DTU

02450 Introduction to Machine Learning and Data Mining

Week 11: Mixture models and density estimation

Bjørn Sand Jensen

22 April 2025

DTU Compute, Technical University of Denmark

Today

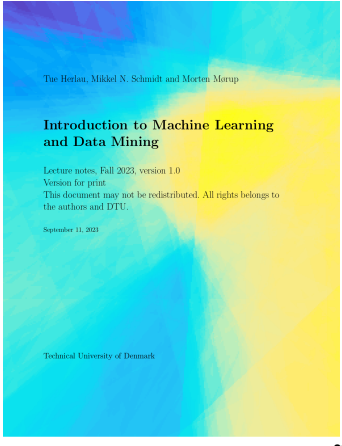
Feedback Groups of the day:

Anders Vesterholm-Lavesen, Stephan Larsen, Hao Du, Arun Kumar Dhuraisamy, Emil Broge Johansen, Shireesha ., Andreas Knorborg Frydensberg Ludvigsen, Ronja Toft Ørbeck, Nikolina Thorleifsson, Zhongan Zhao, Noel Willem Pihl, Kawa Shawki Bilal, Thomas Wijetunga Kærgaard, Lucas Bjerg Frandsen, Roozbeh Salajeghe, Alejandra Caballero Perez, Liza Szalai, Pawas Swain, Christian Madsen, Christian Eeg Nellemose, Markus Kenno Hansen, Maria Stentoft-Christensen, Caroline Lynge Nielsen, Mathilde Wismann Bechgaard, Mathias Munck Nikolajsen, Arnau Sage Costa I Laudisio, Estela Martín Cebrián, Lucio Edward Chavez Stevns, Sif Wittus Kaae Ludvigsen, Simone Sejdenfaden, Mathias Bang, Victor Leth Schmidt, Seyed Soheil Ghoreishi, Oskar Bak Jannings, Aoling Li, Andreas Emil Andersen, Mads August Claussen, Pavlos Neoklis Anglelopoulos, Kassandra Vendeltorp König, Anders Jensen, Eik Lykke Ring, Jawhara Hamoua

Reading/homework material:

Chapter 19 and 20

P20.1, P19.1, P19.2



Lecture Schedule

1. Introduction
   4 February: C1,C2

Data: Feature extraction, and visualization

2. Summary statistics, similarity and visualization
   11 February: C4,C7
3. Computational linear algebra and PCA
   18 February: C3
4. Probability and probability densities
   25 February: C5, C6

Supervised learning: Classification and regression

5. Decision trees and linear regression
   4 March: C8, C9 **(Project 1 due 6 March at 17:00)**
6. Overfitting, cross-validation and Nearest Neighbor
   11 March: C10, C12
7. Performance evaluation, Bayes, and Naive Bayes
   18 March: C11, C13
8. Artificial Neural Networks and Bias/Variance
   25 March: C14, C15
9. AUC and ensemble methods
   1 April: C16, C17

Unsupervised learning: Clustering and density estimation

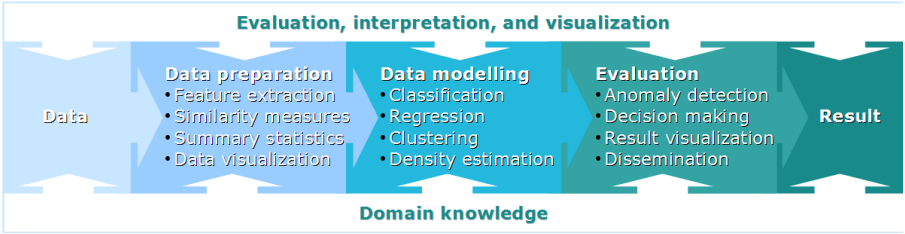
10. K-means and hierarchical clustering
    8 April: C18 **(Project 2 due 10 April at 17:00)**
11. **Mixture models and density estimation**
    22 April: C19, C20
12. Association mining
    29 April: C21

Recap

13. Recap and discussion of the exam
    6 May: C1-C21

Online help: Piazza
Videos of lectures: `https://panopto.dtu.dk`
Streaming of lectures: Zoom (link on DTU Learn)

Learning Objectives



Learning Objectives

- Explain the role of the parameters in the Gaussian Mixture Model (GMM) and how the parameters are updated using the EM-algorithm
- Explain how cross-validation can be used for the GMM
- Understand and apply (kernel) density, K-nearest neighbour density and average relative density estimation for outlier detection

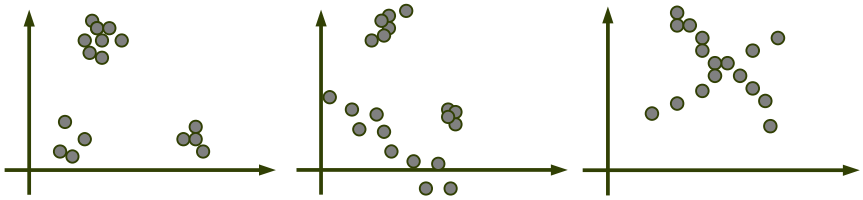
Imagine you observe the world for the first time!

http://www.clipartlord.com/category/baby-clip-art/

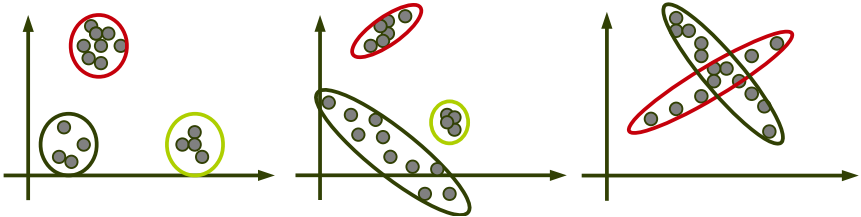
We humans are skilled at dividing objects into groups (clustering), but how do we make computers do the same?

Source (animal images): commons.wikimedia.org

- What are the clusters below and what characterize each cluster?
- Is K-means well suited for modeling the clusters below?
  - Will it always find the optimum solution?
  - Can it model the sizes of the clusters?
  - Can it model the shape of the clusters?
  - How can we determine the number of clusters?



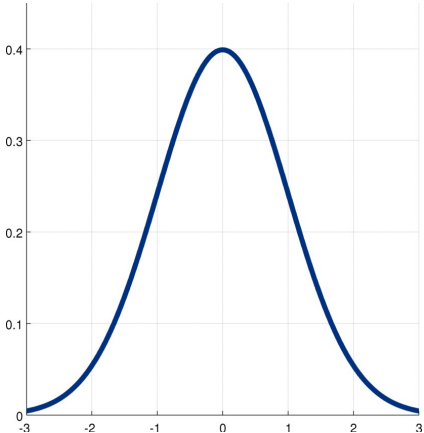
- What are the clusters below and what characterize each cluster?
- Is K-means well suited for modeling the clusters below?
  - Will it always find the optimum solution?
  - Can it model the sizes of the clusters?
  - Can it model the shape of the clusters?
  - How can we determine the number of clusters?



Recall: The Normal distrbution

- Probability density function describes the relative chance of a given value the relative chance of a given value
- Normal distribution characterized by
  - Mean
  - Variance

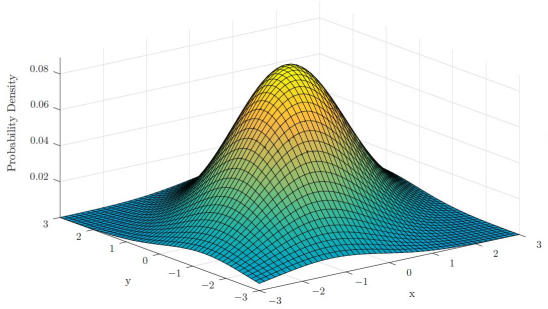
$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

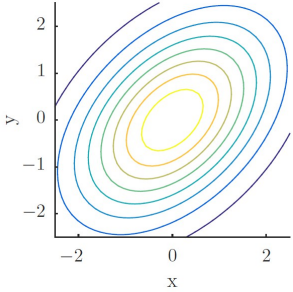


Recall: Multivariate Normal Distrbution

$$\boldsymbol{x} \in \mathbb{R}^M$$

$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^M |\boldsymbol{\Sigma}|}} exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$



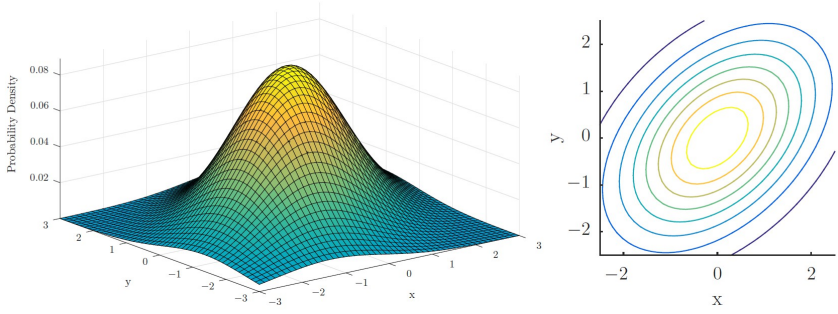


Recall: Multivariate Normal Distrbution

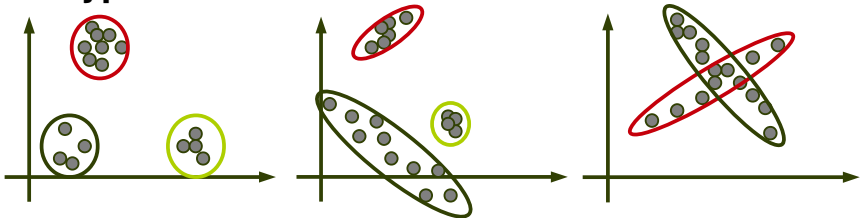
$$\boldsymbol{x} \in \mathbb{R}^M$$

$$\mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^M |\boldsymbol{\Sigma}|}} exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$$

Example: 2-dimensional Normal distribution



Prototypical mixture model

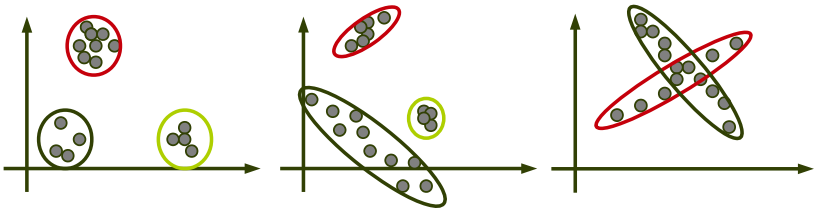


- We want a **density** $p(\boldsymbol{x})$ of our observations $\boldsymbol{x} \in \mathbb{R}^M$
- Suppose we have K clusters and let $z = k$ if $\boldsymbol{x}$ belongs to cluster k
- According to the the basic rules of probability:

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} p(\boldsymbol{x}, z = k) = \sum_{k=1}^{K} p(\boldsymbol{x}|z = k)p(z = k)$$

- If we specify $p(\boldsymbol{x}|z = k)$ and $p(z = k) = w_k$ we have a model

Gaussian Mixture Model (GMM)



- Different locations $\boldsymbol{\mu}_{(k)}$
- Different shape $\boldsymbol{\Sigma}_{(k)}$
- Different sizes w_k

Data density

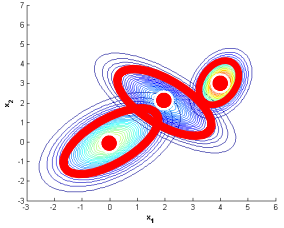
Sum of cluster specific densities assumed normal distributed

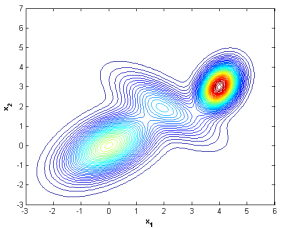
$$p(\mathbf{x}) = \sum_{k=1}^{K} w_k \mathcal{N}(\mathbf{x} \mid \mu_{(k)}, \boldsymbol{\Sigma}_{(k)})$$

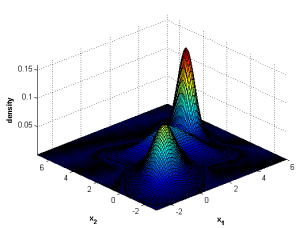
$$\text{s.t.} \sum_{k=1}^{K} w_k = 1, \; w_k \geq 0$$

GMM Example

$$p(\boldsymbol{x}) = 0.5\mathcal{N}(\boldsymbol{x}|\begin{bmatrix}0\\0\end{bmatrix},\begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}) + 0.2\mathcal{N}(\boldsymbol{x}|\begin{bmatrix}2\\2\end{bmatrix},\begin{bmatrix}1 & -0.7\\-0.7 & 1\end{bmatrix}) + 0.3\mathcal{N}(\boldsymbol{x}|\begin{bmatrix}4\\3\end{bmatrix},\begin{bmatrix}0.2 & 0.1\\0.1 & 0.5\end{bmatrix})$$







$\boldsymbol{\mu}_{(k)}$: Cluster center (prototypical example in cluster)

$\boldsymbol{\Sigma}_{(k)}$: Shape of the cluster

w_k : Relative size/density of the cluster

Quiz 01: GMM

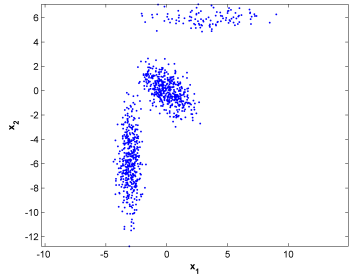


Figure 1: 1000 data observations drawn from a Gaussian Mixture Model (GMM) with three clusters.

In Figure 1 is shown 1000 observations drawn from a density defined by a Gaussian Mixture Model (GMM) with three clusters. Suppose $\mathcal{N}(\boldsymbol{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^2|\boldsymbol{\Sigma}|}} \exp\left(-\frac{1}{2}(\boldsymbol{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\boldsymbol{x}-\boldsymbol{\mu})\right)$ is the multivariate normal distribution, which one of the following GMM densities was used to generate the data?

A.
$$p(x) = 0.1 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 3 \\ 6 \end{bmatrix}, \begin{bmatrix} 5 & 0 \\ 0 & 0.25 \end{bmatrix}) + 0.45 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -3 \\ -6 \end{bmatrix}, \begin{bmatrix} 0.25 & 0 \\ 0 & 5 \end{bmatrix}) + 0.45 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix})$$

B.
$$p(x) = 0.45 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 3 \\ 6 \end{bmatrix}, \begin{bmatrix} 5 & 0 \\ 0 & 0.25 \end{bmatrix}) + 0.1 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -3 \\ -6 \end{bmatrix}, \begin{bmatrix} 0.25 & 0 \\ 0 & 5 \end{bmatrix}) + 0.45 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix})$$

C.
$$p(x) = 0.1 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 3 \\ 6 \end{bmatrix}, \begin{bmatrix} 0.25 & 0 \\ 0 & 5 \end{bmatrix}) + 0.45 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -3 \\ -6 \end{bmatrix}, \begin{bmatrix} 5 & 0 \\ 0 & 0.25 \end{bmatrix}) + 0.45 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix})$$

D.
$$p(x) = 0.1 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 3 \\ 6 \end{bmatrix}, \begin{bmatrix} 0.25 & 0 \\ 0 & 5 \end{bmatrix}) + 0.45 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} -3 \\ -6 \end{bmatrix}, \begin{bmatrix} 5 & 0 \\ 0 & 0.25 \end{bmatrix}) + 0.45 \cdot \mathcal{N}(\boldsymbol{x}|\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix})$$

E. Don't know.

GMM Sanity check time:

Consider the Gaussian mixture model (GMM)

$$p(\boldsymbol{x}) = \sum_{k=1}^{K} w_k \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_{(k)}, \boldsymbol{\Sigma}_{(k)}) \qquad s.t \quad \sum_{k=1}^{K} w_k = 1, w_K \geq 0$$

What is the value of the integral?

$$\int p(\boldsymbol{x}) d\boldsymbol{x}$$

The EM algorithm for GMMs

Select an initial set of model parameters
(mean and covariance for each cluster)

Repeat

- Expectation
  For each object, calculate the probability of belonging to each distribution
- Maximization
  For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change

E-step

$$p(z_n = k \mid \boldsymbol{x}_n) = \frac{w_k \mathcal{N}(\boldsymbol{x}_n \mid \boldsymbol{\mu}_{(k)}, \boldsymbol{\Sigma}_{(k)})}{\sum_{k=1}^{K} w_k \mathcal{N}\boldsymbol{x}_n \mid \boldsymbol{\mu}_{(k)}, \boldsymbol{\Sigma}_{(k)})}$$

M-step

$$N_k = \sum_{n=1}^{N} p(z_n = k \mid \boldsymbol{x}_n)$$

$$w_k = \frac{N_k}{N}$$

$$\boldsymbol{\mu}_{(k)} = \frac{1}{N_k} \sum_{n=1}^{N} \boldsymbol{x}_n p(z_n = k \mid \boldsymbol{x}_n)$$

$$\boldsymbol{\Sigma}_{(k)} = \frac{1}{N_k} \sum_{n=1}^{N} (\boldsymbol{x}_n - \boldsymbol{\mu}_{(k)})(\boldsymbol{x}_n - \boldsymbol{\mu}_{(k)})^\top p(z_n = k|\boldsymbol{x}_n)$$

Quiz 02: GMM

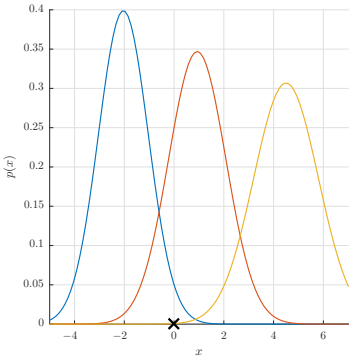


Figure 1: Mixture components in a GMM mixture model with $K = 3$

Consider a 1D GMM mixture model where each of the $K = 3$ (Gaussian) mixture components are illustrated in Figure 1 as the colored curves and the figure also shows a new observation indicated by the cross. Suppose we wish to apply the EM algorithm to this mixture model beginning with the E-step (i.e. assuming the mixture components has the means and variances indicated by Figure 1 and equal weights). According to the EM algorithm, what is the (approximate) probability the black cross is assigned to the blue (left-most) mixture component?

A. 0.05

B. 0.17

C. 0.25

D. 0.02

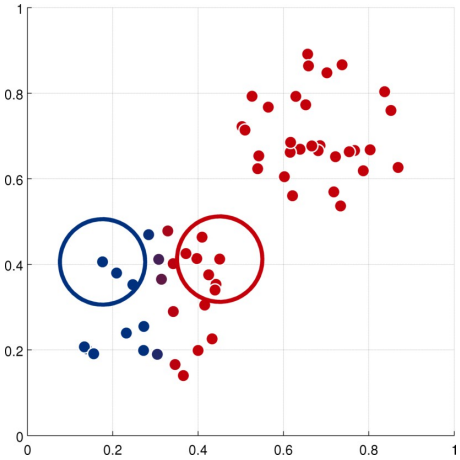
E. Don't know.

The EM algorithm for GMMs

Select an initial set of model parameters (mean and covariance for each cluster)
Repeat

- **Expectation**
  For each object, calculate the probability of belonging to each distribution
- **Maximization**
  For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change

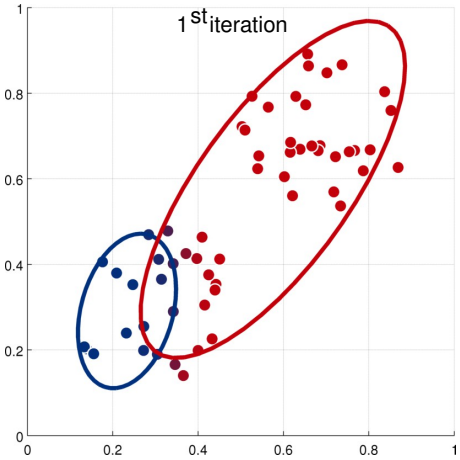


The EM algorithm for GMMs

Select an initial set of model parameters (mean and covariance for each cluster)
Repeat

- **Expectation**
  For each object, calculate the probability of belonging to each distribution
- **Maximization**
  For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change

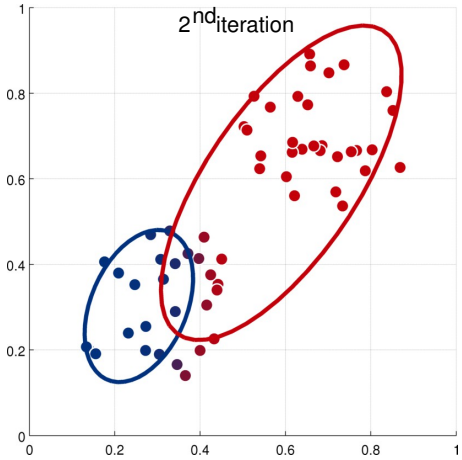


The EM algorithm for GMMs

Select an initial set of model parameters (mean and covariance for each cluster)
Repeat

- **Expectation**
  For each object, calculate the probability of belonging to each distribution
- **Maximization**
  For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change

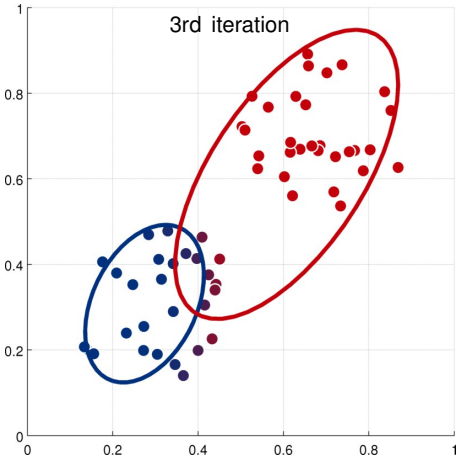


The EM algorithm for GMMs

Select an initial set of model parameters (mean and covariance for each cluster)
Repeat

- **Expectation**
  For each object, calculate the probability of belonging to each distribution
- **Maximization**
  For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change

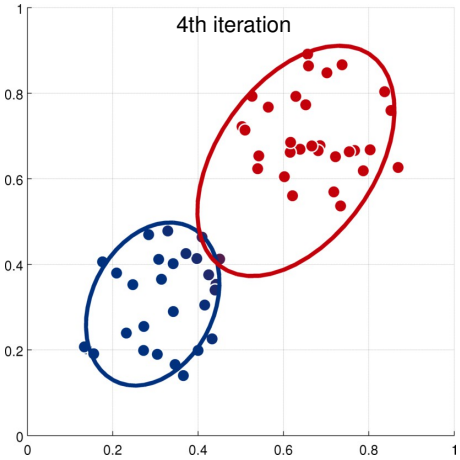


The EM algorithm for GMMs

Select an initial set of model parameters (mean and covariance for each cluster)
Repeat

- **Expectation**
  For each object, calculate the probability of belonging to each distribution
- **Maximization**
  For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change

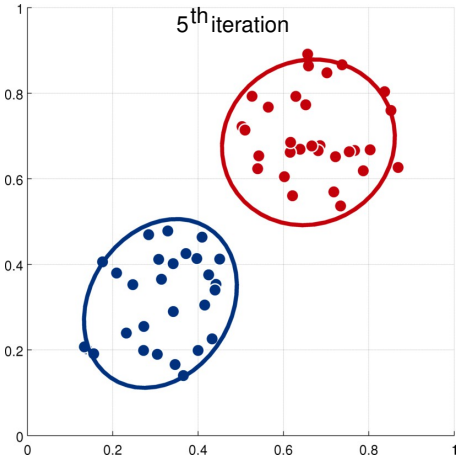


The EM algorithm for GMMs

Select an initial set of model parameters (mean and covariance for each cluster)
Repeat

- **Expectation**
  For each object, calculate the probability of belonging to each distribution
- **Maximization**
  For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change

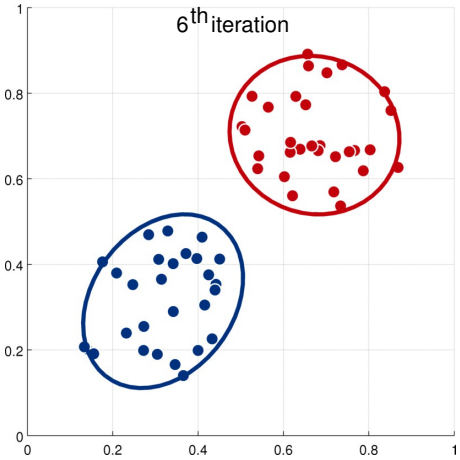


The EM algorithm for GMMs

Select an initial set of model parameters (mean and covariance for each cluster)
Repeat

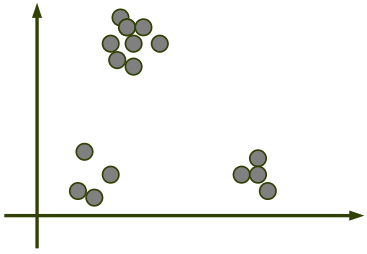
- **Expectation**
  For each object, calculate the probability of belonging to each distribution
- **Maximization**
  For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change

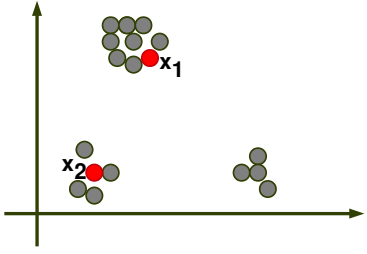


GMM - influence of K

- Consider the data to the right with 16 observations.
  - What would ideally happen if we used a GMM with K=16 clusters to model the data?

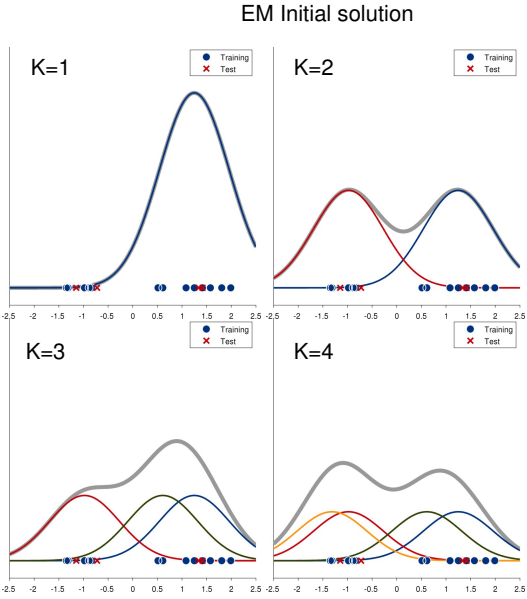


- Imagine we have two **test observations** denoted $\boldsymbol{x}_1$ and $\boldsymbol{x}_2$ (red points) that are not used for training.
  - What happens to $p(\boldsymbol{x}_1)$ and $p(\boldsymbol{x}_2)$ if we use K=3 and K=16 clusters?



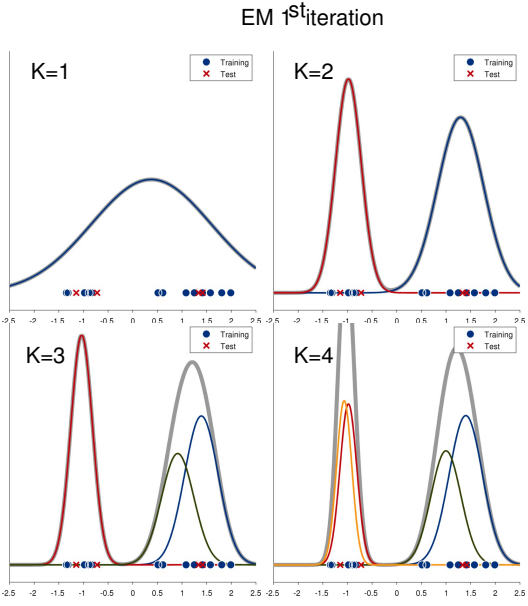
Mixture models

- Selecting complexity using crossvalidation



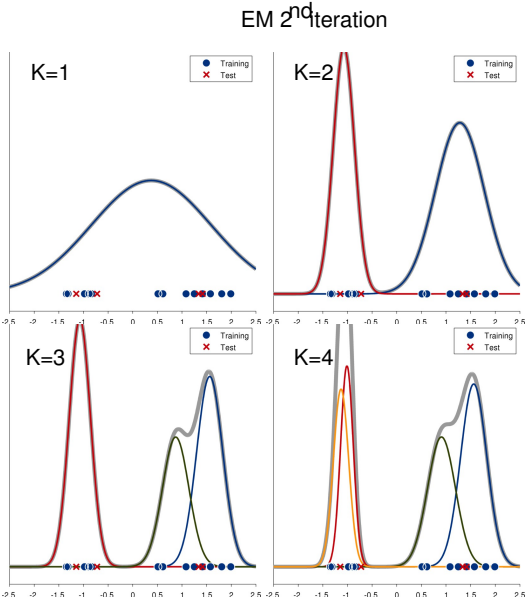
Mixture models

- Selecting complexity using crossvalidation



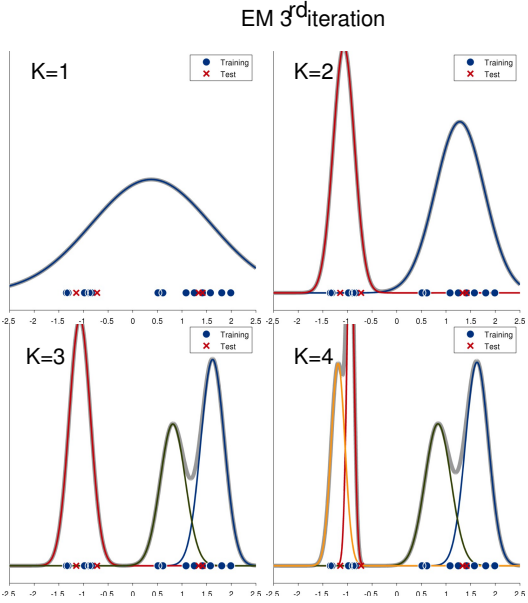
Mixture models

- Selecting complexity using crossvalidation



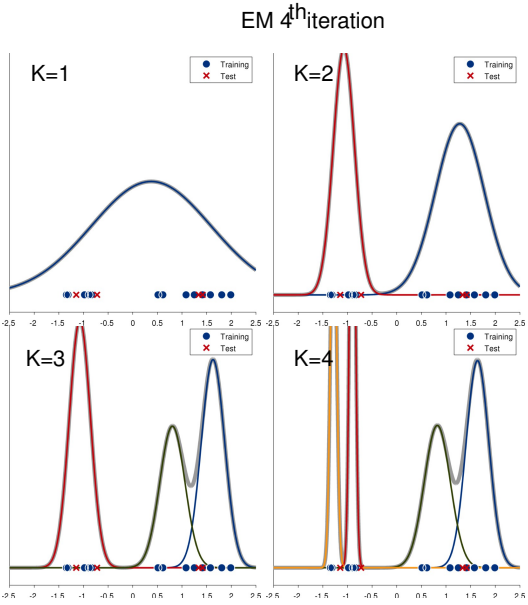
Mixture models

- Selecting complexity using crossvalidation



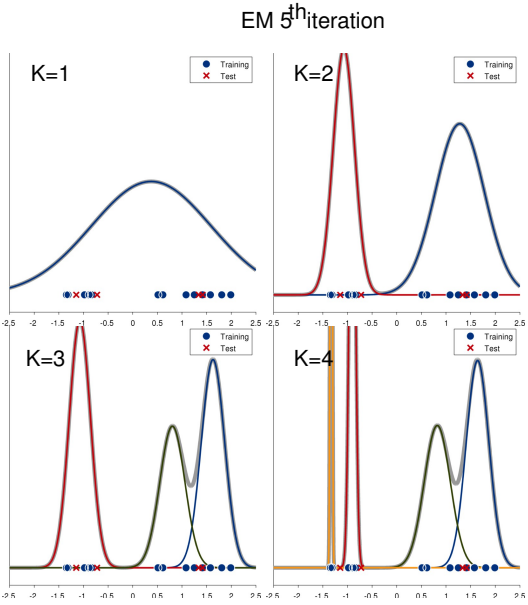
Mixture models

- Selecting complexity using crossvalidation



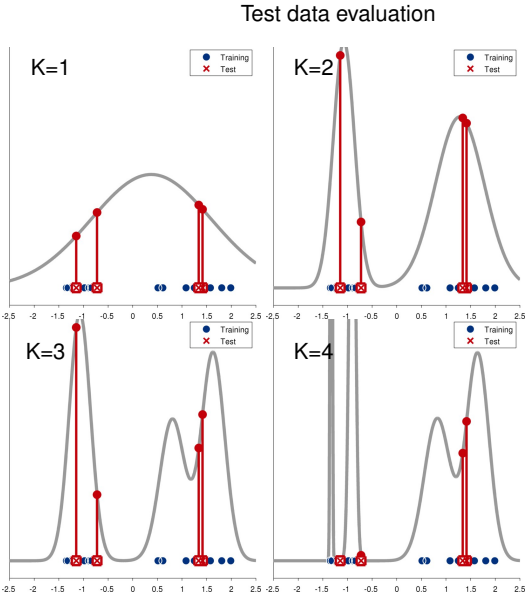
Mixture models

- Selecting complexity using crossvalidation



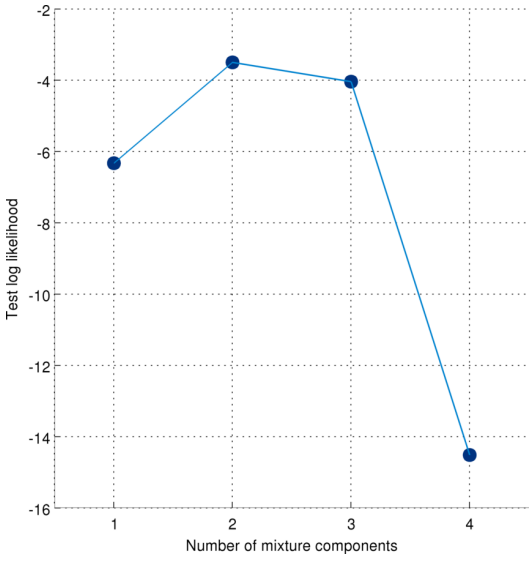
Mixture models

- Selecting complexity using crossvalidation



Mixture models

- Selecting complexity using crossvalidation

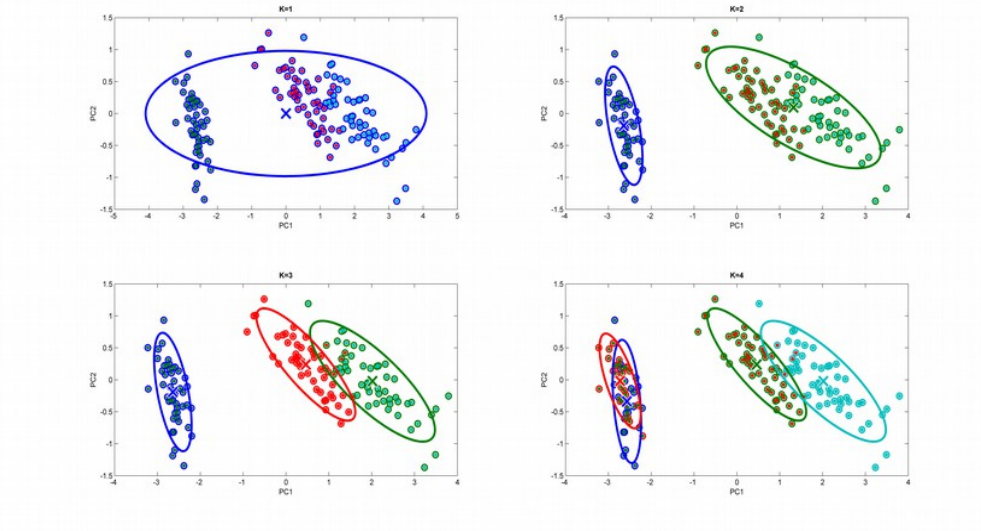


K-means versus GMM

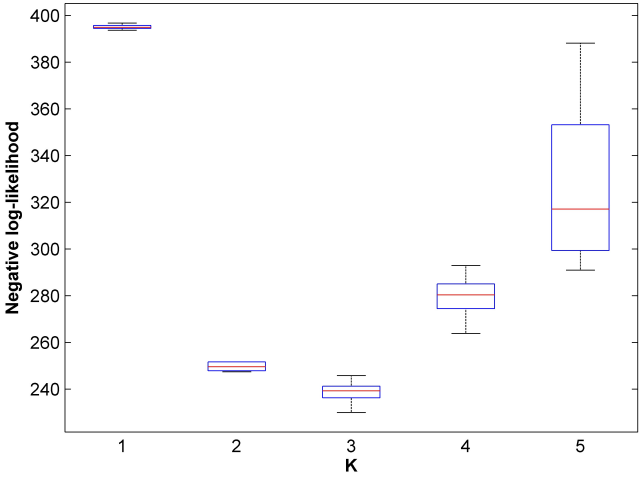
- No guarantee of optimal solution
- Does not model shape of clusters
- Does not model size of clusters
- Difficult to assess the number of clusters to use particularly when there is no ground truth

- No guarantee of optimal solution (even more local minima issues due to the additional model parameters)
- Models shape of cluster as ellipsoid
- Models the size of clusters
- Possible to estimate the number of components by cross-validation

Example: GMM on the Iris dataset



Example: GMM on Iris data



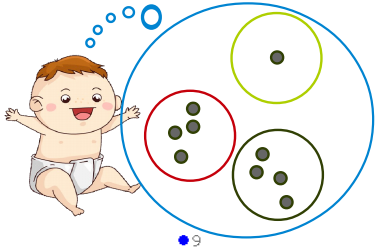
Anomaly detection: Definition

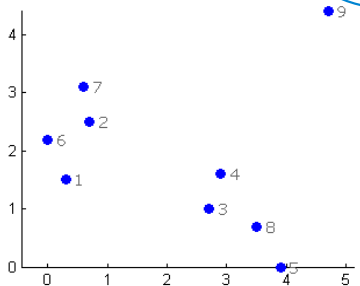
- Given a collection of data objects
  - Each object has associated a number of features
- Detect which objects **deviate from normal** behaviour

Anomaly detection: Examples

- Credit card **fraud detection**
  - Recognize dubious credit card transactions based on the transaction history of the card holder
- Network **intrusion detection**
  - Detect hacker attacks, web crawlers etc.
- **Ecosystem disturbances**
  - Detect hurricanes, floods droughts, heat waves and fires
- **Health and medicine monitoring**
  - Detect abnormal behaviour in populations and patients
- **Fault detection in industry systems**
  - Detect when a wind turbine performs poorly due to ice coating on blades
- Detection of **outliers** in data measurements
  - Remove erroneous measurements due to misreading from an instrument

Data example I: Cats, Dogs and Dinosaurs

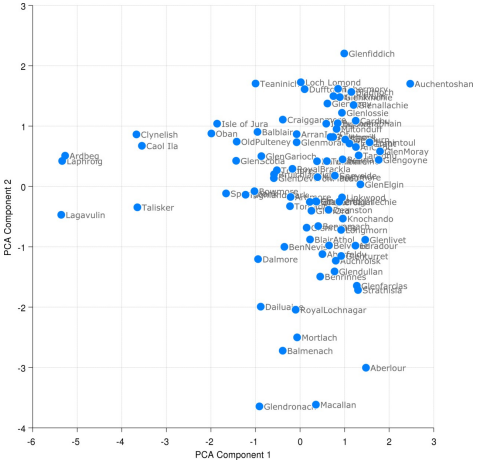




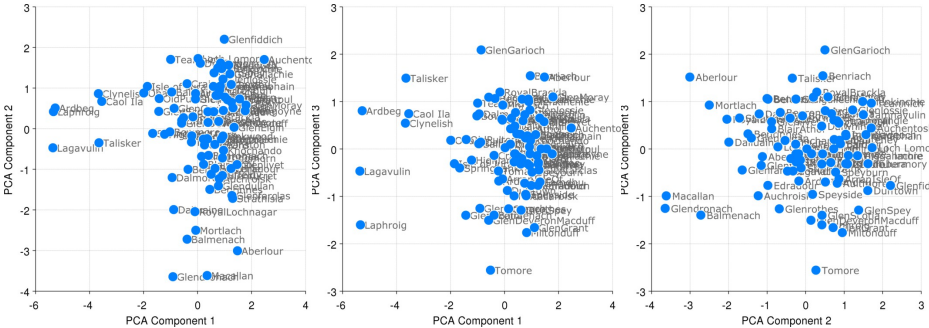
Whisky dataset

Data example II: Whisky

- 86 types of Scotch whisky
- Human ratings 1-5
- 12 taste categories
  - body, sweetness, smoky, medicinal, tobacco, honey, spicy, winey, nutty, malty, fruity, floral



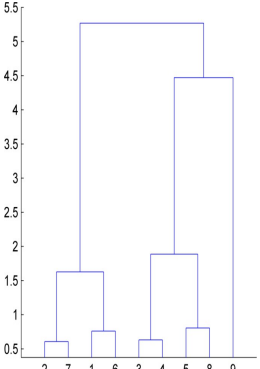
PCA - Whisky


PCA Component 2
PCA Component 1
Glenfiddich
Talisker
Ardbeg
Caol Ila
Clynelish
Lagavulin
Laphroig
Dalmore
Mortlach
Balmenach
Aberlour
Macallan
PCA Component 3
PCA Component 1
GlenGarioch
Tomore
PCA Component 3
PCA Component 2
Aberlour
Benriach
Mortlach
Auchentoshan
Loch Lomond
Glenfiddich
Macallan
Glendronnach
Balmenach
Auchroisk
Speyside
Dufftown
GlenSpey
Tomore

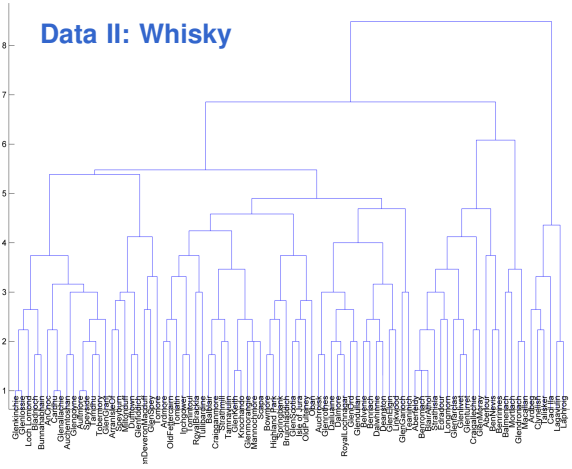
Dendrograms

- Dendrograms can be used to visualize relative distances between the observations

Data I: Cats, Dogs and Dinosaurs



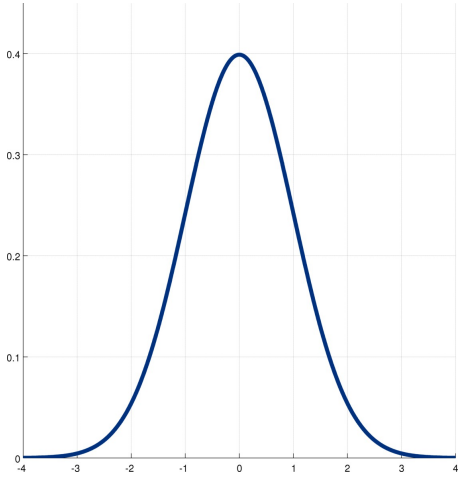
Data II: Whisky



Density based techniques: Univariate normal distribution

- Map attribute to standard normal variable.
  $z = \frac{x-\mu}{\sigma}$
- Choose a threshold:

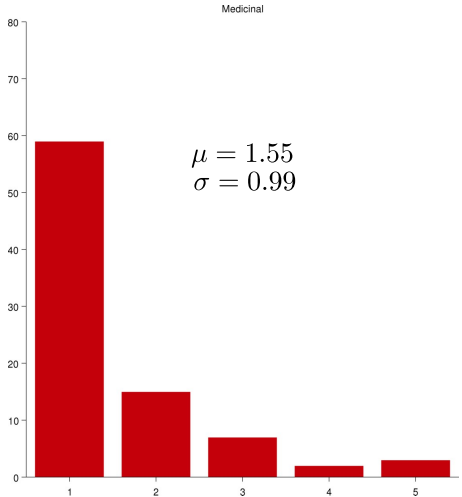
c	p(|z|>c)
1.0	0.3173
1.5	0.1336
2.0	0.0455
2.5	0.0124
3.0	0.0027
3.5	0.0005
4.0	0.0001



Density based techniques: Univariate normal distribution

- Map attribute to standard normal variable.
  $z = \frac{x-\mu}{\sigma}$
- Choose a threshold:

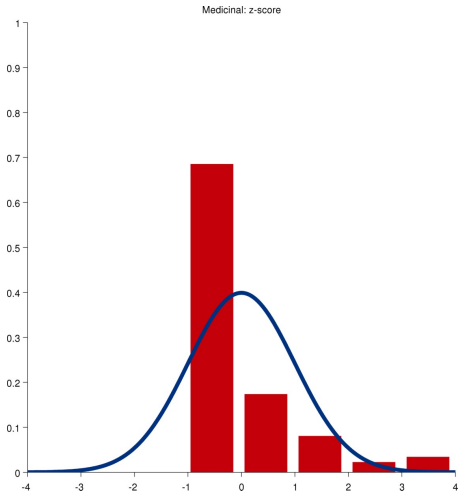
c	p(|z|>c)
1.0	0.3173
1.5	0.1336
2.0	0.0455
2.5	0.0124
3.0	0.0027
3.5	0.0005
4.0	0.0001



Density based techniques: Univariate normal distribution

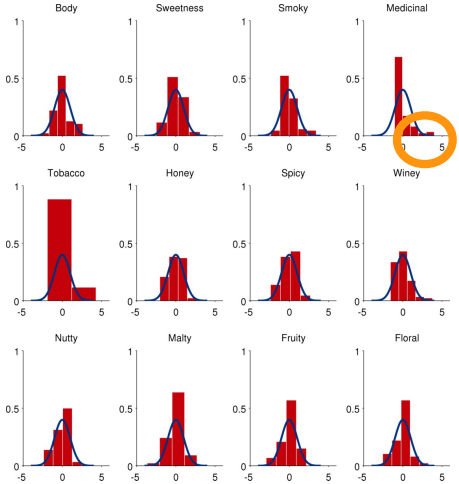
- Map attribute to standard normal variable.
  $z = \frac{x-\mu}{\sigma}$
- Choose a threshold:

c	p(|z|>c)
1.0	0.3173
1.5	0.1336
2.0	0.0455
2.5	0.0124
3.0	0.0027
3.5	0.0005
4.0	0.0001



Density based techniques: Univariate normal distribution

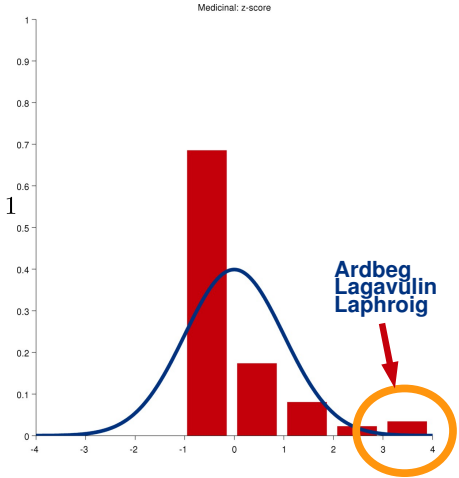
- Map attribute to standard normal variable.
  $z = \frac{x-\mu}{\sigma}$
- Choose a threshold:
  $p(|z| > c) = 0.001$
  $c = 3.2905$



Note: Assumes attributes follow a normal distribution which may not be a valid assumption!

Density based techniques: Univariate normal distribution

- Map attribute to standard normal variable.
  $z = \frac{x-\mu}{\sigma}$
- Choose a threshold:
  $p(|z| > c) = 0.001$
  $c = 3.2905$



Approaches to anomaly detection

- **Density-based techniques**
  1) Estimate the density of data objects
  2) Outliers are: Data objects in low density area

  Densities:
  - 1D normal (as demonstrated)
  - Multivariate normal
  - GMM to evaluate the density of test data.
  - Kernel density estimation

- **Distance-based techniques**:
  - Basic distance
  - Inverse average distance to K nearest neighbours (KNN density)
  - Average relative KNN density

Kernel Density Estimator

Kernel Density estimation with Gaussian **kernel**:

* Consider the GMM with $K = N^{train}$ components (i.e. as many component as traning points)
* For each component:
  - Set the mean $\boldsymbol{\mu}_k = \boldsymbol{x}_n$, i.e., each mean is given by a training point
  - Set the covarince $\boldsymbol{\Sigma} = \sigma^2 \boldsymbol{I}$, i.e., a fixed varince across all components
  - Set $w_k = \frac{1}{N^{train}}$, i.e., all components/observations have equal weight

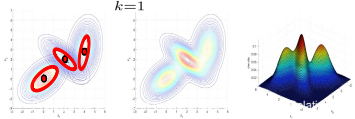
$$p(x) = \sum_{k=1}^{N^{train}} \frac{1}{N^{train}} \mathcal{N}(\boldsymbol{x} \mid \boldsymbol{\mu}_k = \boldsymbol{x}_k, \boldsymbol{\Sigma}_k = \sigma^2 \boldsymbol{I})$$

Recall the Gaussian Mixture Model(GMM)

Data density — Sum of cluster specific densities

$$p(\boldsymbol{x}) = \sum_{K=1}^{K} w_k \mathcal{N}(\boldsymbol{x} | \boldsymbol{\mu}_{(k)}, \boldsymbol{\Sigma}_{(k)})$$

$$(s.t. \sum_{k=1}^{K} w_k = 1, \quad w_k \geq 0)$$



$\boldsymbol{\mu}_{(k)}$: Cluster center (prototypical example in cluster)

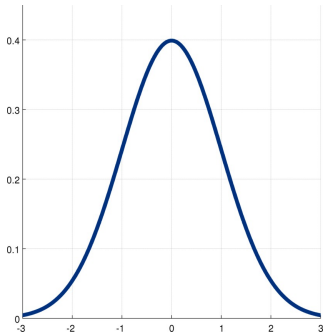
$\boldsymbol{\Sigma}_{(k)}$: Shape of the cluster

w_k : Relative density of the cluster

Quiz 03: Kernel density

$$p(x \mid \mu, \sigma) = \frac{1}{\sqrt{2\pi\sigma^2}} exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right)$$

.



Consider five observations of an attribute x given by

$$\boldsymbol{X} = \{2, 3, 5, 10, 12\}.$$

Based on the five observations, what is the Gaussian kernel density estimate at $x = 4$ using $\sigma^2 = 4$?

A. $\frac{1}{\sqrt{8\pi}^5} \exp(-\frac{53}{4})$

B. $\frac{1}{5\sqrt{8\pi}} \exp(-\frac{53}{4})$

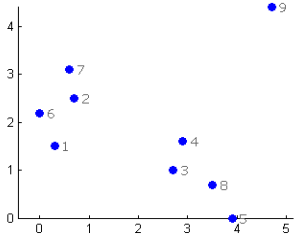
C. $\frac{1}{5\sqrt{8\pi}}(\exp(-\frac{1}{2}) + 2 \cdot \exp(-\frac{1}{8}) + \exp(-\frac{9}{2}) + \exp(-8))$

D. $\frac{1}{5\sqrt{8\pi}}(\exp(-1) + 2 \cdot \exp(-\frac{1}{4}) + \exp(-9) + \exp(-16))$

E. Don't know.

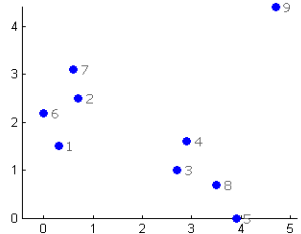
KDE: How do we determine σ^2?

Data I: Cats, Dogs and Dinosaurs

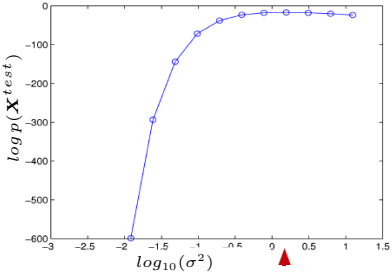


KDE: How do we determine σ^2?

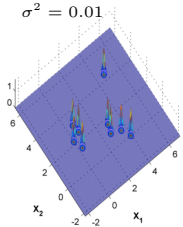
Data I: Cats, Dogs and Dinosaurs

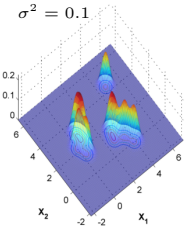


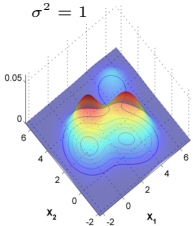
Density of test set based on leave-one-out cross validation

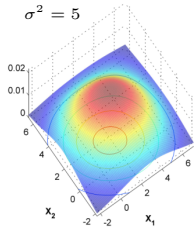


Optimal $\sigma^2 = 1.55$



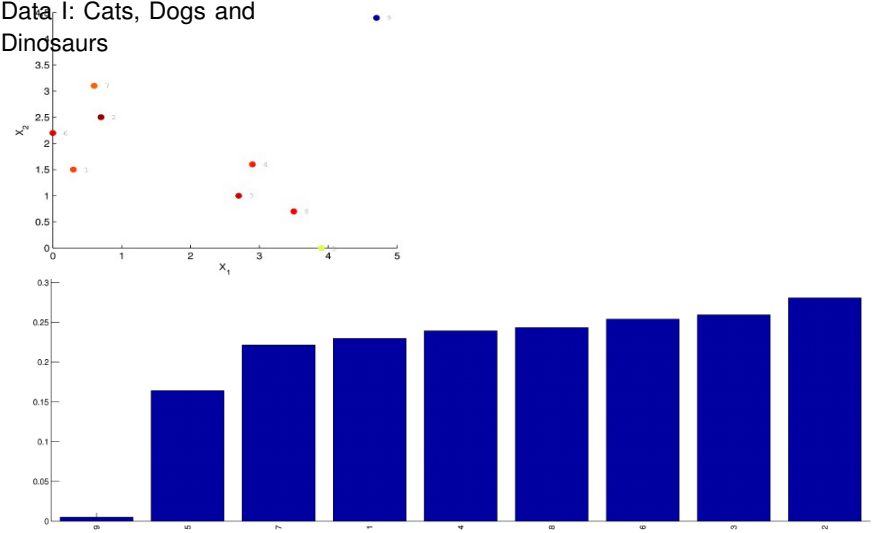






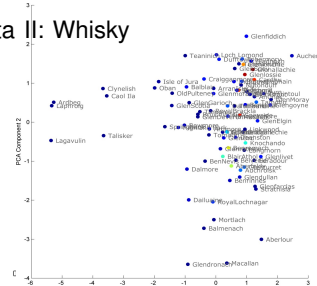
KDE: Leave-one-out density evaluated at each point

Data I: Cats, Dogs and Dinosaurs

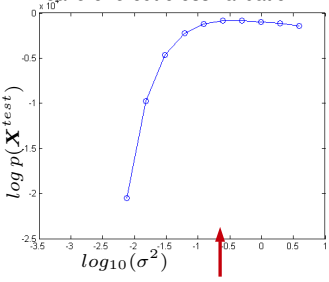


KDE: Leave-one-out density evaluated at each point

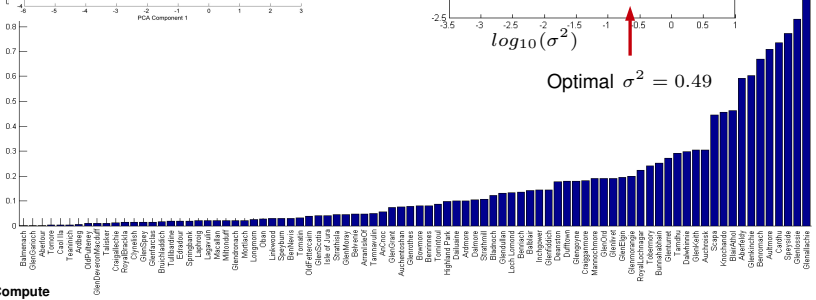
Data II: Whisky



Density of test set based on leave-one-out cross validation

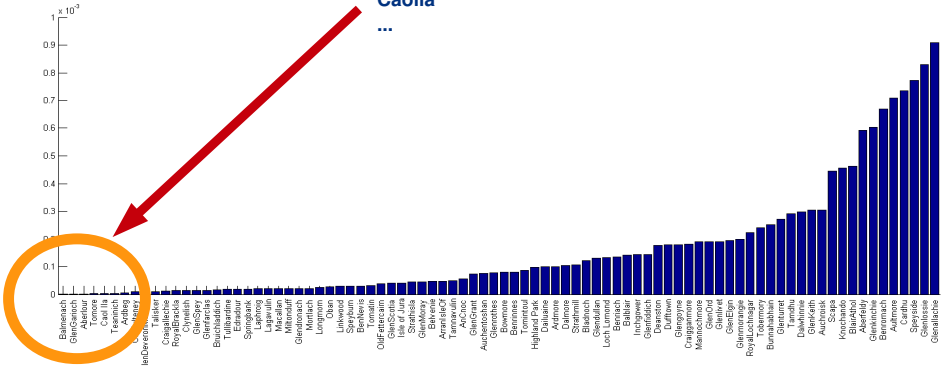


Optimal $\sigma^2 = 0.49$



Data II: Whisky

Balmenach
Glen Garioch
Aberlour
Tomore
Caolla
...



Inverse distance density estimation

- **Distance based measure of density**
  - Density is inverse proportional to avereage distance to k nearest neighbors
  - Density is low if nearest neighbors are far away

$$\text{density}_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K) = \frac{1}{\frac{1}{K}\sum_{\boldsymbol{x}' \in N_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K)} d(\boldsymbol{x}_i, \boldsymbol{x}')}$$

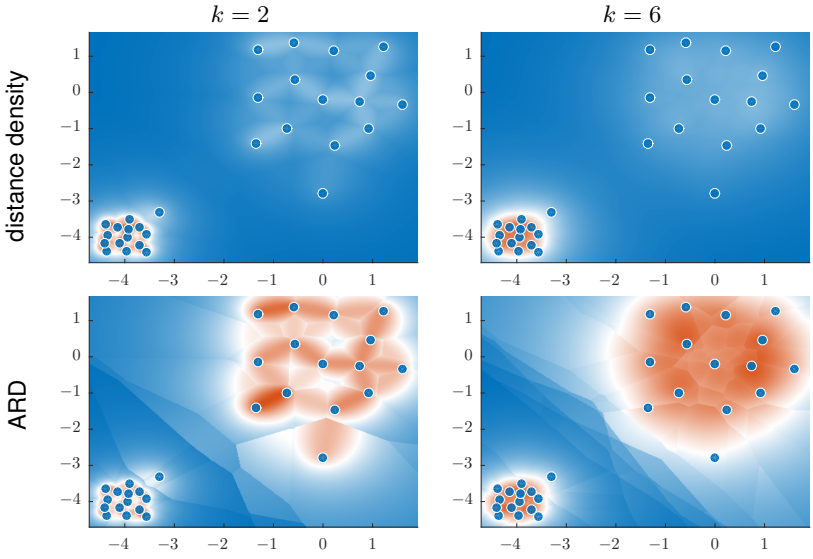
- **Relative density**
  - Density compared to density at nearest neighbors

$$\text{ard}_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K) = \frac{\text{density}_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K)}{\frac{1}{K}\sum_{\boldsymbol{x}_j \in N_{\boldsymbol{X}_{\setminus i}}(\boldsymbol{x}_i, K)} \text{density}_{\boldsymbol{X}_{\setminus j}}(\boldsymbol{x}_j, K)}$$

$N_{\boldsymbol{X}}(\boldsymbol{x}, K) = \{\text{The K observations in } \boldsymbol{X} \text{which are nearest to } \boldsymbol{x}\}$

$\boldsymbol{X}_{\setminus i}^\top = [\boldsymbol{x}_1\ \boldsymbol{x}_2\ \cdots \boldsymbol{x}_{i-2}\ \boldsymbol{x}_{i-1}\ \boldsymbol{x}_{i+1}\ \boldsymbol{x}_{i-2}\ \boldsymbol{x}_N]$

These measures are taken from: “Introduction to Data Mining” by Pang-Ning Tan, Michael Steinbach, Vipin Kumar


$k = 2$
$k = 6$
distance density
ARD
1
0
−1
−2
−3
−4
−4
−3
−2
−1
0
1

Quiz 04: ARD

	O1	O2	O3	O4	O5	O6	O7	O8	O9	O10
O1	0	393.5	68.1	165.4	271.8	200.6	210.9	206.1	166.3	365.0
O2	393.5	0	411.3	361.8	478.6	490.9	409.2	382.3	391.1	37.4
O3	68.1	411.3	0	119.8	208.4	136.6	152.8	154.3	111.1	387.1
O4	165.4	361.8	119.8	0	137.5	130.8	62.1	44.7	32.5	346.2
O5	271.8	478.6	208.4	137.5	0	99.0	76.8	101.0	116.4	468.5
O6	200.6	490.9	136.6	130.8	99.0	0	100.1	124.0	100.5	473.8
O7	210.9	409.2	152.8	62.1	76.8	100.1	0	29.5	45.2	396.8
O8	206.1	382.3	154.3	44.7	101.0	124.0	29.5	0	44.6	370.1
O9	166.3	391.1	111.1	32.5	116.4	100.5	45.2	44.6	0	375.1
O10	365.0	37.4	387.1	346.2	468.5	473.8	396.8	370.1	375.1	0

Table 1: Pairwise Euclidean distance between the 10 first observations in the PM10 (air polution) dataset. Red/green observations corresponds to high/low pollution levels.

We suspect that observation O1 in Table 1 may be an outlier. In order to assess if this is the case we will calculate the average relative density (ARD) based on the distances in the table using the definitions:

$$\text{density}(\boldsymbol{x}, K) = \left(\frac{1}{K}\sum_{\boldsymbol{y}\in N(\boldsymbol{x},K)} distance(\boldsymbol{x},\boldsymbol{y})\right)^{-1},$$

$$\text{a.r.d.}(\boldsymbol{x}, K) = \frac{\text{density}(\boldsymbol{x}, K)}{\frac{1}{K}\sum\limits_{\boldsymbol{y}\in N(\boldsymbol{x},K)} \text{density}(\boldsymbol{y}, K)},$$

where $N(\boldsymbol{x}, K)$ is the set of K nearest neighbors of observation $\boldsymbol{x}$ and a.r.d$(\boldsymbol{x}, K)$ is the average relative density of $\boldsymbol{x}$ using K nearest neighbors. What is ARD for observation O1 for $K = 2$ nearest neighbors?

A. 0.01

B. 0.02

C. 0.23

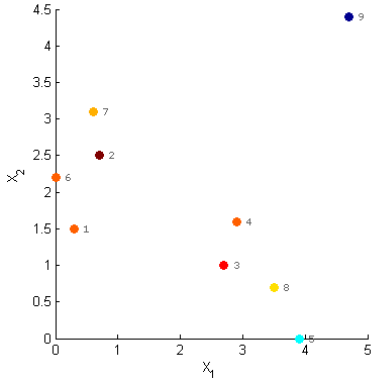
D. 0.46

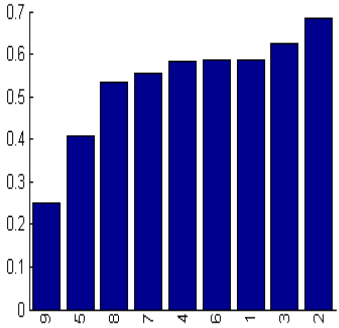
E. Don't know.

Inverse distance density estimation

- KNN density (5 nearest neighbors)

Data I: Cats , dogs and dinosaurs



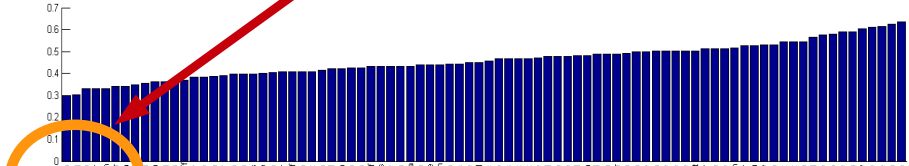


Inverse distance density estimation

- KNN density (5 nearest neighbors)

Data II: Whisky

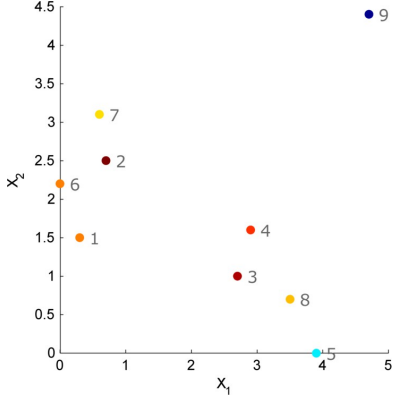
Balmenach
Laphroig
GlenGarioch
Aberlour
Lagevulin
...

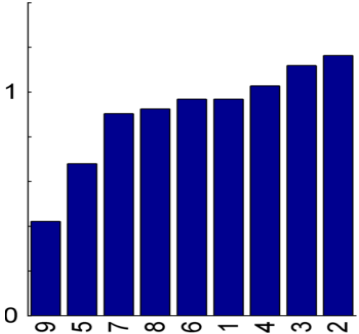


Average Relative density

- Average relative KNN density (5 nearest neighbors)

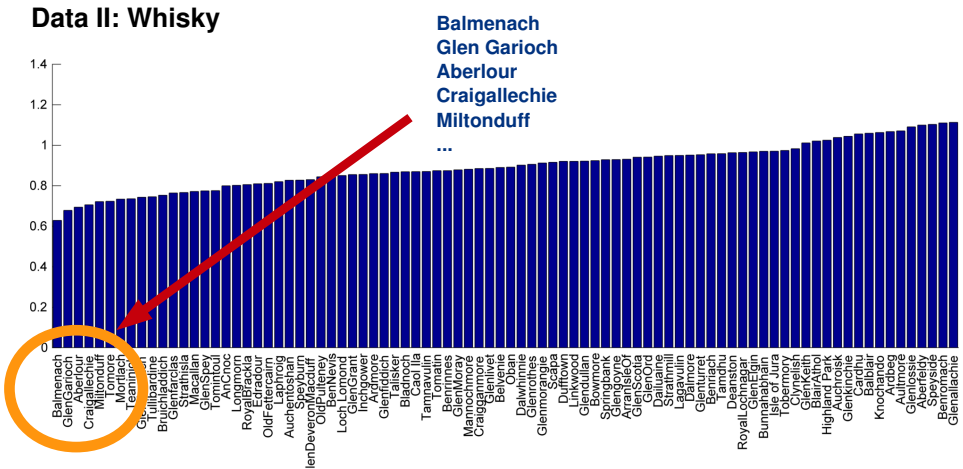
Data I: Cats , dogs and dinosaurs





Average Relative density

- Average relative KNN density (5 nearest neighbors)



Results using different methods

- Kernel Density Estimation
  - Balmenach
  - Glen Garioch
  - Aberlour
  - Tomore
  - Caolla
- Distance density
  - Balmenach
  - Laphroig
  - Glen Garioch
  - Aberlour
  - Lagavulin
- Average relative density
  - Balmenach
  - Glen Garioch
  - Aberlour
  - Craigallechie
  - Miltonduff
- **Common**: Balmenach, Glen Garioch, Aberlour

Summary

- **Density modelling**
  Idea: *Model* $p(x)$
  - Normal (aka Gaussian) distribution
  - Gaussian mixture model
    - EM algorithm for find $\{w_k, \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k\}^{k=1:K}$
    - Cross-validation used to select K based on likelihood on held out dataset.
    - **Applications:** Density estimation/analysis/visulizaiton, anomaly detection, clustering (indirectly).
  - Kernel density estimation estimator - *a special mixture model*

    ($\boldsymbol{\mu}_k$ and K given by traning points). Cross-validation for selecting σ based on likelihood on held out data (easy as **no retraining** needed).
    - **Applications:** Density estimation/analysis/visualizaiton, anomaly detection.
- **Anomaly detection**
  Idea: *Consider which observations are abnormal*
  - Density-based (simple univariate distrbutions, GMMs, KDEs)
    Observations with low (test) likelihood are potential outliers
  - Distance-based (Inverse distance density, average relative density)

Resources

https://www.youtube.com Nice explanation of expectation maximization for the Guassian Mixture Model (https://www.youtube.com/watch?v=WaKNSBeDLTw)