. Background information in the form of clusterings or classifications is used for example to improve the quality of discovered rules .... thus validating the correctness of that background information is essential.‘

KTH VETENSKAP OCH KONST

KTH Matematik

McKinsey Global Institute Report Big data: The next frontier for innovation, competition, and productivity. Ch. Big data techniques and technologies

Cluster analysis. A statistical method for classifying objects that splits a diverse group into smaller groups of similar objects, whose characteristics of similarity are not known in advance. An example of cluster analysis is segmenting consumers into self-similar groups for targeted marketing. This is a type of unsupervised learning because training data are not used. This technique is in contrast to classification, a type of supervised learning. Used for **data mining**.

KTH Matematik

McKinsey Global Institute Report Big data: The next frontier for innovation, competition, and productivity

Data mining. A set of techniques to extract patterns from large datasets by combining methods from statistics and machine learning with database management. These techniques include association rule learning, cluster analysis, classification, and regression. Applications include mining customer data to determine segments most likely to respond to an offer, mining human

KTH VETENSKAP OCH KONST
KTH Matematik

Introduction to Clustering (1)

One goal of clustering is to discover structure in the data as a whole. Clustering techniques are used for combining observed examples into clusters (groups) which satisfy two main criteria:

- each group or cluster is homogeneous; examples that belong to the same group are similar to each other.
- each group or cluster should be different from other clusters, that is, examples that belong to one cluster should be different from the examples of other clusters.

KTH VETENSKAP OCH KONST
KTH Matematik

Clustering as Explorative Data Analysis

Statistical methods are concerned with combining information from different observational units, and with making inferences from the resulting summaries to prospective measurements on the same or other units.
These operations will be useful only when the units are judged to be *similar* (comparable or homogeneous)[1]

[1] D. Draper, J.S. Hodges, C.L. Mallows & D. Pregibon: Exchangeability and Data Analysis, *Journal of the Royal Statistical Society*, A, 156, pp. 9−37, 1993.

Introduction to Clustering (2)

Clusters can be expressed in different ways:

- identified clusters may be exclusive, so that any example belongs to only one cluster.
- they may be overlapping; an example may belong to several clusters.
- they may be probabilistic, whereby an example belongs to each cluster with a certain probability.
- clusters might have hierarchical structure, having crude division of examples at highest level of hierarchy, which is then refined to sub-clusters at lower levels.

KTH VETENSKAP OCH KONST
KTH Matematik

Introduction to Clustering (3)

Clustering methods can be hierarchic or partitioning. We will first explain here the basics of the simplest (and most fundamental) of the partitioning clustering methods: **k-means algorithm**, which seems a best choice for the illustration of the main principles. After this we will go over to a generalized partitioning method: mainly those clustering models, which are probabilistic versions of the k-means algorithm.

KTH VETENSKAP OCH KONST

KTH Matematik

(A) k-means algorithm :

This algorithm has as an input a predefined number of clusters, that is the k in its name. Means stands for an average, an average location of all the members of a particular cluster.
The value of each attribute of an example represents a distance of the example from the origin along the attribute axes. Of course, in order to use this geometry efficiently, the values in the data set must all be numeric and should be normalized in order to allow fair computation of the overall distances in a multi-attribute space.
In these lectures the multi-attribute space is R^d and the examples are (with the exception in discussion of Jackard's distance) d-dimensional vectors $\mathbf{x}$ in R^d, also known as **feature vectors**.

KTH VETENSKAP OCH KONST
KTH Matematik

k-means algorithm: the centroid

k-means algorithm is a simple, iterative procedure, in which a crucial concept is the one of centroid. **Centroid** is an artificial point in the space of items which represents an average location of the particular cluster. The coordinates of this point are averages of attribute values of all examples that belong to the cluster.

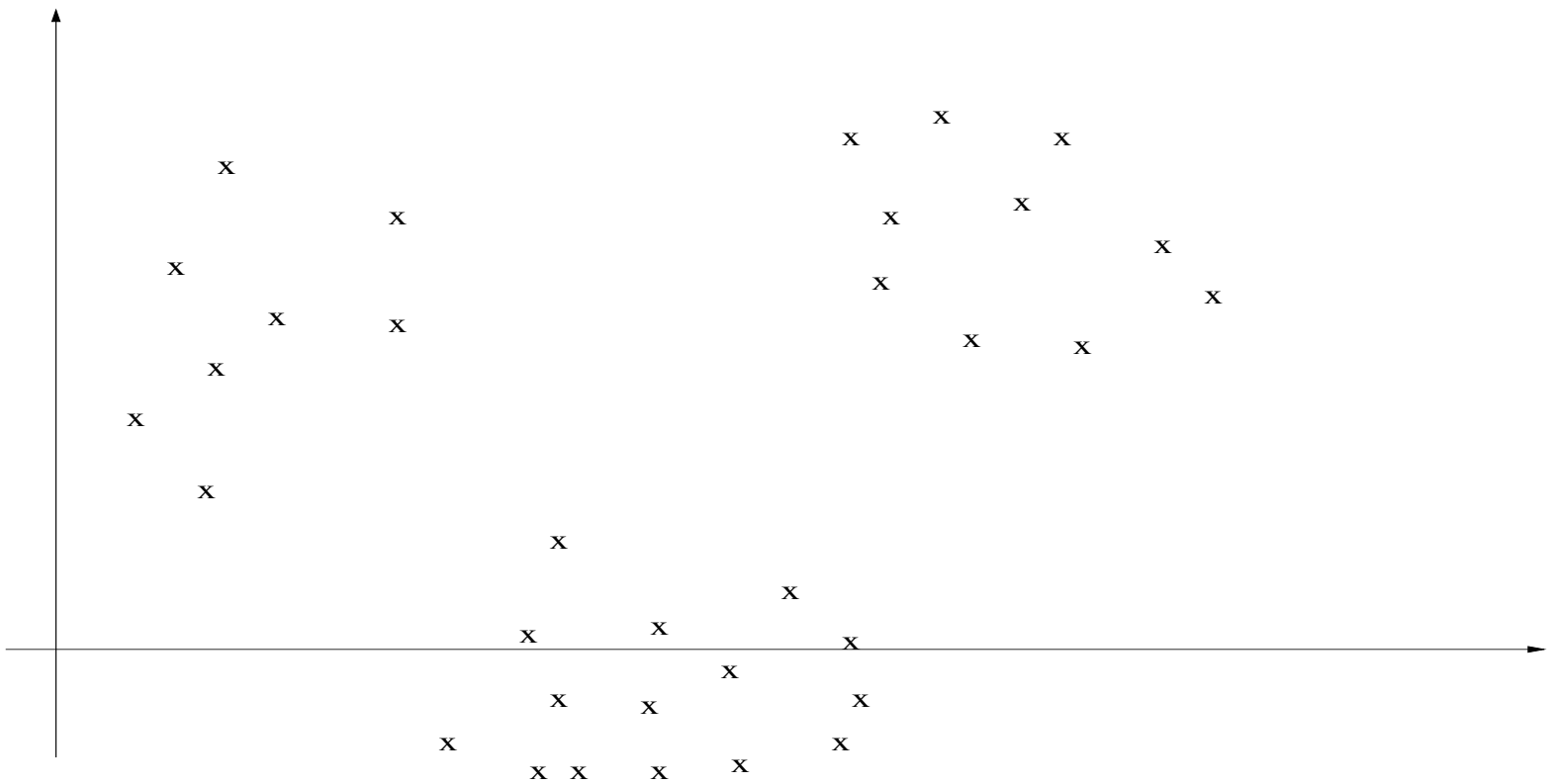
KTH VETENSKAP OCH KONST
KTH Matematik

The steps of the k-means algorithm are:

1. Select randomly k points (it can be also examples) to be the seeds for the centroids of k clusters.
2. Assign each example to the centroid closest to the example, forming in this way k exclusive clusters of examples.
3. Calculate new centroids of the clusters. For that purpose average all attribute values of the examples belonging to the same cluster (centroid).
4. Check if the cluster centroids have changed their "coordinates". If yes, start again from the step 2. . If not, cluster detection is finished and all examples have their cluster memberships defined.

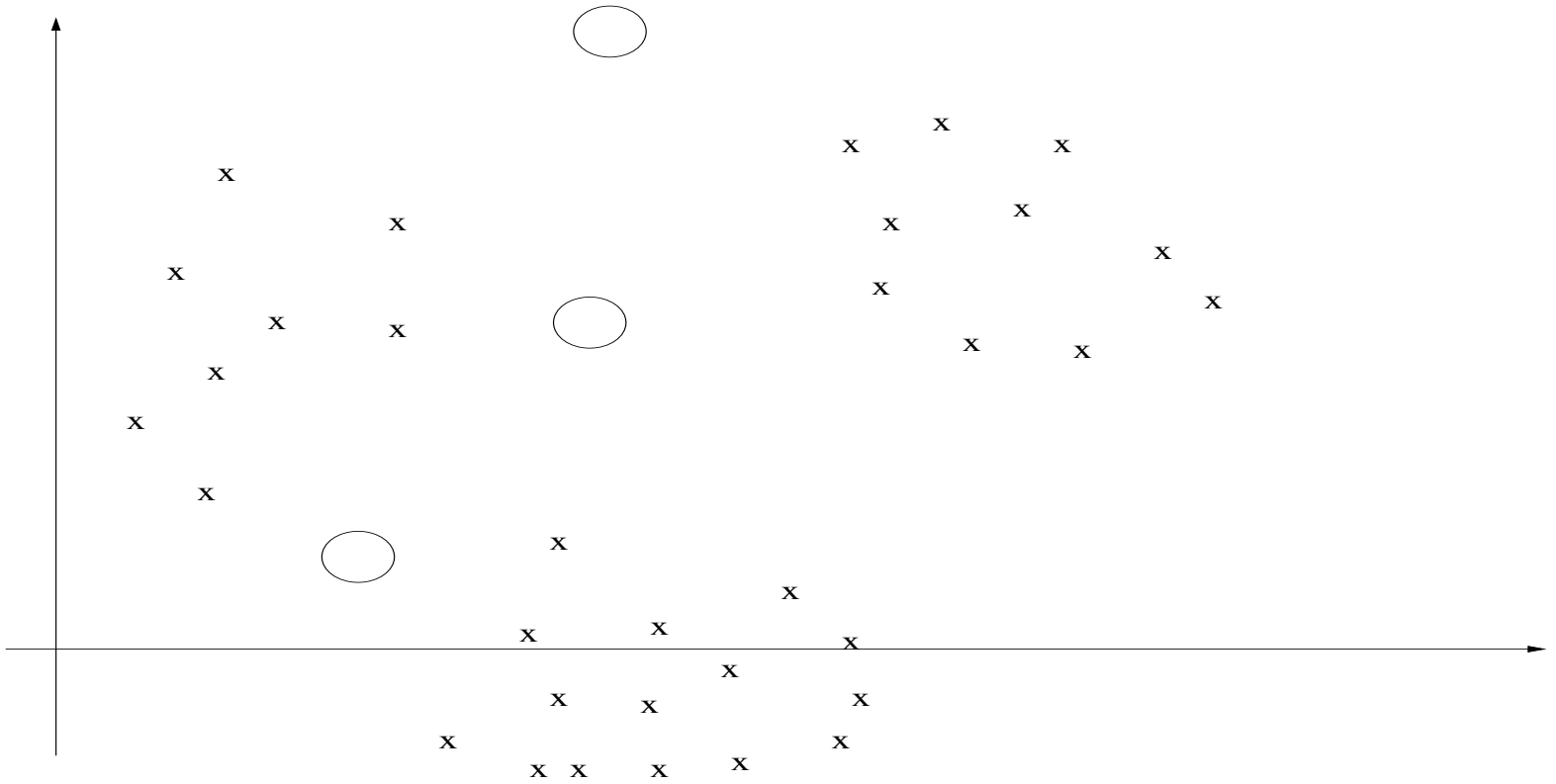
KTH VETENSKAP OCH KONST
KTH Matematik

k-means clustering: the data

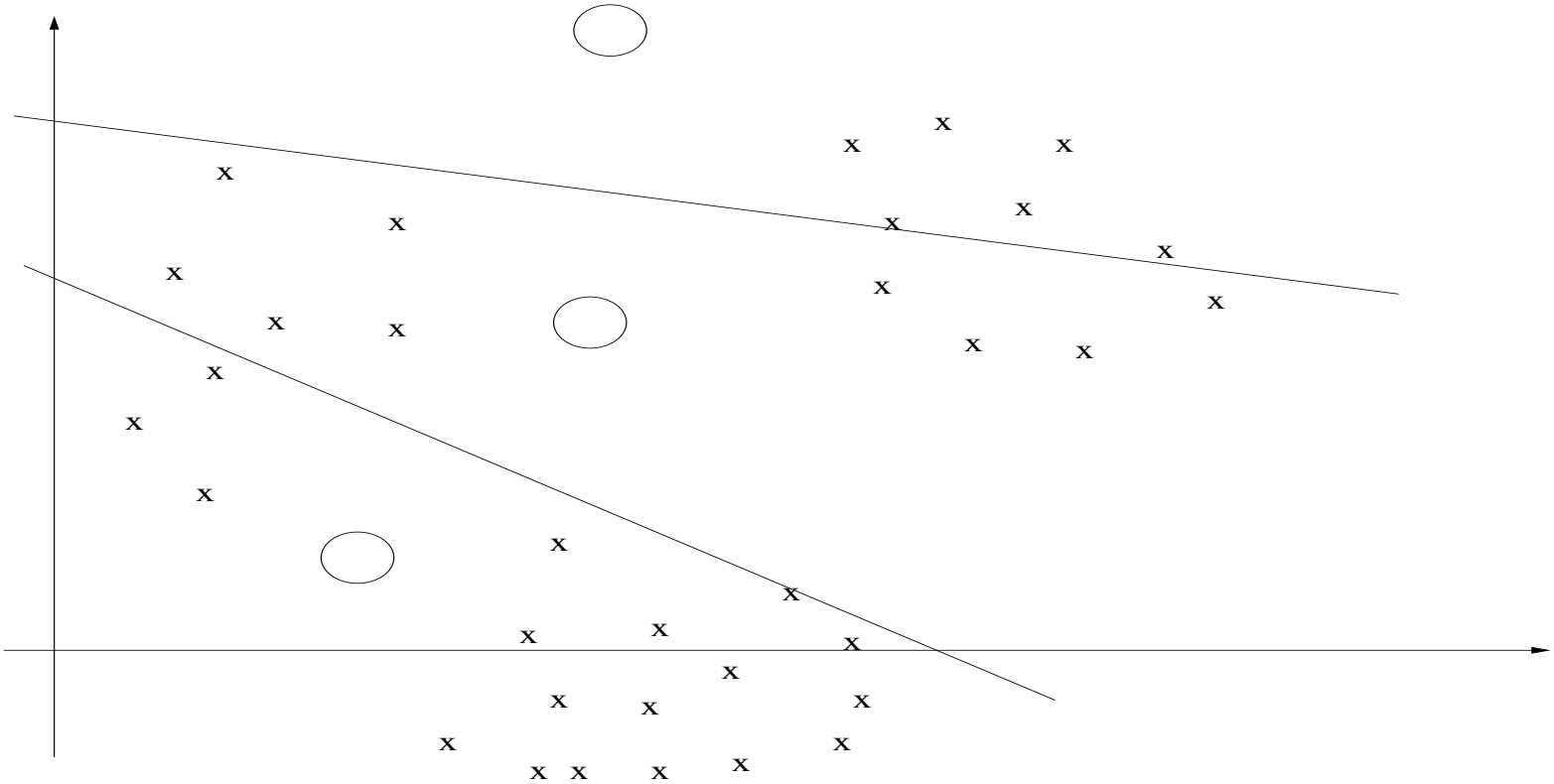


k-means clustering with $k = 3$

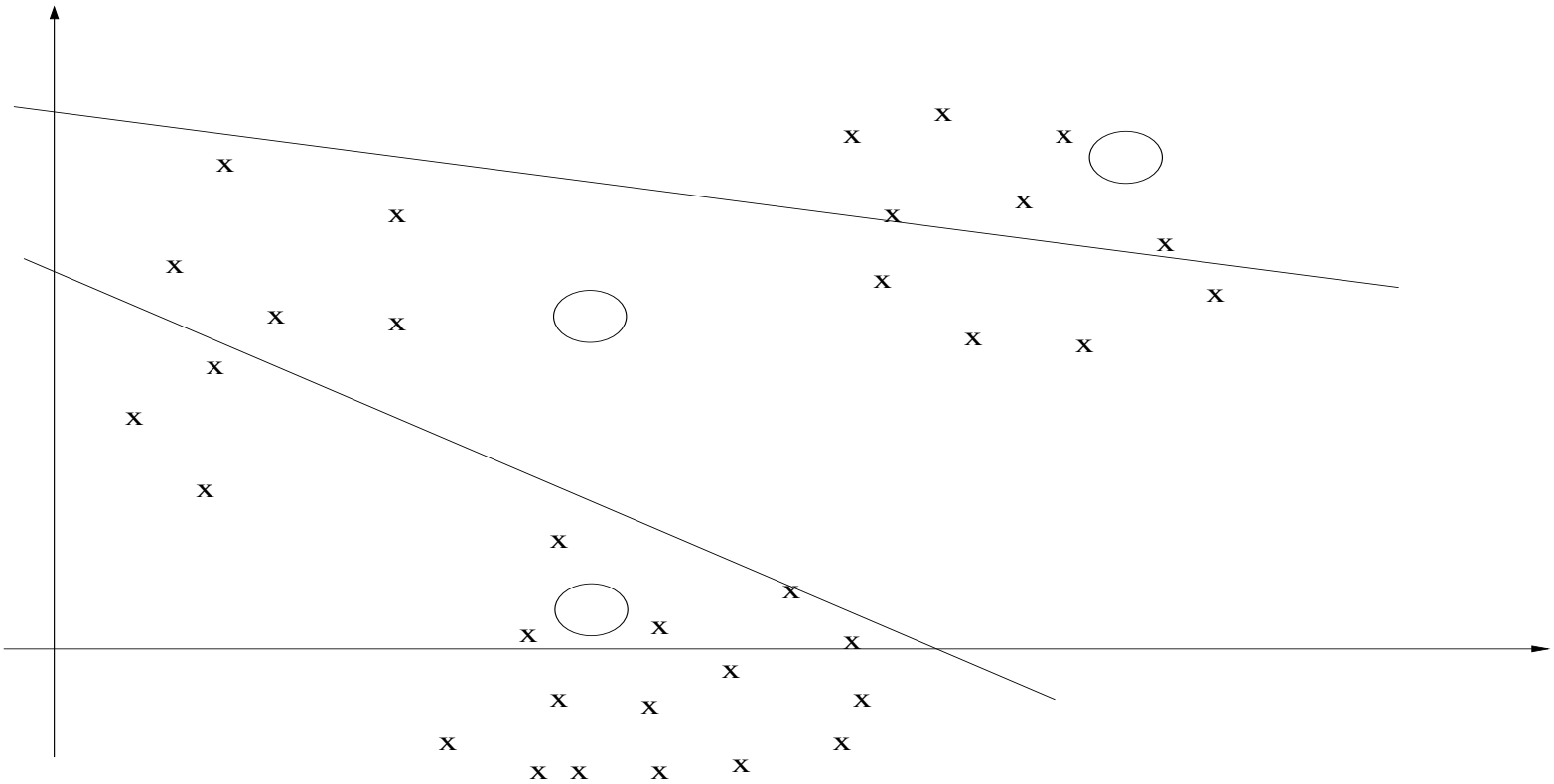
The randomly chosen centroids



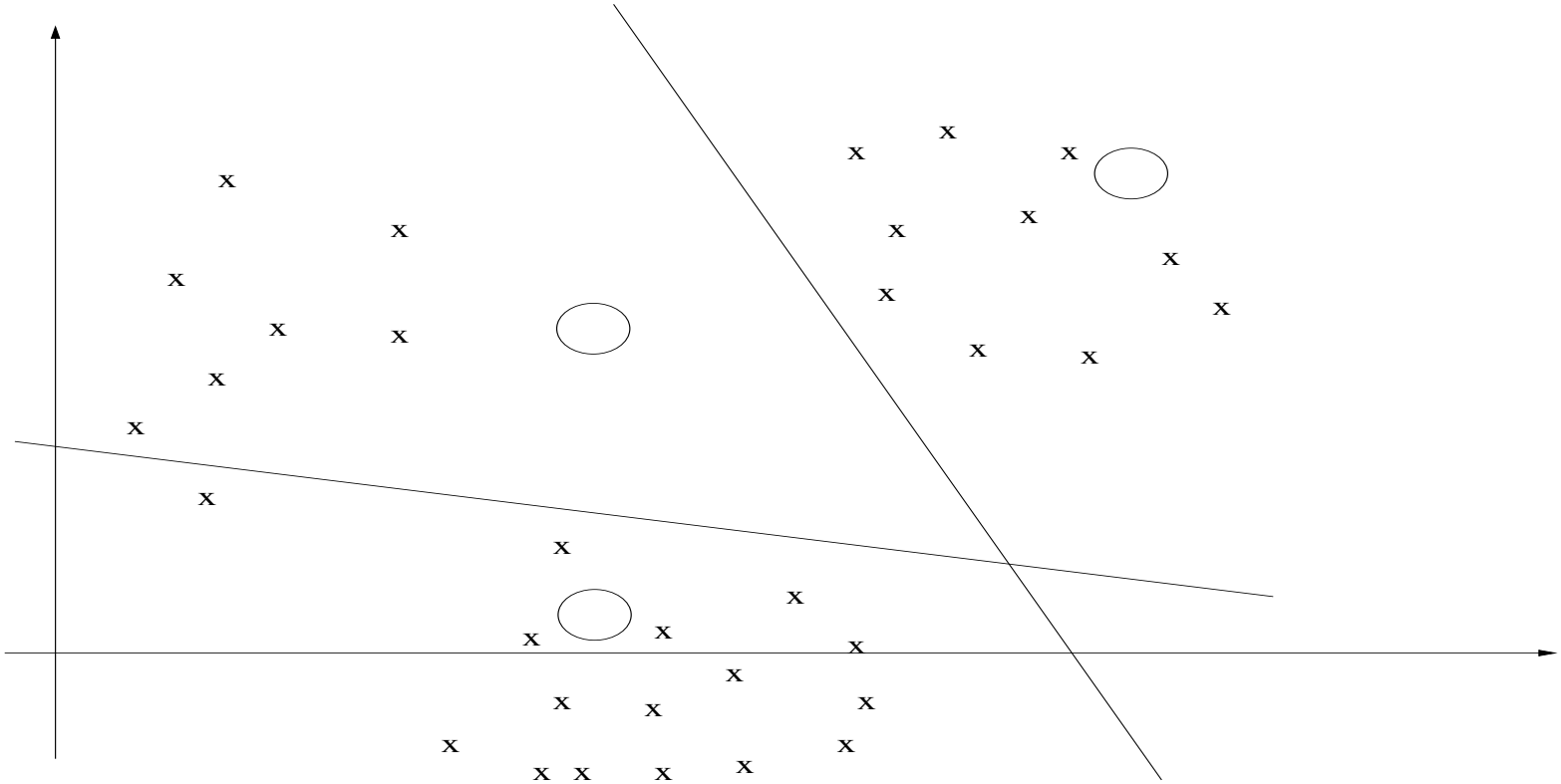
k-means clustering: Iteration 1



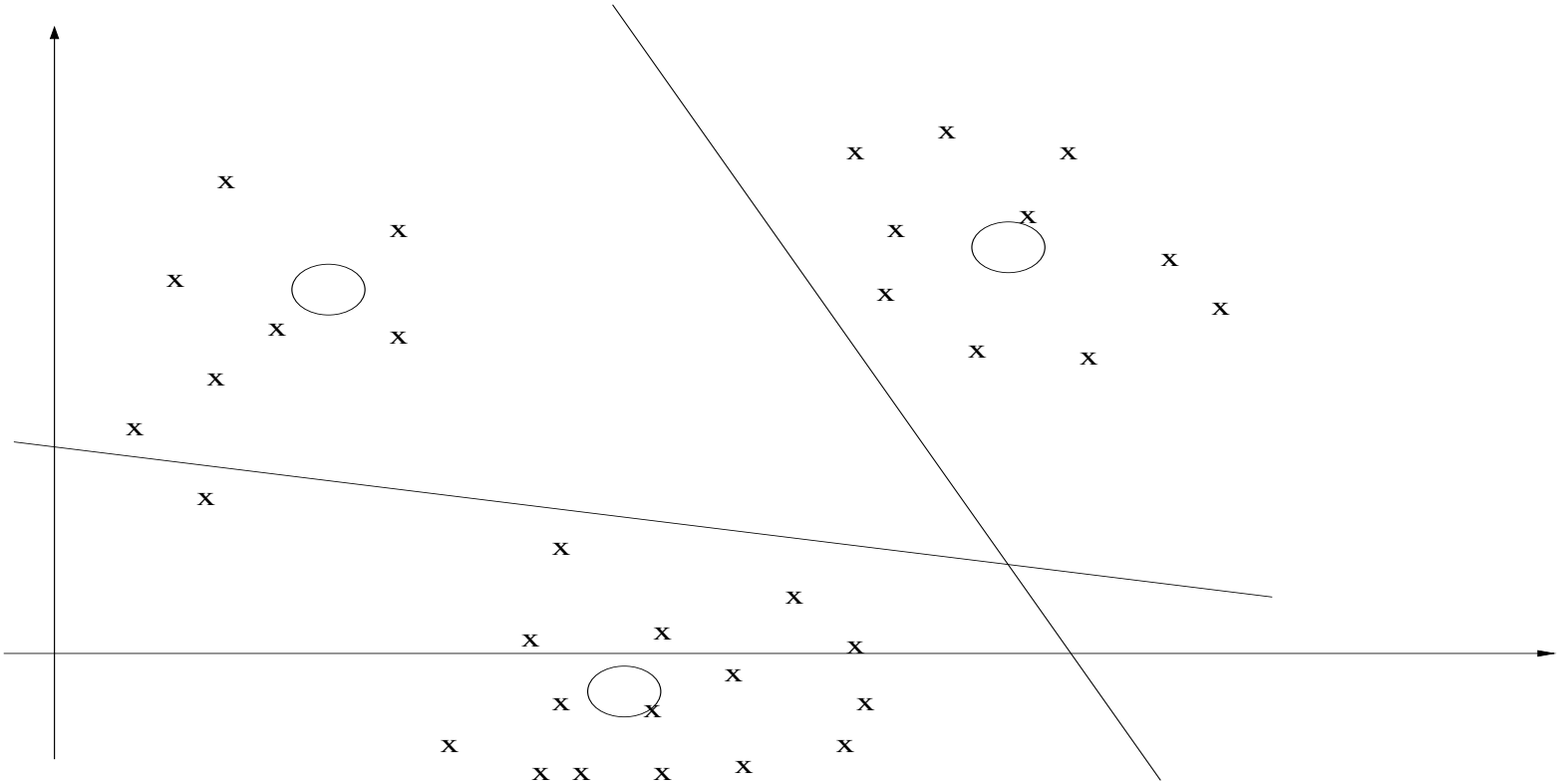
k-means clustering



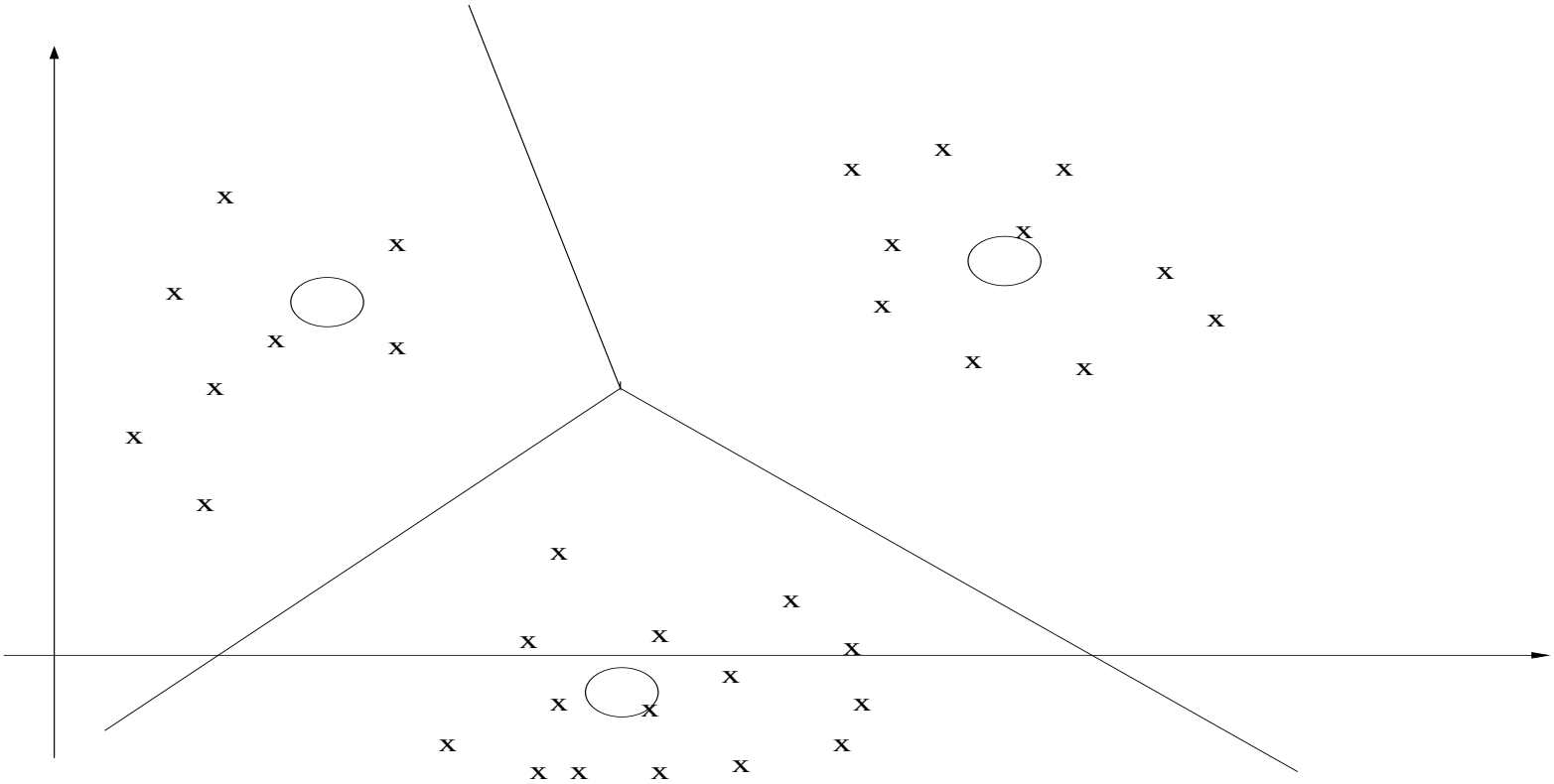
k-means clustering: Iteration 2



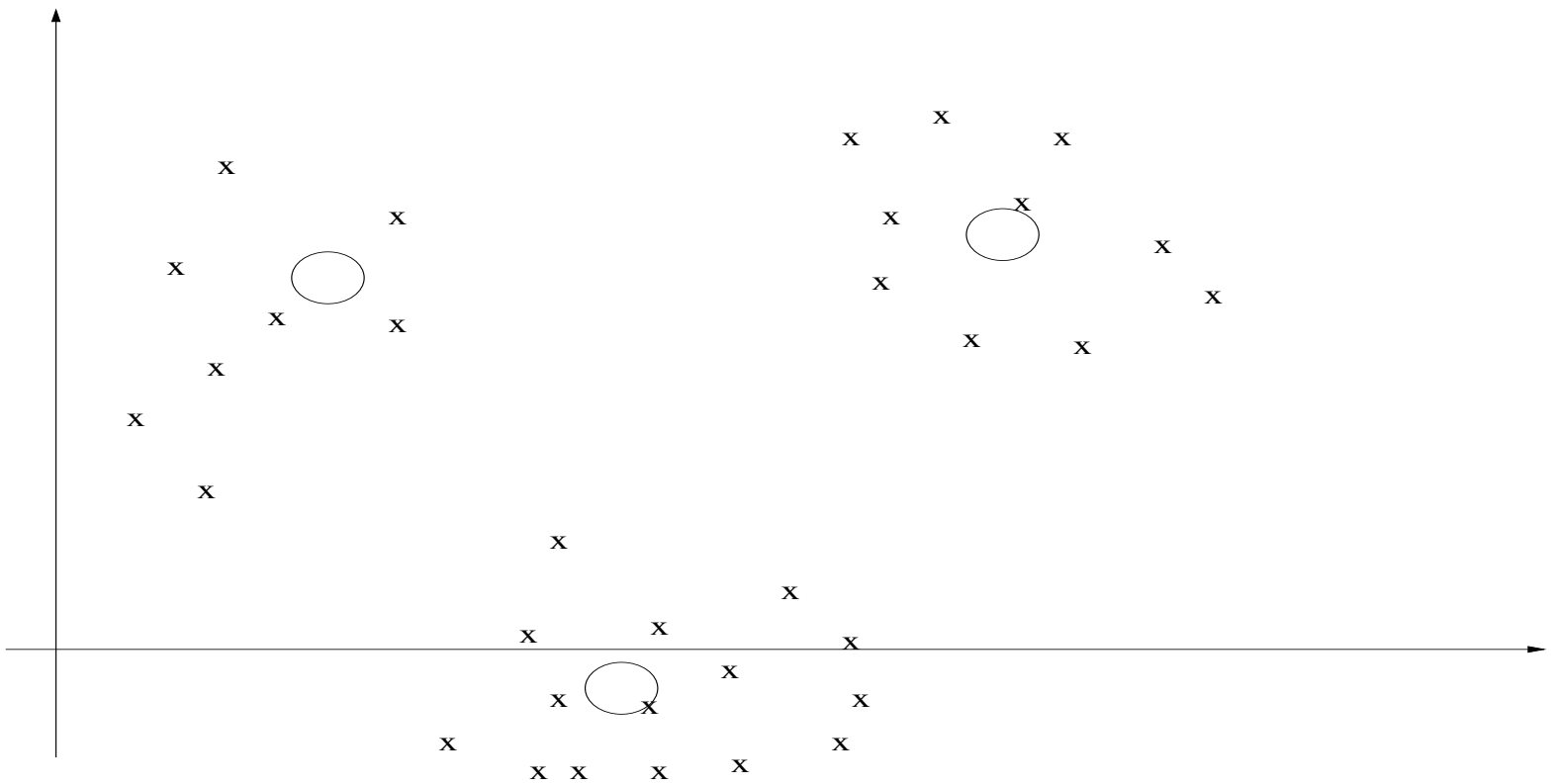
k-means clustering



k-means clustering: Iteration 3

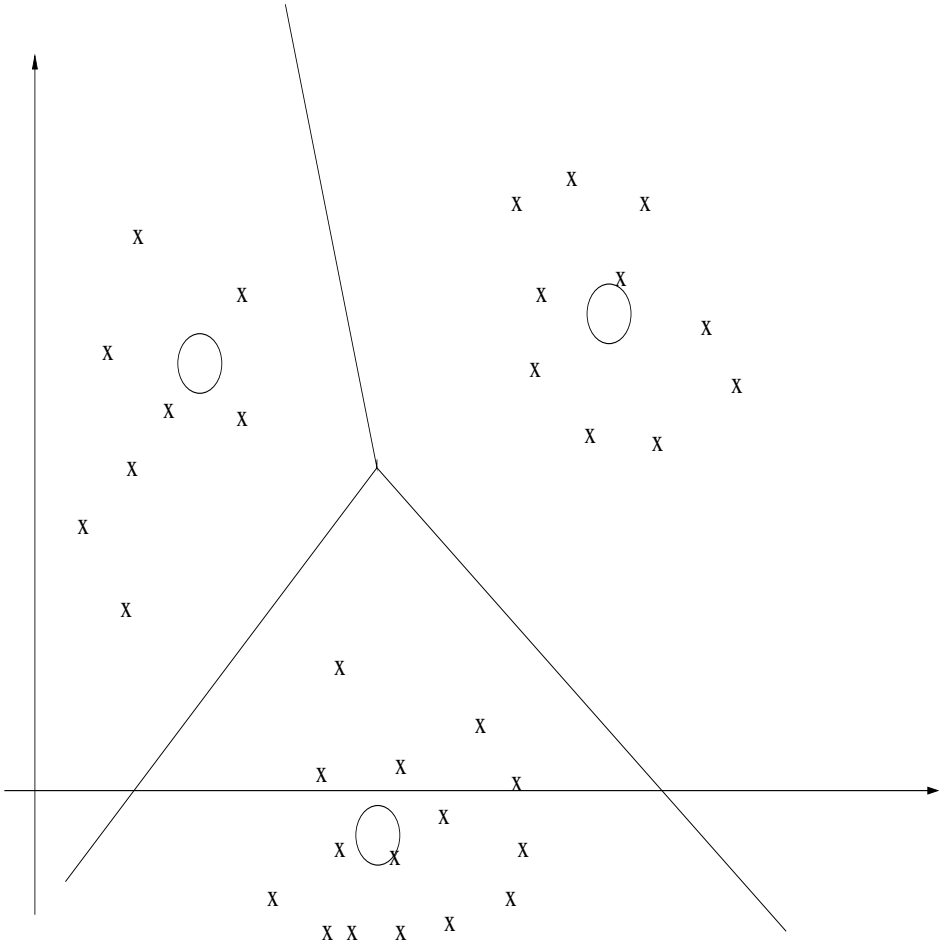


k-means clustering



Convergence

k-means clustering: Iteration 4



Convergence

Usually this iterative procedure of redefining centroids and reassigning the examples to clusters needs only a few iterations to converge. There are several strict mathematical proofs of this convergence.

KTH VETENSKAP OCH KONST

KTH Matematik

Important issues in automatic cluster detection :

The kinds of questions we want to be answered in data preparation for successful application.

(1) Distance measure

Most clustering techniques use for the distance measure the Euclidean distance formula (square root of the sum of the squares of distances along each attribute axes). More about distance, to be formulated as a **metric**, will be said soon.

Non-numeric variables must be transformed and scaled before the clustering can take place.

KTH Matematik

(2) Choice of the right number of clusters :

If the number of clusters k in the k-means method is not chosen so as to match the natural structure of the data, the results will not be good. The proper way to alleviate this is to experiment with different values for k. In principle, the best k value will exhibit the smallest intra-cluster distances and largest inter-cluster distances. More sophisticated techniques or software platforms measure these qualities automatically, and optimize the number of clusters in a separate loop

(3) Cluster interpretation

Once the clusters are discovered they have to be interpreted in order to have some value for data mining. Often this is done by letting a domain expert study the clustering. Some descriptive data mining technique (like decision trees) can be used to find descriptions of clusters.
The question of the validity of clustering will be addressed in the final part of this lecture.

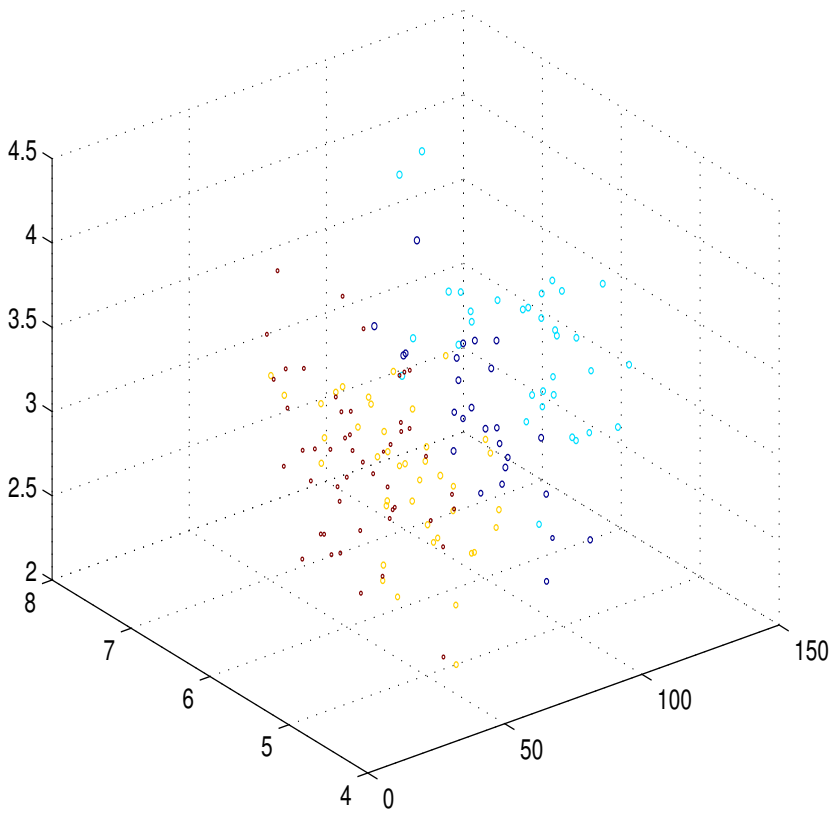
KTH VETENSKAP OCH KONST
KTH Matematik

(3) Cluster interpretation (cont'nd)

There are different ways to utilize clustering results:
Cluster membership can be used as a label for the separate classification problem, c.f., clustering lecture 2 in this course.
Clusters can be visualized using 2D and 3D scatter graphs or some other visualization technique (e.g., principal components analysis).
Differences in attribute values among different clusters can be examined, one attribute at a time.

KTH Matematik

2D scatter graph



(4) Application issues

Clustering techniques are used when we expect natural groupings (like biological species or genera) in examples of the data. Clusters should then represent groups of items (products, events, biological organisms) that have a lot in common. Creating clusters prior to application of some data analysis technique (decision trees, neural networks) might reduce the complexity of the problem by dividing the space of examples. These partitions can be studied separately and such a two step procedure might exhibit improved results (descriptive or predictive) as compared to analysis without using clustering.

KTH Matematik

Distance

Distance is usually a metric, which is a general mathematical concept formulating the requirements for measuring distance in an intelligent way. A metric is a function $D(\mathbf{x}, \mathbf{y})$ which maps two elements $\mathbf{x}$ and $\mathbf{y}$ of a set to a real number, such that

$$D(\mathbf{x}, \mathbf{x}) = 0,$$

$$D(\mathbf{x}, \mathbf{y}) > 0, \text{ if } \mathbf{x} \text{ is not equal to } \mathbf{y}$$

$$D(\mathbf{x}, \mathbf{y}) = D(\mathbf{y}, \mathbf{x}),$$

$$D(\mathbf{x}, \mathbf{y}) + D(\mathbf{y}, \mathbf{z}) \geq D(\mathbf{x}, \mathbf{z}) \text{ 'triangle inequality'}$$

KTH Matematik

Example: The Jaccard Distance

We have two examples (objects) described by d binary (0 or 1) features. They score simultaneously 1 on A features, both simultaneously 0 on D features, object 1 scores 1 while object 2 scores 0 on C features, and object 1 scores 0 while object 2 scores 1 on D variables. Thus $A + B + C + D = d$, the total number of features.

	1	0
1	A	B
0	C	D

This kind of four-table and the distance or similarity measures based on it are at the heart of much of clustering work done for biological systematics.

Example: The Jaccard Distance (cont'nd)

Jaccard's matching coefficient:

$$\frac{A}{A+B+C}$$

The Dice coefficient gives more weight for positive agreement:

$$\frac{2A}{2A+B+C}$$

It can be proved that

$$1-\frac{A}{A+B+C}=\frac{B+C}{A+B+C}$$

is actually a metric (for binary d-vectors) called the **Jaccard distance** or **Jaccard's mismatch coefficient**.

KTH VETENSKAP OCH KONST

KTH Matematik

Manhattan distance:

The Manhattan distance: $\mathbf{x}$ and $\mathbf{y}$ are d dimensional column vectors $\mathbf{x} = (x_1, \ldots, x_d)^T$, $\mathbf{y} = (y_1, \ldots, y_d)^T$,

$$D(\mathbf{x}, \mathbf{y}) = \sum_{i=1}^{d} |x_i - y_i|$$

KTH Matematik

More Examples:

Euclidean Distance: $\mathbf{x}$ and $\mathbf{y}$ are d dimensional column vectors $\mathbf{x} = (x_1, \ldots, x_d)^T$, $\mathbf{y} = (y_1, \ldots, y_d)^T$, (^{T} denotes vector transpose)

$$D(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{d} \mid x_i - y_i \mid^2}$$

Mahalanobis Distance: A distance metric that takes into account the 'covariance of the distribution of the data' (?).

$$D_\Sigma(\mathbf{x}, \mathbf{y}) = \sqrt{\sum_{i=1}^{d} \sum_{j=1}^{d} w_{ij} (x_i - y_i)(x_j - y_j)}$$

where w_{ij} is the element in the position (i, j) in a matrix Σ^{-1}, where Σ is a positive definite and symmetric $d \times d$-matrix (a covariance matrix).

KTH Matematik

More on the Mahalanobis Distance

We write for some future purposes the Mahalanobis Distance as

$$D_\Sigma(\mathbf{x},\mathbf{y}) = \sqrt{\sum_{i=1}^{d}\sum_{j=1}^{d} w_{ij}\,(x_i - y_i)\,(x_j - y_j)} =$$

$$\sqrt{(\mathbf{x}-\mathbf{y})^T\Sigma^{-1}(\mathbf{x}-\mathbf{y})} = \|\Sigma^{-1/2}\,(\mathbf{x}-\mathbf{y})\,\|,$$

where now ^{T} denotes the transpose of a column vector. Then we see that the Euclidean distance is obtained by $\Sigma = I$ ($d \times d$ identity matrix).

$$D_I(\mathbf{x},\mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T(\mathbf{x}-\mathbf{y})} = \|\mathbf{x}-\mathbf{y}\|.$$

KTH VETENSKAP OCH KONST

KTH Matematik

Euclidean Norm

$$D_I(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x} - \mathbf{y})^T (\mathbf{x} - \mathbf{y})} = \|\mathbf{x} - \mathbf{y}\|.$$

Recall that if $\mathbf{x} = (x_1, \ldots, x_d)$, then

$$\|\mathbf{x}\| = \sqrt{\sum_{i=1}^{d} x_i^2}.$$

is the Euclidean norm, so that

$$\|\mathbf{x} - \mathbf{y}\| = \sqrt{\sum_{i=1}^{d} (x_i - y_i)^2}$$

KTH VETENSKAP OCH KONST

KTH Matematik

Gaussian Probability Density

- A Gaussian density can now be written as

$$p\left(\mathbf{x};\mu,\Sigma\right)=\frac{1}{\left(2\pi\right)^{d/2}\sqrt{\det\Sigma}}e^{-\frac{1}{2}D_{\Sigma}(\mathbf{x},\mu)^{2}}$$

where $\mathbf{x}=\left(x_1,x_2,\cdots,x_d\right)^T$. Here μ is the mean vector of the density and Σ is the covariance matrix assumed invertible.

Different methods can be used to cluster the same set of data. One way to compare results by different clustering methods is to define (computable) distances between clusterings. This is also a means of validating clusterings. For example, certain of these metrics measure the distance as the length of the shortest sequence of admissible transformations required to obtain one partition from another.

KTH VETENSKAP OCH KONST

KTH Matematik

k-means clustering: a formal summary (1)

Let us suppose now that our data is

$$\mathcal{X}^t = \left\{ \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(t)} \right\}.$$

Here $\mathcal{X}^t$ can be called a **data matrix**.
Then, as an exercise in formalities, we have that the k-means algorithm can be summarized as a method of finding k centroids $\mathbf{a}_j$, $j = 1, \ldots, k$, such that

$$W\left(\mathbf{a}_1, \ldots, \mathbf{a}_k\right) = \sum_{l=1}^{t} \min_{1 \leq j \leq k} D\left(\mathbf{x}^{(l)}, \mathbf{a}_j\right)$$

is minimized, where $D\left(\mathbf{x}^{(l)}, \mathbf{a}_j\right)$ can be any generic metric.
Mathematically speaking we can only hope to obtain a local maximum of this expression.

KTH VETENSKAP OCH KONST
KTH Matematik

k-means clustering: a formal summary (2)

Let $\mathcal{X}^t = \left\{\mathbf{x}^{(l)}\right\}_{l=1}^{t}$, where $x^{(l)} \in R^d$. A clustering of $\mathcal{X}^t$ is any subdivision of $\mathcal{X}^t$ into k disjoint subsets c_j ($\subseteq \mathcal{X}^t$). We define the cluster membership functions (class label vector)

$$u_j^{(l)} := \begin{cases} 1 & \text{if } x^{(l)} \in c_j \\ 0 & \text{otherwise,} \end{cases}$$

and put these in the matrix

$$U = \left\{u_j^{(l)}\right\}_{l=1,j=1}^{t,k}.$$

The steps of the k-means algorithm are:

1. Select randomly k points $\mathbf{a}_1, \ldots, \mathbf{a}_k$ to be the seeds for the centroids of k clusters.
2. Assign each example to the centroid closest to the example,

$$u_{j^*}^{(l)} = 1 \Leftrightarrow j^* = \arg \min_{1 \leq j \leq k} D\left(\mathbf{x}^{(l)}, \mathbf{a}_j\right)$$

   forming in this way k exclusive clusters c_j^* of $\mathcal{X}^t$.
3. Calculate new centroids of the clusters

$$\mathbf{a}_j^* = \frac{1}{t_j} \sum_{l=1}^{t} u_{j^*}^{(l)} \mathbf{x}^{(l)}, t_j = \sum_{l=1}^{t} u_{j^*}^{(l)},$$

The steps of the k-means algorithm are:

4. Check if the cluster centroids have changed their "coordinates", i.e., if $\mathbf{a}_j^* \neq \mathbf{a}_j$. If yes, start again from the step 2. $\mathbf{a}_j^* \mapsto \mathbf{a}_j$. If not, cluster detection is finished and all $\mathbf{x}^{(l)}$ have their cluster memberships $u_{j^*}^{(l)}$ defined.

KTH VETENSKAP OCH KONST
KTH Matematik

Of course, in principle clustering is straightforward: just find all possible partitions of $\mathcal{X}^t = \left\{\mathbf{x}^{(l)}\right\}_{l=1}^{t}$ into k subsets, calculate the centroids $\mathbf{a}_j$, evaluate

$$W\left(\mathbf{a}_1, \ldots, \mathbf{a}_k\right) = \sum_{l=1}^{t} \min_{1 \leq j \leq k} D\left(\mathbf{x}^{(l)}, \mathbf{a}_j\right),$$

and take the partition that gives the least value of $W\left(\mathbf{a}_1, \ldots, \mathbf{a}_k\right)$. Unfortunately the number of partitions of a finite number of items is given by a Stirling's number of the second kind, and these grow fast to an astronomical size with t.

k-means clustering produces Voronoi Regions

Let R^d be a euclidean space with distance function D. Let K be a set of indices and let $(a_l)_{l \in K}$ be an ordered collection of vectors in R^d.

KTH VETENSKAP OCH KONST

KTH Matematik

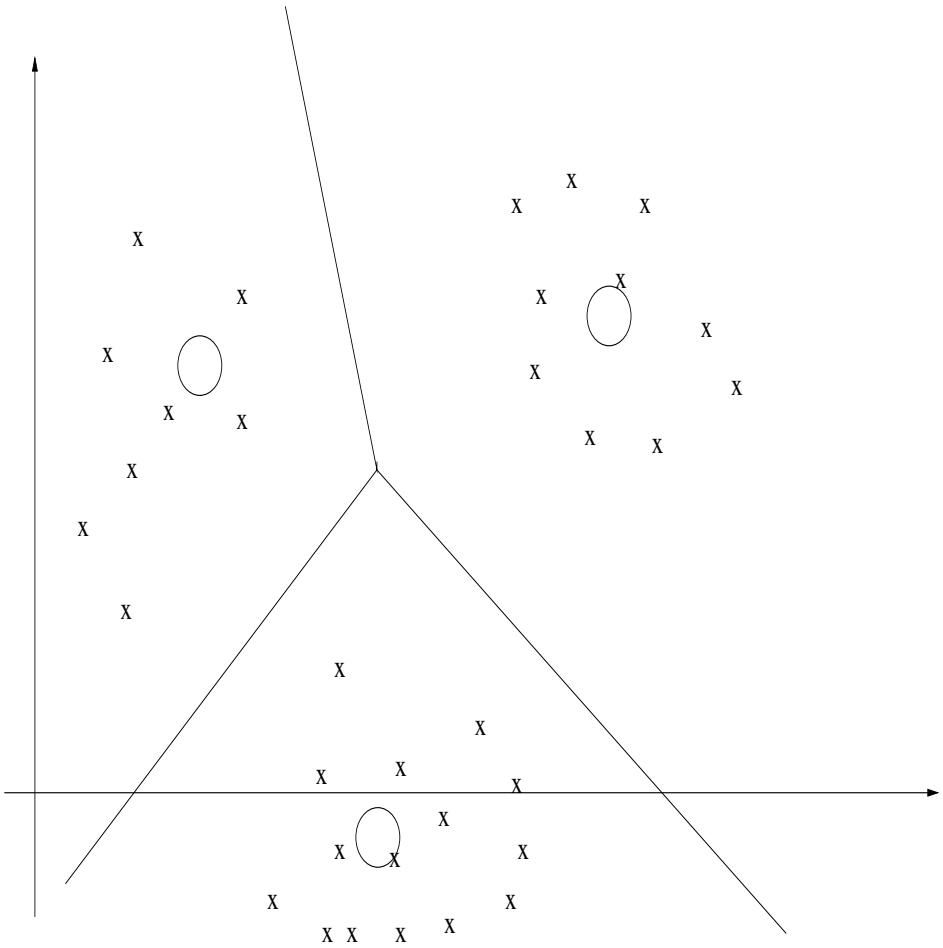
Voronoi Region

The **Voronoi cell**, or **Voronoi region**, R_l, associated with the centroid P_l is the set of all points in R^d whose distance to a_k is not greater than their distance to the other centroids a_j, where j is any index different from k. In other words, if $D(x, A) = \inf\{D(x, y) \mid y \in A\}$ denotes the distance between the point x and the subset A, then

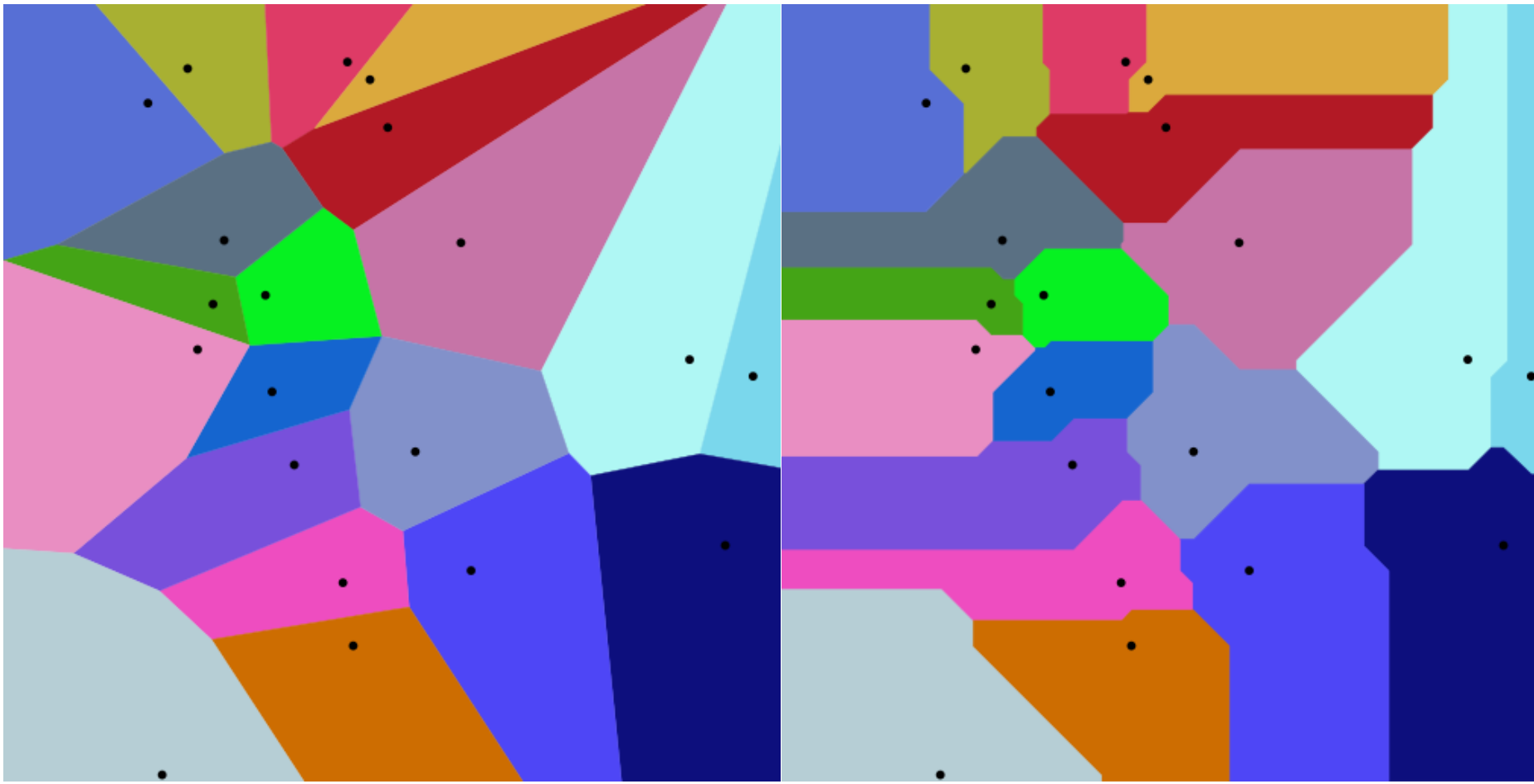
$$R_l = \{x \in R^d \mid D(x, a_l) \leq D(x, a_j) \text{ for all } j \neq l\}$$

The Voronoi diagram is simply the tuple of cells $(R_l)_{l \in K}$.

Voronoi regions



Voronoi regions w.r.t. Euclidean (l), Manhattan (r)



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.

Aristotle: Partition and Prediction

With description of a class or Voronoi cell, or by knowing the centroid, we should be able to predict (well) the properties of an item in this class without having investigated the item $\rightarrow$ a 'natural classification' $\rightarrow$ uncertainty in the prediction.

KTH Matematik

FAMILLES DES PLANTES.

Par M. ADANSON, de l'Académie des Sciences, de la Société Roiale de Londres, Censeur Roial.

I. PARTIE.

Contenant une Préface Istorike sur l'état ancien & actuel de la Botanike, & une Téorie de cette Science.

Tot generibus Herbarum, utilitatibus hominum aut voluptatibus genitis recenſitis, quantò plura reſtant, quantò quæ mirabiliora inventa! PLIN. Hiſt. nat. Lib. 22 Procem.

A PARIS,
Chez VINCENT, Imprimeur-Libraire de Mgr le Comte de PROVENCE, rue S. Severin.

M DCC LXIII.

AVEC APPROBATION, ET PRIVILEGE DU ROI.

1911 Encyclopdia Britannica: Adanson, Michel

Adanson's **Familles des plantes** *(1763) described his classification system for plants, which was much opposed by Carolus Linnaeus, the Swedish botanist who had proposed his own classification system based on the reproductive organs of plants. He founded his classification of all organized beings on the consideration of each individual organ. As each organ gave birth to new relations, so he established a corresponding number of arbitrary arrangements. Those beings possessing the greatest number of similar organs were referred to one great division, and the relationship was considered more remote in proportion to the dissimilarity of organs (i.e. a statistical method in botanical classification). His system of classification was not widely successful, and it was superseded by the Linnaean system.*

Partition as a Prediction

Probability theory provides us models to handle systematically the uncertainty of an item with regard to (a cell in) a partition, and methods to learn the classification. This can be seen as a pragmatic way of doing algorithmic clustering and learning.

KTH VETENSKAP OCH KONST

KTH Matematik

(C) Gaussian Mixture Models (1)

- Gaussian mixture models lead to probabilistic ('soft') versions of the k-means algorithm.
- Multivariate Gaussian Density
  $p(\mathbf{x}; \mu, \Sigma)$ is a multivariate Gaussian density if

$$p(\mathbf{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma}} e^{-\frac{1}{2} D_\Sigma(\mathbf{x},\mu)^2}$$

  where $\mathbf{x} = (x_1, x_2, \cdots, x_d)^T$. Here μ is the mean vector of the density and Σ is the covariance matrix.

Gaussian Mixture Models (2)

- A Class-conditional Multivariate Gaussian Density:

$$p\left(\mathbf{x} \mid u_j = 1; \mu_j, \Sigma_j\right) =$$

$$\frac{1}{(2\pi)^{d/2}\sqrt{\det \Sigma_j}} e^{-\frac{1}{2} D_{\Sigma_j}(\mathbf{x}, mu_j)^2}$$

- Mixing distribution: $\lambda_j = P\left(u_j = 1\right)$,

Gaussian mixture:

$$p\left(\mathbf{x}\right) = \sum_{j=1}^{k} \lambda_j p\left(\mathbf{x} \mid u_j = 1; \mu_j, \Sigma_j\right)$$

KTH Matematik

Gaussian Mixture Models (3)

'Generative' (or sampling) interpretation of clustering using the Gaussian mixture:

- Draw a label j using the distribution $\lambda_j, j = 1, \ldots, k$
- Draw a sample $\mathbf{x}^{(l)}$ from the class-conditional multivariate Gaussian density pointed out by the label j.
- This is repeated to yield $\mathcal{X}^t = \left\{\mathbf{x}^{(l)}\right\}_{l=1}^{t}$.

The density of any $\mathbf{x}^{(l)}$ is by this given by:

$$p\left(\mathbf{x}^{(l)}\right) = \sum_{j=1}^{k} \lambda_j p\left(\mathbf{x}^{(l)} \mid u_j = 1; \mu_j, \Sigma_j\right)$$

NOTE that the mixture density

$$p(\mathbf{x}) = \sum_{j=1}^{k} \lambda_j p(\mathbf{x} \mid u_j = 1; \mu_j, \Sigma_j)$$

is in general NOT a Gaussian density any more.
In practice we do not have access to the labels and we need also to find μ_j and Σ_j, $j = 1, \ldots, k$ using the samples $\mathcal{X}^t = \left\{\mathbf{x}^{(l)}\right\}_{l=1}^{t}$.
Describing clusters by means fitting a Gaussian mixture to data is an instance of **unsupervised statistical learning**. We shall now show how this is done using the EM-algorithm.

The EM-algorithm for Gaussian Mixture Models (1)

The EM-algorithm is an iteration for fitting a Gaussian mixture that alternates between two steps:

- Compute the 'responsibility' [2] of each class-conditional Gaussian density for each sample vector
- The means of the class-conditional Gaussian densities are moved to the centroid of the data, weighted by the responsibilities, the covariance matrices are up-dated weighted by the responsibilities.

[2] terminology due to Michael I. Jordan

KTH Matematik

The formal EM-algorithm for Gaussian Mixture Models (2)

The two steps are called the 'E-step' and 'M-step' for estimation and maximization, respectively. Suppose that we have performed iteration at step r and have thus obtained initial estimates $\mu_j^{(r)}$ and $\Sigma_j^{(r)}$ and proceed to step $r+1$.

E-step The responsibilities $p^{(r+1)}\left(j|\mathbf{x}^{(l)}\right)$ of each cluster/class for the data point $\mathbf{x}^{(l)}$ are computed as

$$p^{(r+1)}\left(j|\mathbf{x}^{(l)}\right)$$

$$= \frac{w_j^{(r)} p\left(\mathbf{x}^{(l)} \mid u_j = 1; \mu_j^{(r)}, \Sigma_j^{(r)}\right)}{\sum_{j=1}^{k} w_j^{(r)} p\left(\mathbf{x}^{(l)} \mid u_j = 1; \mu_j^{(r)}, \Sigma_j^{(r)}\right)}$$

KTH VETENSKAP OCH KONST
KTH Matematik

We note that in more conventional terms of probability calculus $p^{(r+1)}\left(j|\mathbf{x}^{(l)}\right)$ is the up-date of the conditional probability of the class label j given $\mathbf{x}^{(l)}$.
This probability can be used to find a 'soft' partition.
In addition we have the update

$$w_j^{(r+1)} = \frac{1}{t}\sum_{l=1}^{t} p^{(r)}\left(j|\mathbf{x}^{(l)}\right).$$

A Method of Unsupervised Classification: The CEM-algorithm for Gaussian Mixture Models (1)

Suppose that we have performed iteration at step r and have thus obtained initial estimates $\mu_j^{(r)}$ and $\Sigma_j^{(r)}$ and proceed to step $r+1$. The E- and M-steps are as before, we recall the E-step:

E-step The posterior probabilities $p^{(r+1)}\left(j|\mathbf{x}^{(l)}\right)$ of each cluster/class for the data point $\mathbf{x}^{(l)}$ are computed as

$$p^{(r+1)}\left(j|\mathbf{x}^{(l)}\right)$$

$$= \frac{w_j^{(r)} p\left(\mathbf{x}^{(l)} \mid u_j = 1; \mu_j^{(r)}, \Sigma_j^{(r)}\right)}{\sum_{j=1}^{k} w_j^{(r)} p\left(\mathbf{x}^{(l)} \mid u_j = 1; \mu_j^{(r)}, \Sigma_j^{(r)}\right)}$$

KTH VETENSKAP OCH KONST
KTH Matematik

- M-step contn'd: The updated covariance matrices are computed as

$$\Sigma_j^{(r+1)}$$

$$= \frac{\sum_{l=1}^{t} \left(\mathbf{x}^{(l)} - \mu_j^{(r+1)}\right) \left(\mathbf{x}^{(l)} - \mu_j^{(r+1)}\right)^T p^{(r)}\left(j|\mathbf{x}^{(l)}\right)}{\sum_{l=1}^{t} p^{(r)}\left(j|\mathbf{x}^{(l)}\right)}.$$

The so called '*Sundberg equations*',c.f.,
Rolf Sundberg (1974): Maximum Likelihood Theory for Incomplete Data from an Exponential Family. *Scandinavian Journal of Statistics*, 1, pp. 49−58.

Some comments on the EM-algorithm (1)

Let us set

$$p\left(\mathbf{x}^{(l)};\mathbf{\Sigma},\Lambda\right)=\sum_{j=1}^{k}\lambda_j p\left(\mathbf{x}^{(l)}\mid u_j=1;\mu_j,\Sigma_j\right)$$

and for $\mathcal{X}^t=\left\{\mathbf{x}^{(l)}\right\}_{l=1}^{t}$

$$P\left(\mathcal{X}^t|\mathbf{\Sigma},\Lambda\right)=\prod_{l=1}^{t}p\left(\mathbf{x}^{(l)};\mathbf{\Sigma},\Lambda\right)$$

If we set $L_{\mathcal{X}^t}\left(\mathbf{\Sigma},\Lambda\right)\equiv P\left(\mathcal{X}^t|\mathbf{\Sigma},\Lambda\right)$ (likelihood function), then we have

1. *For every iteration of the* **EM** *- algorithm*

$$L_{\mathcal{X}^t}\left(\mathbf{\Sigma}^{(r+1)},\Lambda^{(r+1)}\right)\geq L_{\mathcal{X}^t}\left(\mathbf{\Sigma}^{(r)},\Lambda^{(r)}\right)$$

2. $\left(\Theta^{(r)},\Lambda^{(r)}\right)$ *converges to a stationary point of* $L_{\mathcal{X}^t}\left(\mathbf{\Sigma},\Lambda\right)$ *as* $r\to\infty$.

The EM-algorithm for Gaussian Mixture Models & Add Clustering

Once the algorithm has converged, make clusters by assigning $\mathbf{x}^{(l)}$ to a class satisfying

$$j^{opt} = \text{argmax}_{1 \leq j \leq k} p^{(*)} \left(j | \mathbf{x}^{(l)} \right) .$$

KTH VETENSKAP OCH KONST

KTH Matematik

Validity of Clustering

A clustering software applied to any set of data will always produce a clustering. How can we be assured of that such a result is a 'valid clustering' ?
This would, from the point of view of a statistician, require a null model for data, and a procedure for computing the distribution of some test statistic under this null model. There would seem to be several difficulties in this approach.
For a survey of the topic, especially in the context of hierarchic methods, we refer to
A.D. Gordon (1994): Identifying genuine clusters in a classification. *Computational Statistics & Data Analysis*, 18, pp. 561−581.

KTH Matematik

Validity of Clustering, one more comment

A method for the validation of k-means that is much used
W.M. Rand (1973): Objective criteria for the evaluation of clustering methods. *Journal of American Statistical Association*, 66, pp. 846−850.
There is one very simple approach to secure valid clustering. This is based on Bayesian classification with rejection and is to be presented in the end if this lecture.

KTH VETENSKAP OCH KONST
KTH Matematik

A Method of Unsupervised Classification: The CEM-algorithm for Gaussian Mixture Models (1)

Suppose that we have performed iteration at step r and have thus obtained initial estimates $\mu_j^{(r)}$ and $\Sigma_j^{(r)}$ and proceed to step $r+1$. The E- and M-steps are as before, we recall the E-step:

E-step The posterior probabilities $p^{(r+1)}\left(j|\mathbf{x}^{(l)}\right)$ of each cluster/class for the data point $\mathbf{x}^{(l)}$ are computed as

$$p^{(r+1)}\left(j|\mathbf{x}^{(l)}\right)$$

$$= \frac{w_j^{(r)} p\left(\mathbf{x}^{(l)} \mid u_j = 1; \mu_j^{(r)}, \Sigma_j^{(r)}\right)}{\sum_{j=1}^{k} w_j^{(r)} p\left(\mathbf{x}^{(l)} \mid u_j = 1; \mu_j^{(r)}, \Sigma_j^{(r)}\right)}$$

KTH VETENSKAP OCH KONST
KTH Matematik

The CEM-algorithm for Gaussian Mixture Models (2): Classification

C-step After the E and M-steps have converged, we let

$$\mathbf{P}^* = [p^{(*)}(j|\mathbf{x}^{(l)})]_{j=1}^{k}{}_{l=1}^{t}.$$

We classify $\mathbf{x}^{(l)}$ to c_{j_*} if

$$j_* = \arg \max_{1 \leq j \leq k} p^*(j|\mathbf{x}^{(l)})$$

for $l = 1, \ldots, t$.

Here we note a difference to the learning methods of supervised learning outlined in a previous lecture: all of the data is used to estimate all of the statistical parameters.

KTH Matematik

A Method of Unsupervised Classification: Classification Maximum Likelihood (CML) (1)

Let us again consider the class-conditional Gaussian densities for $\mathcal{X}^t = \left\{\mathbf{x}^{(l)}\right\}_{l=1}^{t}$. A log likelihood function is defined as

$$l\left(\mu, \Sigma, \Lambda, U\right)$$

$$= \sum_{j=1}^{k} \sum_{l=1}^{t} u_j^{(l)} \left[\log \lambda_j p\left(\mathbf{x}^{(l)} \mid u_j = 1; \mu_j, \Sigma_j\right)\right],$$

where

$$u_j^{(l)} := \begin{cases} 1 & \text{if } x^{(l)} \in c_j \\ 0 & \text{otherwise,} \end{cases}$$

KTH VETENSKAP OCH KONST

KTH Matematik

Classification Maximum Likelihood (CML) (2)

We estimate the statistical parameters and the clusters by maximum likelihood:

$$\left(\widehat{\mu}, \widehat{\Sigma}, \widehat{\Lambda}, \widehat{U}\right) = \arg \max_{\mu,\Sigma,\Lambda,U} l\left(\mu, \Sigma, \Lambda, U\right).$$

Algorithmically this can be done using an alternating algorithm as follows:

KTH Matematik

Classification Maximum Likelihood (CML) (3)

1. Assign randomly k clusters for $\mathcal{X}^t$.
2. Find $\widehat{\Sigma}_j$, $\widehat{\mu}_j$, $\widehat{\Lambda}$ using maximum likelihood with the data in each cluster.
3. Calculate new class memberships using the plug-in discriminant functions.

KTH Matematik

Classification Maximum Likelihood (CML) (4)

4. Check if the cluster memberships have changed. If yes, start again from the step 2. If not, clustering is finished and all $\mathbf{x}^{(l)}$ have their cluster memberships defined by CML.

Classification Maximum Likelihood with Binary Vectors (1)

The

$$\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(t)}$$

are binary vectors,

$$x_i^{(l)} \in [0,1], i \in [1, \ldots, d]$$

and

$$f(\mathbf{x}|\theta_j) = \prod_{i=1}^{d} \theta_{ij}^{x_i}(1-\theta_{ij})^{1-x_i}; x_j \in \{0,1\}; 0 \leq \theta_{ij} \leq 1.$$

and

$$\theta_j = (\theta_{1j}, \ldots, \theta_{dj})$$

KTH Matematik

Classification Maximum Likelihood with Binary Vectors (1)

(Steps 0-2) (Estimate parameters):

Steg 0.
Take $\Lambda^{(0)}$ and $\Theta^{(0)}$ (random), so that

$$P\left(\mathbf{x}^{(l)}|\Lambda^{(0)},\Theta^{(0)}\right)=\sum_{j=1}^{k}\lambda_j^{(0)}f(\mathbf{x}^{(l)}|\theta_j^{(0)})>0;$$

$l=1,\ldots,t$ and compute

$$\mathbf{P}^{(0)}=[p^{(0)}(j|\mathbf{x}^{(l)})]_{j=1}^{k}{}_{l=1}^{t}$$

$$p^{(0)}(j|\mathbf{x}^{(l)})=\frac{\lambda_j^{(0)}f(\mathbf{x}^{(l)}|\theta_j^{(0)})}{P\left(\mathbf{x}^{(l)}|\Lambda^{(0)},\Theta^{(0)}\right)}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Step 1. Find $\Theta^{(s+1)}$ and $\Lambda^{(s+1)}$ for $s = 1, 2, \ldots$ by the aid of

$$\lambda_j^{(s+1)} = \frac{1}{t} \sum_{l=1}^{t} p^{(s)}(j|\mathbf{x}^{(l)});$$

and

$$\theta_j^{(s+1)} = \frac{1}{\sum_{l=1}^{t} p^{(s)}(j|\mathbf{x}^{(l)})} \sum_{l=1}^{t} \mathbf{x}^{(l)} p^{(s)}(j|\mathbf{x}^{(l)});$$

$j = 1, 2, \ldots, k$.

Step 2. Compute $\mathbf{P}^{(s+1)}$ from $\Theta^{(s+1)}$ and $\Lambda^{(s+1)}$

$$\mathbf{P}^{(s+1)} = [p^{(s+1)}(j|\mathbf{x}^{(l)})]_{j=1}^{k}{}_{l=1}^{t};$$

$$p^{(s+1)}(j|\mathbf{x}^{(l)}) = \frac{\lambda_j^{(s+1)} f(\mathbf{x}^{(l)}|\theta_j^{(s+1)})}{P\left(\mathbf{x}^{(l)}|\Lambda^{(s+1)}, \Theta^{(s+1)}\right)}$$

Halt if $\lambda_j^{(s+1)} = \lambda_j^{(s)}$, $\theta_j^{(s+1)} = \theta_j^{(s)}$, else continue with Step 1.

KTH Matematik

Step 3. Classification: After step 2 we denote

$$\mathbf{P}^* = [p^{(*)}(j|\mathbf{x}^{(l)})]_{j=1}^{k}{}_{l=1}^{t}.$$

We put $x^{(l)}$ in the class C_{j_*} if

$$j_* = \arg\max_{1\le j\le k} p^*(j|\mathbf{x}^{(l)})$$

for $l = 1, \ldots, t$.

Well-Classified

We say that $\mathbf{x}^{(l)}$ is 'well-classified', if

$$(\dagger) p^*(j|\mathbf{x}^{(l)}) \geq 0.995 (p^*(j|\mathbf{x}^{(l)}) \geq 1 - \epsilon)$$

'.. the problem of evaluating the quality of cluster assignment of individual items ..' .

- Run CEM for $\mathcal{X}$.
- Throw away those $\mathbf{x}^{(l)}$, that do not live up to $(\dagger)$ and run CEM with those that remain.

This improves in terms of likelihood the result for k fixed.

KTH VETENSKAP OCH KONST
KTH Matematik

Schnell's method of unsupervised clustering (1)

Schnell's method (to be presented) is not a recent idea, the original paper was published in 1964[3]. We choose to present the method, since it is quite different from the methods presented so far and does not seem to have gained much attention in the literature or in the textbooks or in practice (?).

[3] For this reference and the results below we refer to J. Grim (1981): An algorithm for maximizing a finite sum of positive functions and its applications to cluster analysis. *Problems of Control and Information Theory*, 10, pp. 427−437.

KTH VETENSKAP OCH KONST
KTH Matematik

Schnell's method of unsupervised clustering (2)

The data matrix is

$$\mathcal{X}^t = \left\{ \mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(t)} \right\}.$$

We introduce a recursive algorithm

$$\mathbf{z}_{s+1} = \sum_{l=1}^{t} p\left(l \mid \mathbf{z}_s\right) \mathbf{x}^{(l)}$$

which can be proved to possess a finite number $(= k)$ of points of convergence denoted by

$$\left\{ \mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k \right\}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Schnell's method of unsupervised clustering (3)

In the recursion $\mathbf{z}_{s+1} = \sum_{l=1}^{t} p\left(l \mid \mathbf{z}_s\right) \mathbf{x}^{(l)}$ we take

$$p\left(l \mid \mathbf{z}_s\right) = \frac{f_l\left(\mathbf{z}_s\right)}{f\left(\mathbf{z}_s\right)},$$

and

$$f_l\left(\mathbf{z}_s\right) = \frac{1}{(2\pi)^{d/2}\sqrt{\det h \cdot \Sigma}} e^{-\frac{1}{2} D_{h\cdot\Sigma}\left(\mathbf{z}_s, \mathbf{x}^{(l)}\right)^2},$$

where $h > 0$ and

$$f\left(\mathbf{z}_s\right) = \frac{1}{t} \sum_{l=1}^{t} f_l\left(\mathbf{z}_s\right)$$

We can think of $\Sigma = I_d$.

KTH VETENSKAP OCH KONST

KTH Matematik

Schnell's method of unsupervised clustering (4)

$$\mathbf{z}_{s+1} = \sum_{l=1}^{t} p\left(l \mid \mathbf{z}_s\right) \mathbf{x}^{(l)}$$

has as its limit points the vectors

$$\left\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k\right\}.$$

Then the method of Schnell is to cluster the data matrix into k clusters c_j defined by

$$c_j = \left\{\mathbf{x}^{(l)} \mid \lim_{s\to\infty} \mathbf{z}_s = \mathbf{a}_j, \mathbf{z}_0 = \mathbf{x}^{(l)}\right\}.$$

In words this means that the algorithm takes each $\mathbf{x}^{(l)}$ as its initial value and then $\mathbf{x}^{(l)}$ is clustered to c_j if the algorithm converges to $\mathbf{a}_j$.

KTH Matematik

Schnell's method of unsupervised clustering (5)

$$\mathbf{z}_{s+1} = \sum_{l=1}^{t} p\left(l \mid \mathbf{z}_s\right) \mathbf{x}^{(l)}$$

$$c_j = \left\{ \mathbf{x}^{(l)} \mid \lim_{s \to \infty} \mathbf{z}_s = \mathbf{a}_j, \mathbf{z}_0 = \mathbf{x}^{(l)} \right\}.$$

The points $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k\}$ are local minima of an unknown density $p(\mathbf{x})$ generating the data.
We approximate $p(\mathbf{x})$ by $f(\mathbf{x})$ which is a sum of Gaussian kernels.
Does Schnell's method produce reasonable results ???

KTH VETENSKAP OCH KONST
KTH Matematik

Learning Vector Quantization (LVQ)

We have a training set

$$\mathcal{X}^t = \left\{ \mathbf{x}^{(l)} \right\}_{l=1}^{t}, U = \left\{ u_j^{(l)} \right\}_{j=1,l=1}^{k,t}.$$

We have also a set of reference vectors, $\{\mathbf{a}_1, \mathbf{a}_2, \ldots, \mathbf{a}_k\}$ for each of the classes, respectively.
LVQ is a kind of on-line version of k-means clustering.
LVQ gained popularity during the 1980's, but a first version of an algorithm like this was given by
McQueen, J.B.(1967): Some methods of classification and analysis of multivariate observations, *Proc. 5th Berkeley Symposium in Mathematics, Statistics and Probability*, vol 1., pp. 281-296, Univ. of California, Berkeley

KTH VETENSKAP OCH KONST
KTH Matematik

1. Assign each $\mathbf{x}^{(l)}$ to the closest $\mathbf{a}_c$ using, e.g., Euclidean distance.
2.
   - If $\mathbf{x}^{(l)}$ is correctly classified, i.e., its label in U is the one corresponding to $\mathbf{a}_c$ Update $\mathbf{a}_c(r)$ to form $\mathbf{a}_c(r+1)$ using $\mathbf{x}^{(l)}$ as follows

$$\mathbf{a}_c(r+1) = \mathbf{a}_c(r) + \alpha(k)\left[\mathbf{x}^{(l)} - \mathbf{a}_c(r)\right]$$

   - If $\mathbf{x}^{(l)}$ is incorrectly classified, then update $\mathbf{a}_c(r)$ as follows

$$\mathbf{a}_c(r+1) = \mathbf{a}_c(r) - \alpha(k)\left[\mathbf{x}^{(l)} - \mathbf{a}_c(r)\right]$$

   - $\mathbf{a}_i(r+1) = \mathbf{a}_i(r)$ for $i \neq c$.
3. Repeat step 2, and let $\alpha(k)$ decrease at each step toward zero.

But what is a cluster ?

J.A. Hartigan (1985) defines 'statistical cluster *as a (maximally connected) subset of a high density region'. Or, for a population with p.d.f. f in d dimensions to be the maximal connected sets of form* $\{x|f(x) \geq c\}$.

Probability rests on similarity between what we know and what we are guessing!

KTH VETENSKAP OCH KONST

KTH Matematik