

02450: Introduction to Machine Learning and Data Mining

Mixture models and density estimation

Bjørn Sand Jensen

DTU Compute, Technical University of Denmark (DTU)

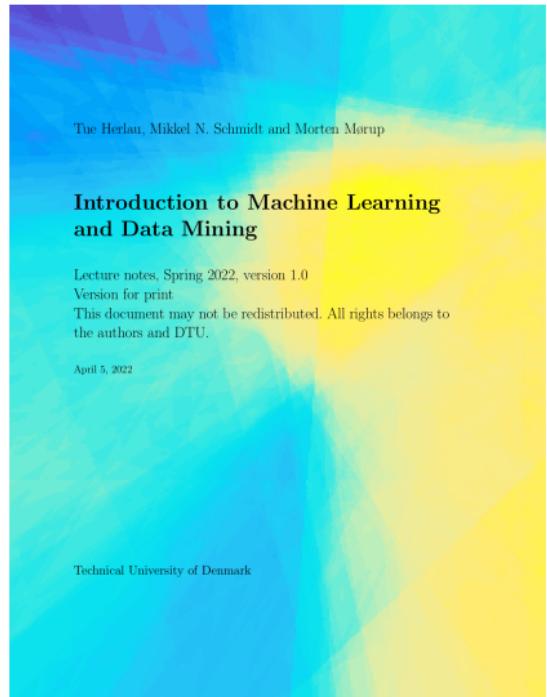
Today Remember you can evaluate the course+TAs on DTU Inside

Feedback Groups of the day:

Beatriz Saragoça, Carina Adriana Anna Schoenmaker,
Astrid Lund Paulsen, Signe Djernis Olsen, Sofie
Kodal Larsen, Eline Dorothea Siegumfeldt, Emil Gejl
Lauritzen, Christian Mundbjerg Nykjær, Lukas
Kofoed Petersen, Josefine Grønset Løken, Nojan
Rezvani, Laura Figueiredo Tor, Ana Pastor
Mediavilla, Bastian Gulstad Ravn, Aleksander Wind,
Usha Thapa, Laurits Munck, Sarah Sophie Simon,
Diogo Matos Da Silva, Jordan Shopp, Natalie Valett,
Ka Chun So, Karolina Zofia Tudelska, Adam Bo
Mandel, Konrad Kacper Uscilo, Sofie Rømø, Bára
Savcicová, Lotta Marie Schlau, Arqum Shahid, Ali
Houssein Yousef Shanoof, Ólafur Thorri Sigurjónsson,
Matija Sipek, Lucas Loft Skals, Patriks Mikus
Sniedze, Rasmus Saugmann Sørensen

Reading material:

Chapter 19, Chapter 20



Lecture Schedule

1 Introduction

31 January: C1

Data: Feature extraction, and visualization

2 Data, feature extraction and PCA

7 February: C2, C3

3 Measures of similarity, summary statistics and probabilities

14 February: C4, C5

4 Probability densities and data visualization

21 February: C6, C7

Supervised learning: Classification and regression

5 Decision trees and linear regression

28 February: C8, C9

6 Overfitting, cross-validation and Nearest Neighbor

7 March: C10, C12 (Project 1 due before 13:00)

7 Performance evaluation, Bayes, and Naive Bayes

14 March: C11, C13

8 Artificial Neural Networks and Bias/Variance

21 March: C14, C15

9 AUC and ensemble methods

28 March: C16, C17

Unsupervised learning: Clustering and density estimation

10 K-means and hierarchical clustering

11 April: C18

11 Mixture models and density estimation

18 April: C19, C20 (Project 2 due before 13:00)

12 Association mining

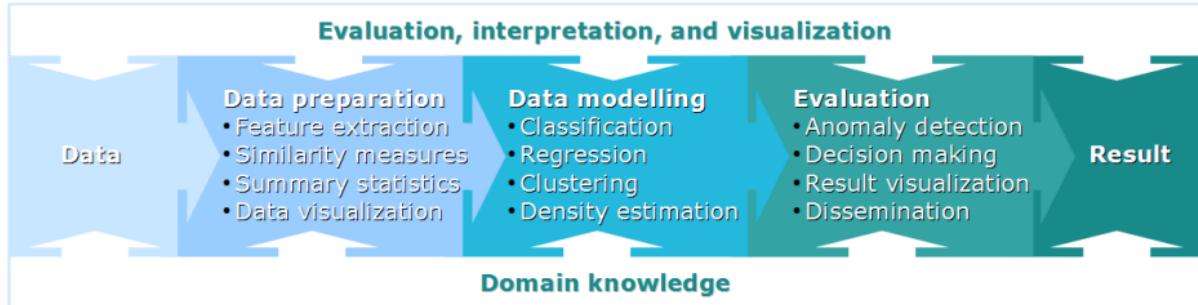
25 April: C21

Recap

13 Recap and discussion of the exam

2 May: C1-C21

Online 24/7 help: Discussion Forum/Piazza
Streaming & Videos: <https://panopto.dtu.dk/>
Online exercises: MS Teams



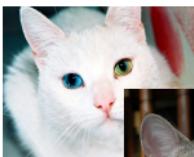
Learning Objectives

- 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 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!



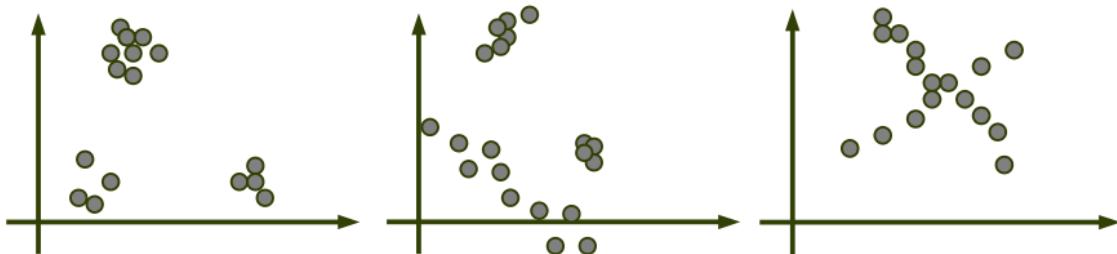
<http://www.clipartlord.com/category/baby-clip-art/>



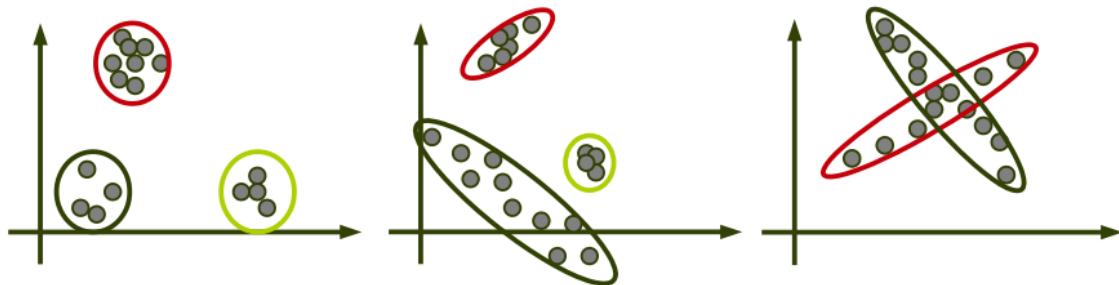
We humans are skilled at dividing objects into groups (clustering), but how do we make computers do the same?

http://commons.wikimedia.org/wiki/File:Abessinier_sorrel.jpg
http://commons.wikimedia.org/wiki/File:Cat_Eyes.jpg
http://commons.wikimedia.org/wiki/File:Black_white_cat_on_fence.jpg
http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg
<http://commons.wikimedia.org/wiki/File:MasPrl-Astro-SVE.jpg>
http://commons.wikimedia.org/wiki/File:GermanShorthPt_wb.jpg
<http://commons.wikimedia.org/wiki/File:Cat002.jpg>
<https://commons.wikimedia.org/w/index.php?title=Dog&oldid=5100000>
<http://commons.wikimedia.org/wiki/File:Saurier2.jpg>

- 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?



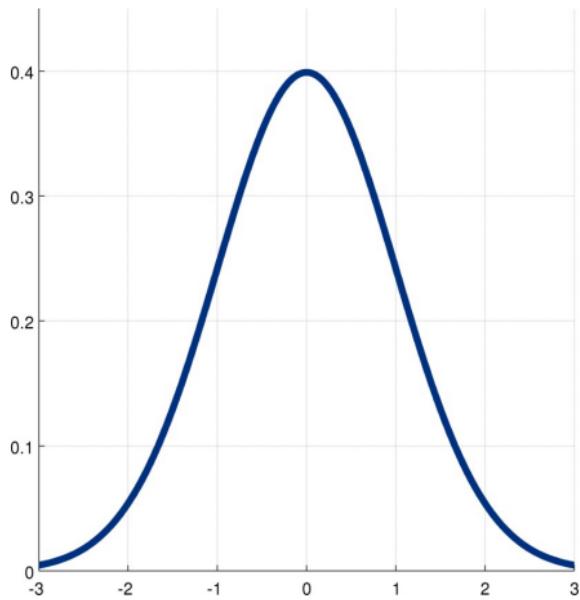
- What are the clusters below and what characterize each cluster?
- Is **k-means** well suited for modeling the clusters below?
 - Will it always find the optimum solution?
 - Can it model the sizes of the clusters?
 - Can it model the shape of the clusters?
 - How can we determine the number of clusters?



Normal distribution

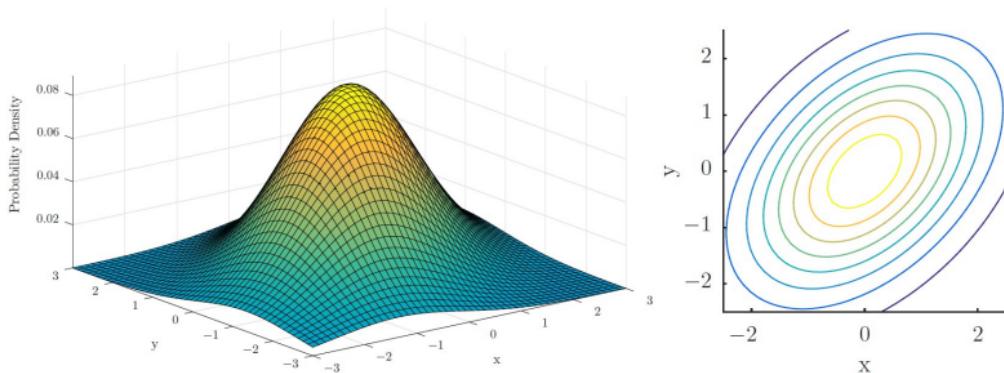
- Probability density function describes the relative chance of a given value to occur
- Normal distribution characterized by
 - Mean
 - Variance

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



Multivariate Normal distribution

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{\sqrt{(2\pi)^d |\Sigma|}} e^{-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu)}$$



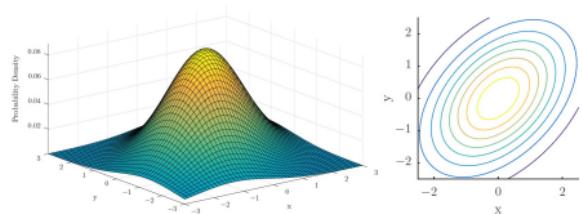
Multivariate Normal distribution

$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{\sqrt{(2\pi)^d |\boldsymbol{\Sigma}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}-\boldsymbol{\mu})}$$

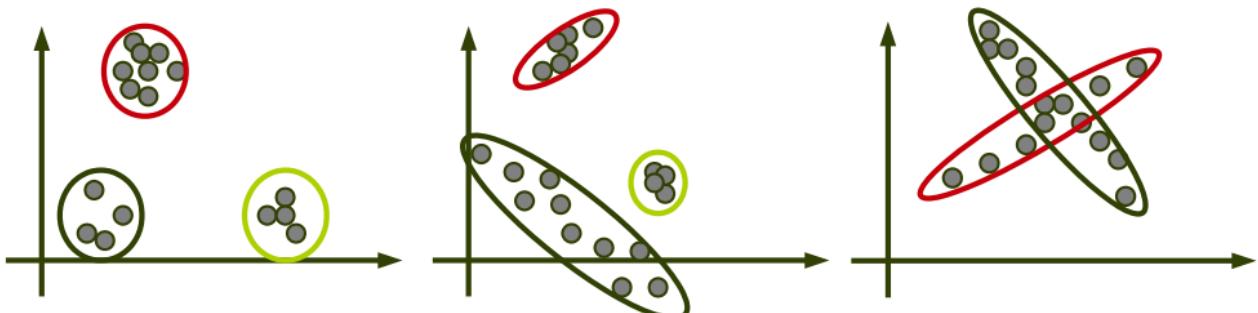
- Example: 2-dimensional Normal distribution

$$\boldsymbol{\mu} = \begin{bmatrix} \bar{x}_1 \\ \bar{x}_2 \end{bmatrix}$$

$$\boldsymbol{\Sigma} = \begin{bmatrix} \text{var}(x_1) & \text{cov}(x_1, x_2) \\ \text{cov}(x_2, x_1) & \text{var}(x_2) \end{bmatrix}$$



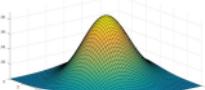
Prototypical mixture model



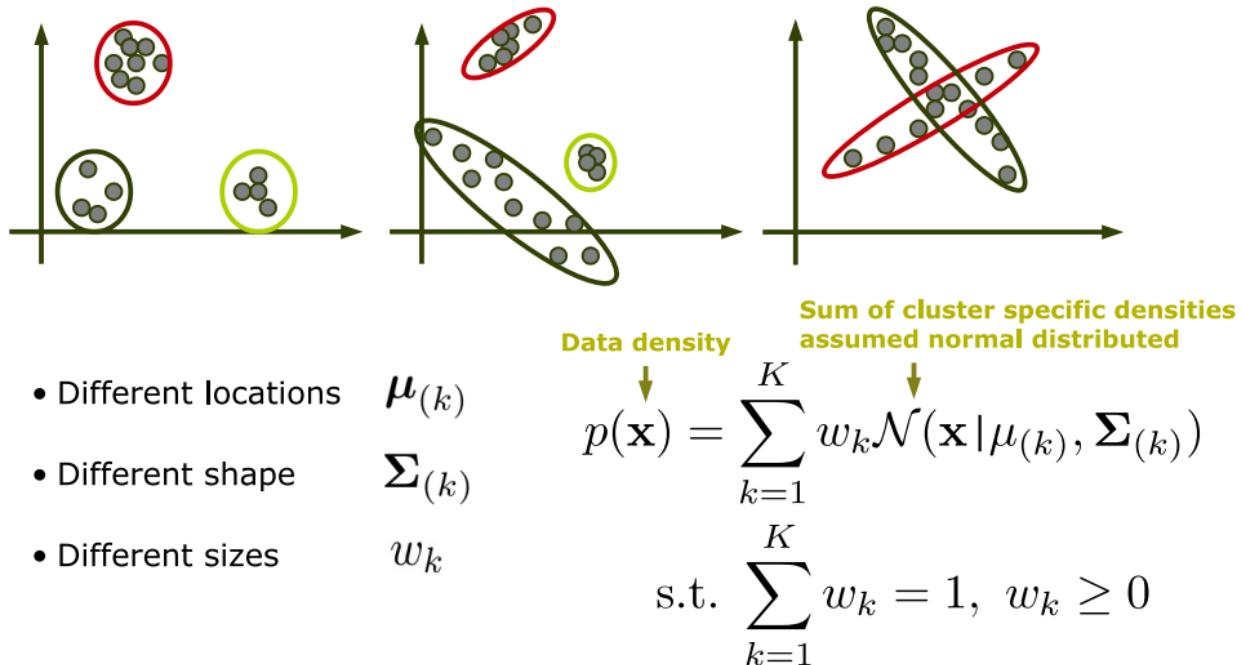
- We want a **density** $p(\mathbf{x})$ of our observations $\mathbf{x} \in \mathbb{R}^M$
- Suppose we have K clusters and let $z = k$ if \mathbf{x} belongs to cluster k
- According to the basic rules of probability:

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

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

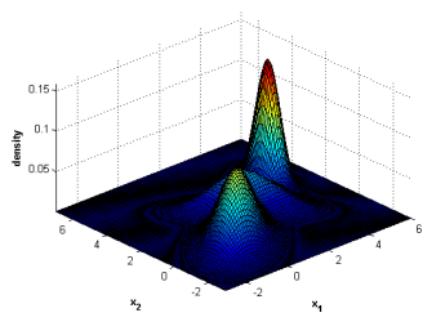
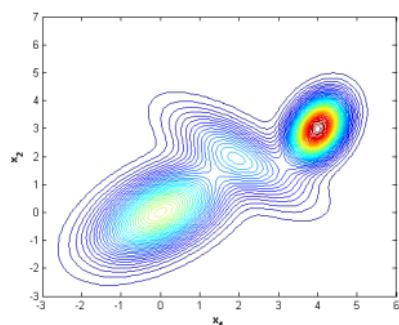
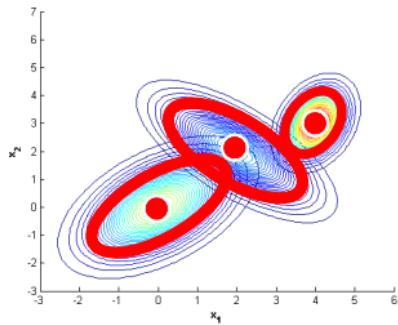


The Gaussian Mixture Model (GMM)



GMM example

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



$\mu_{(k)}$: Cluster center (prototypical example in cluster)

$\Sigma_{(k)}$: Shape of the cluster

w_k : Relative size/density of the cluster

Quiz 01: GMM

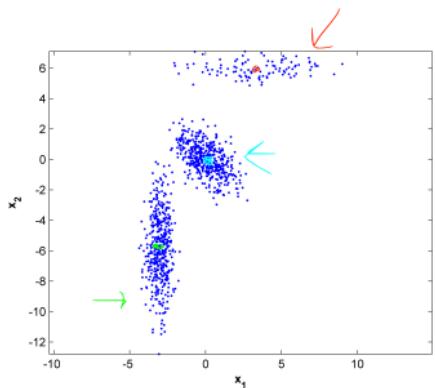


Figure 1: 1000 data observations drawn from a Gaussian Mixture Model (GMM) with three clusters.

In Figure 1 is shown 1000 observations drawn from a density defined by a Gaussian Mixture Model (GMM) with three clusters. Suppose

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

is the multivariate normal distribution, which one of the following GMM densities was used to generate the data?

Størrelse på cluster (antal punkter)
mu
Varians/kovarians.
Varians: spredning på x og y akse
Kovarians: hældning

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

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

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

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

E. Don't know.

Sanity check time:

- Consider the Gaussian mixture model (GMM)

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

- What is the value of the integral?

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

Gaussian mixture models, EM algorithm

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

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

M-step

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

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

$$w_k = \frac{N_k}{N} = p(z = k)$$

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

Quiz 02: GMM

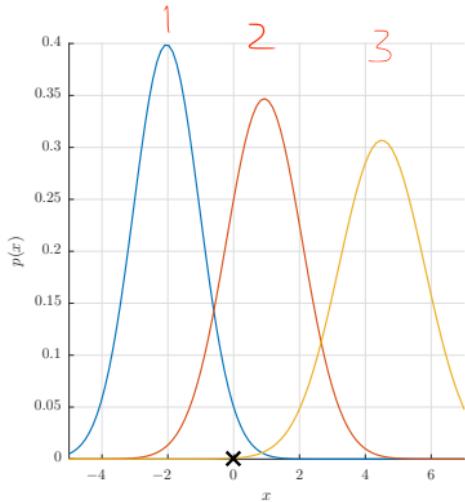


Figure 1: Mixture components in a GMM mixture model with $K = 3$

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.

Gaussian mixture models, EM algorithm

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

Repeat

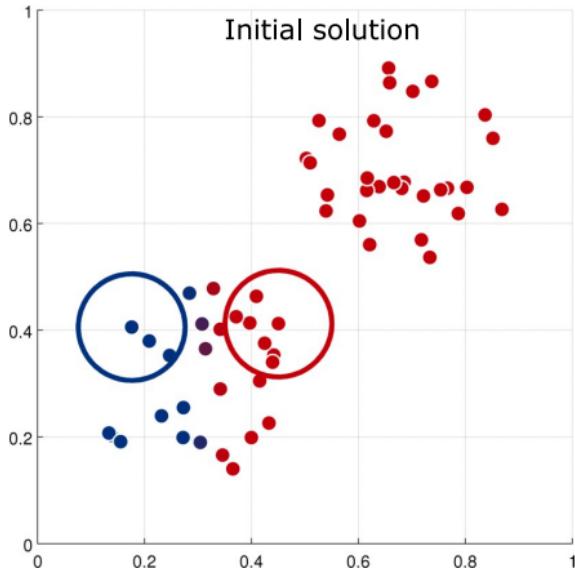
- **Expectation**

- 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



Gaussian mixture models, EM algorithm

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

Repeat

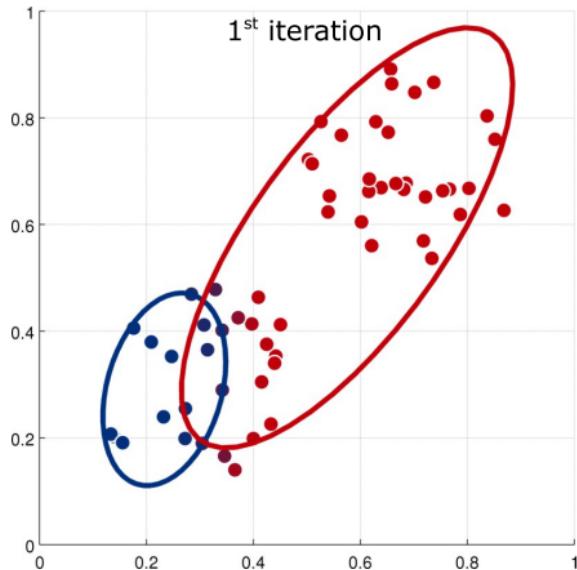
• Expectation

- 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



Gaussian mixture models, EM algorithm

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

Repeat

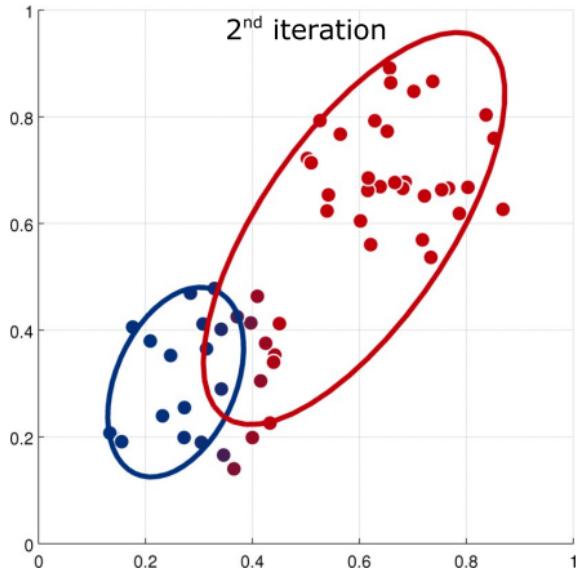
- **Expectation**

- 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



Gaussian mixture models, EM algorithm

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

Repeat

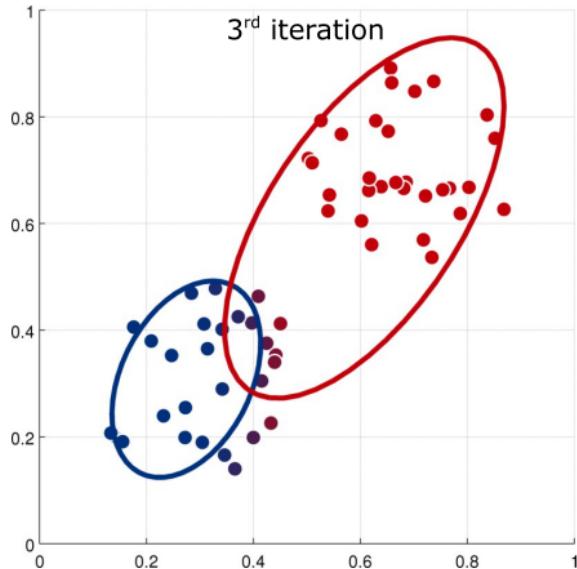
• Expectation

- 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



Gaussian mixture models, EM algorithm

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

Repeat

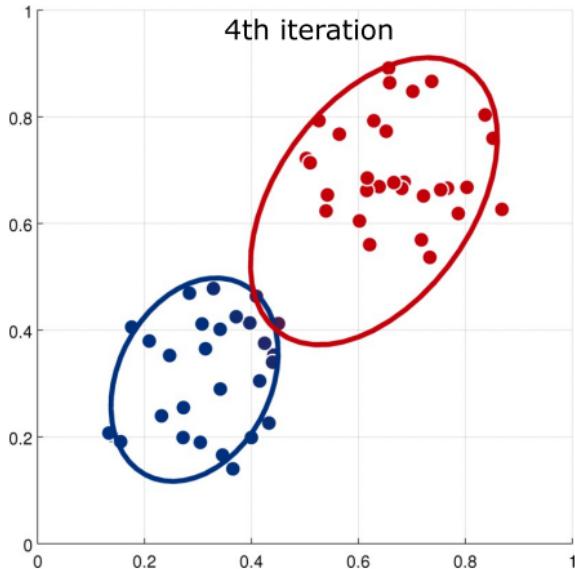
- **Expectation**

- 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



Gaussian mixture models, EM algorithm

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

Repeat

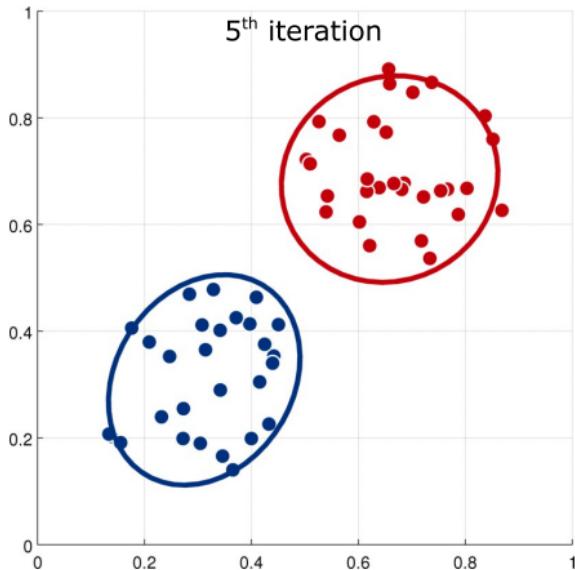
- **Expectation**

- 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



Gaussian mixture models, EM algorithm

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

Repeat

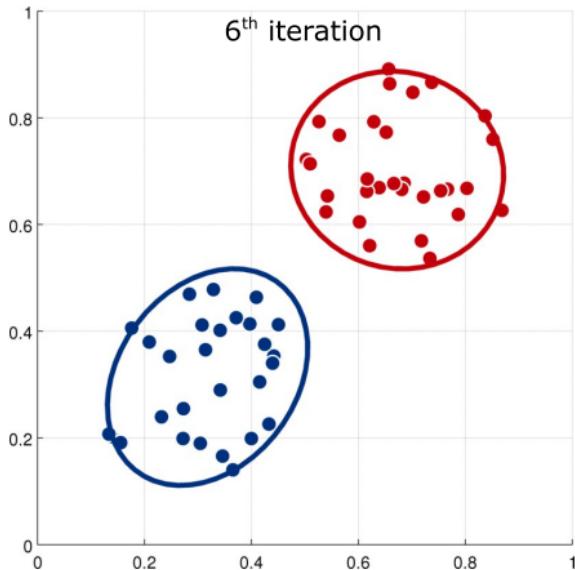
- **Expectation**

- For each object, calculate the probability of belonging to each distribution

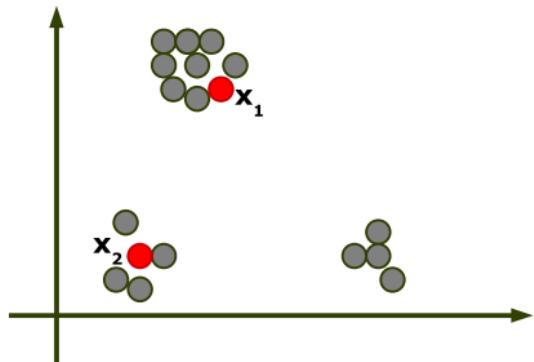
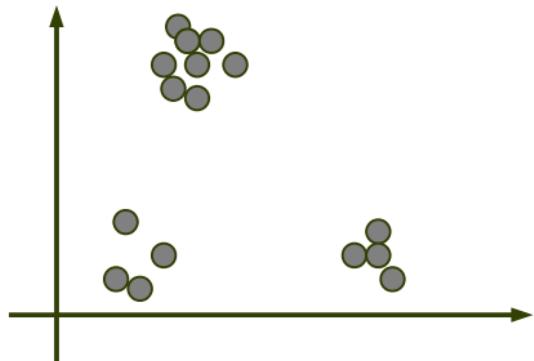
- **Maximization**

- For each probability distribution, estimate parameters by maximum likelihood

Until the parameters do not change



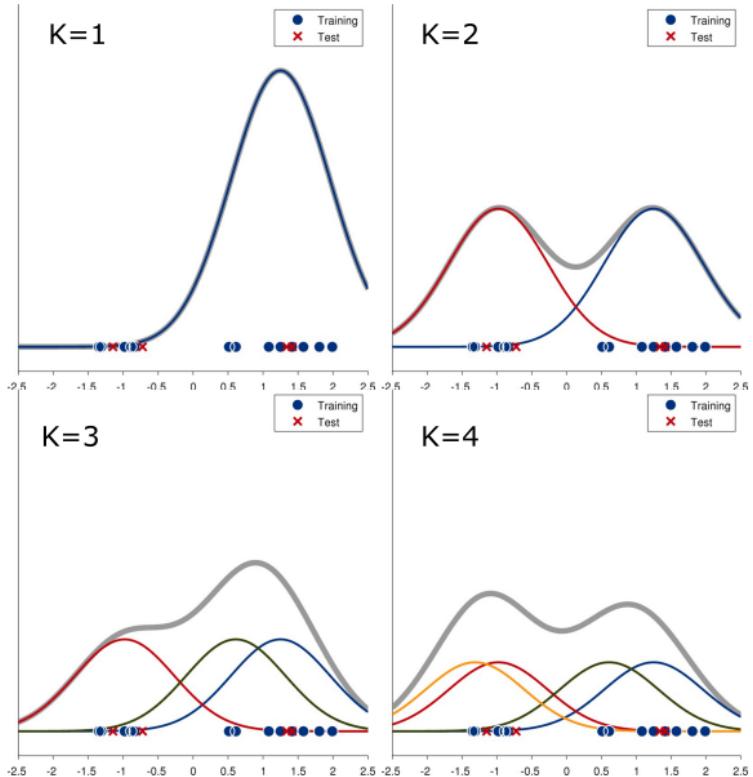
- Consider the data to the right with 16 observations.
 - What would ideally happen if we used a GMM with $K=16$ clusters to model the data?
- Imagine we have two **test observations** denoted \mathbf{x}_1 and \mathbf{x}_2 (red points) that are not used for training.
 - What happens to $p(\mathbf{x}_1)$ and $p(\mathbf{x}_2)$ if we use $K=3$ and $K=16$ clusters?



EM Initial solution

Mixture models

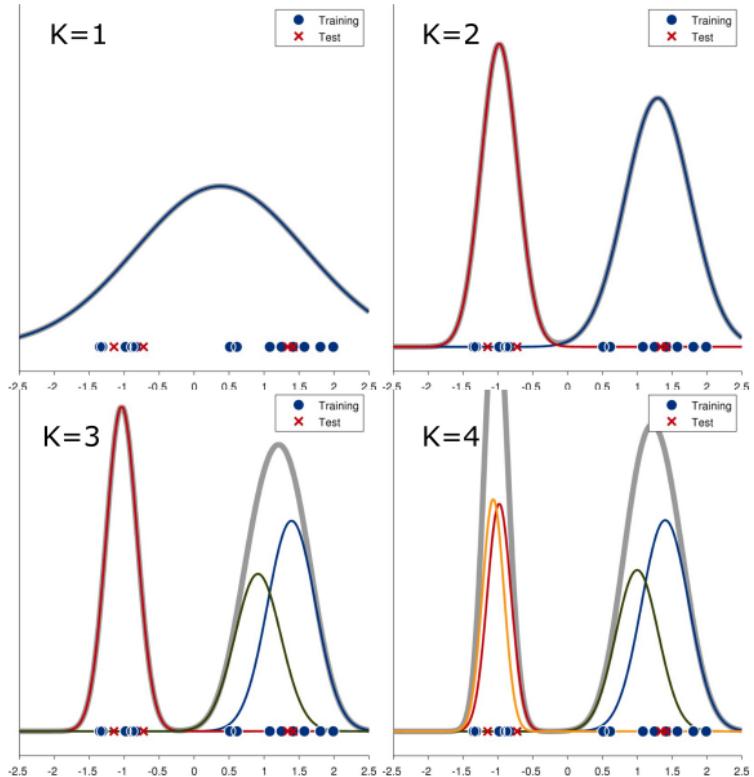
- Selecting complexity using crossvalidation



EM 1st iteration

Mixture models

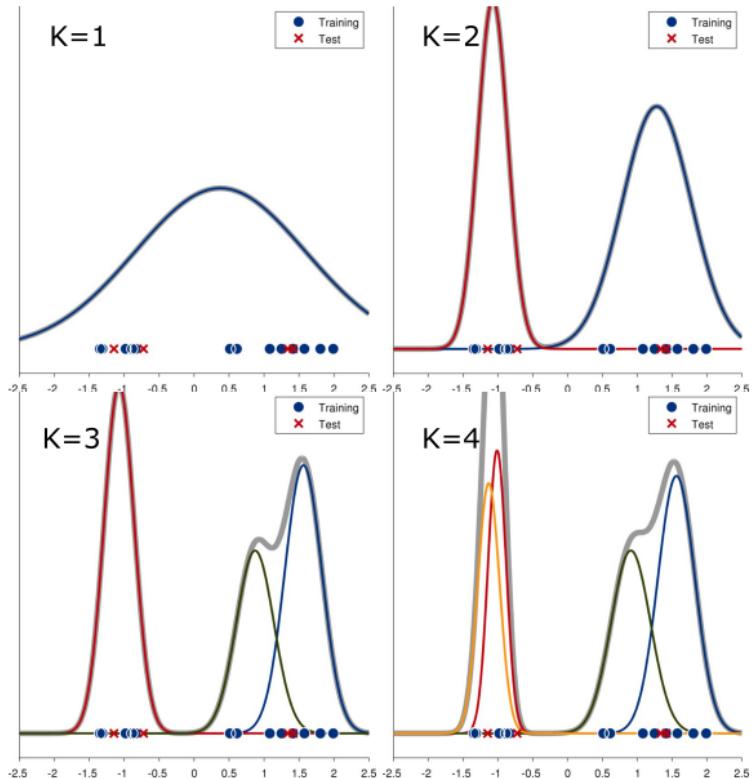
- Selecting complexity using crossvalidation



EM 2nd iteration

Mixture models

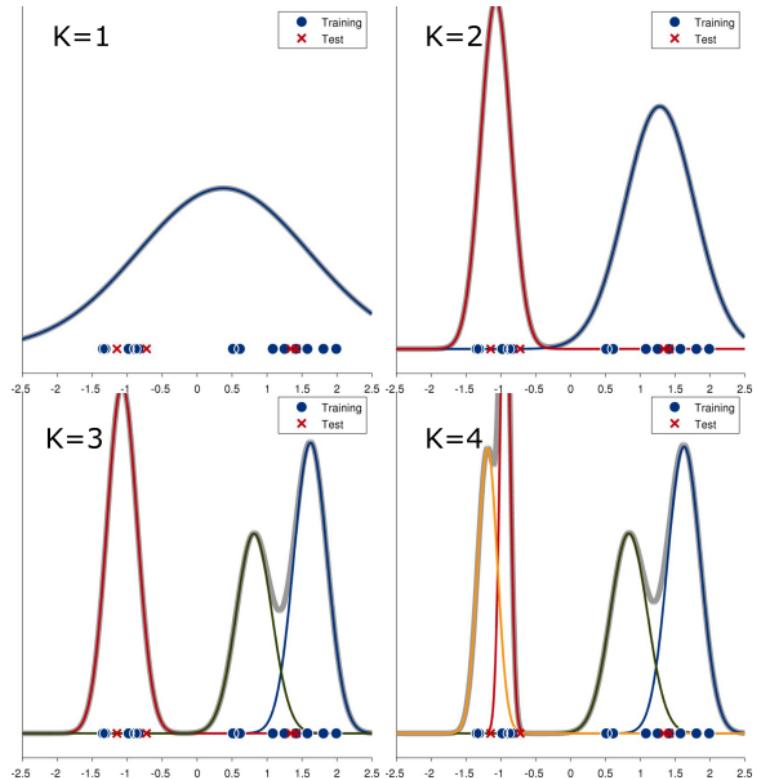
- Selecting complexity using crossvalidation



EM 3rd iteration

Mixture models

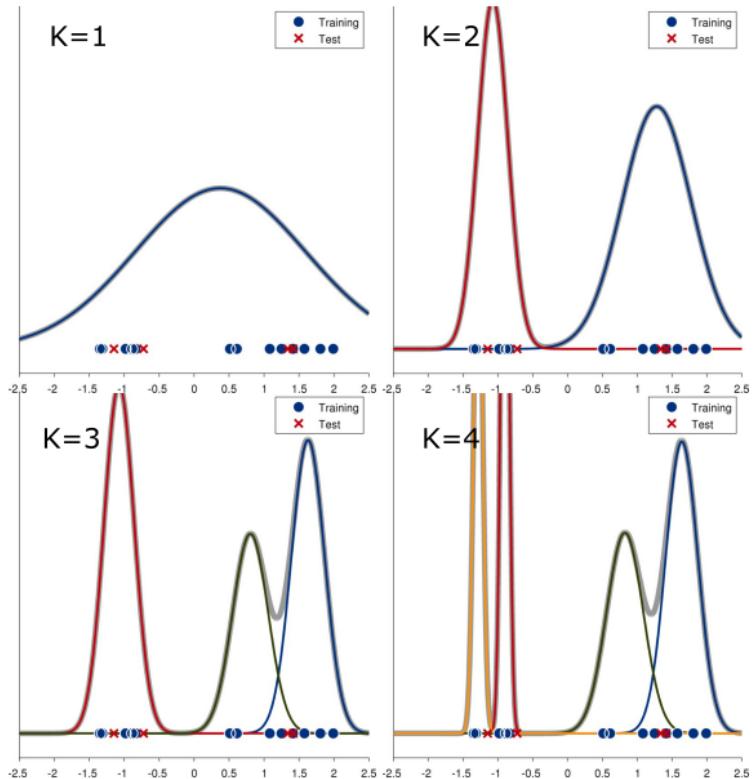
- Selecting complexity using crossvalidation



EM 4th iteration

Mixture models

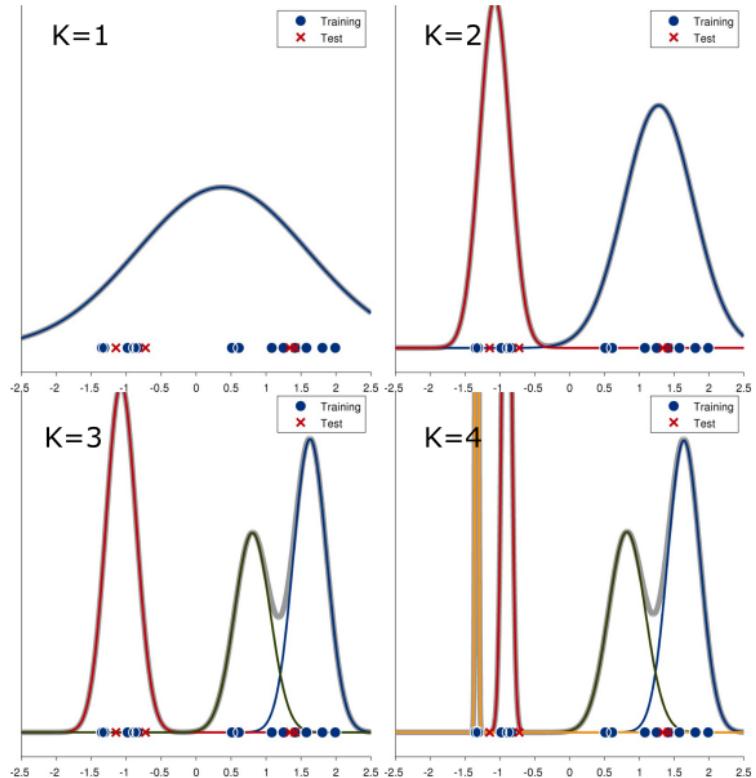
- Selecting complexity using crossvalidation



EM 5th iteration

Mixture models

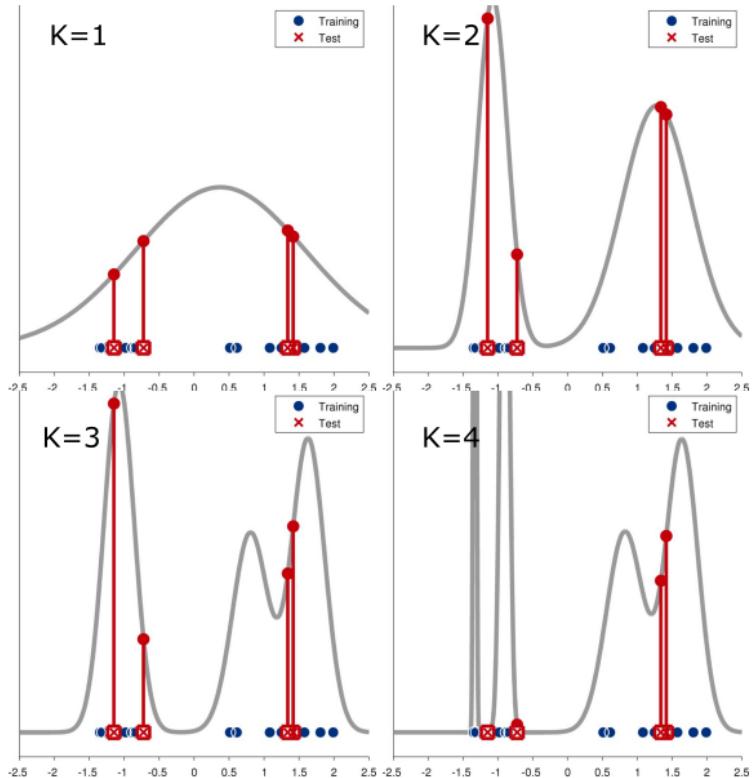
- Selecting complexity using crossvalidation



Test data evaluation

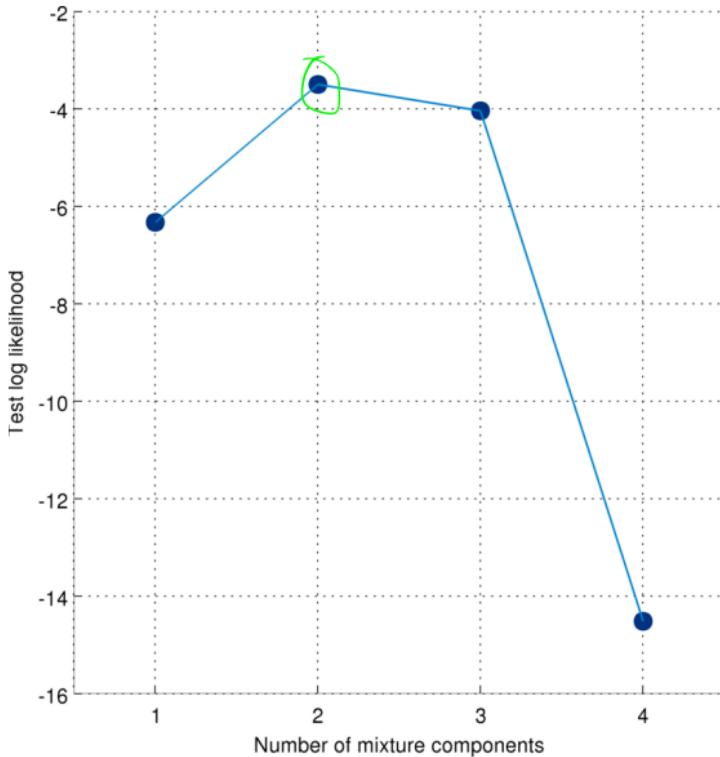
Mixture models

- Selecting complexity using crossvalidation



Mixture models

- Selecting complexity using crossvalidation



K-means versus GMM

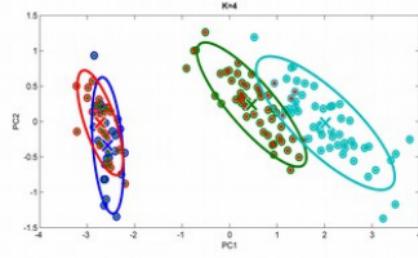
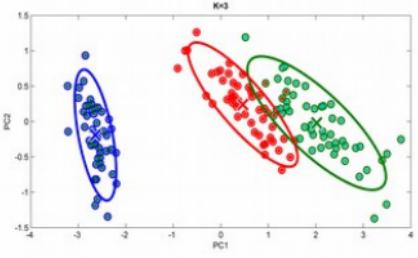
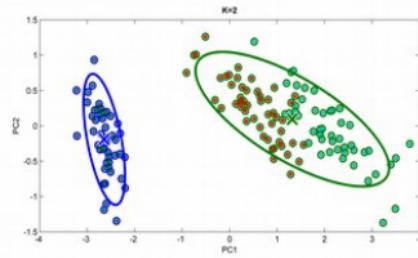
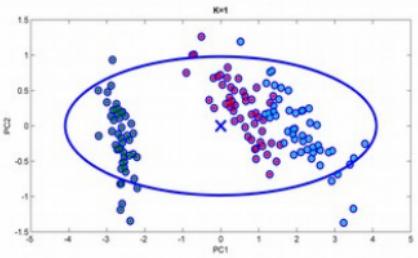
K-means

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

Gaussian mixture model (GMM)

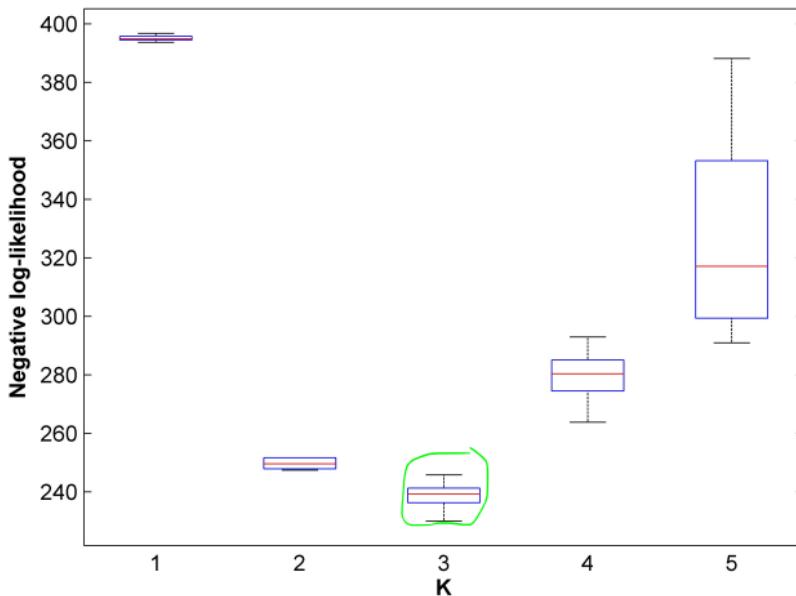
- No guarantee of optimal solution (even more local minima issues due to the additional model parameters)
- Models shape of cluster as ellipsoid
- Models the size of clusters
- Possible to estimate the number of components by cross-validation

GMM on Iris data using 1,2,3 and 4 components



Recap of GMM on Iris data

GMM 10 fold cross-validation on Iris data repeated five times where the five runs are plotted using box-plots.



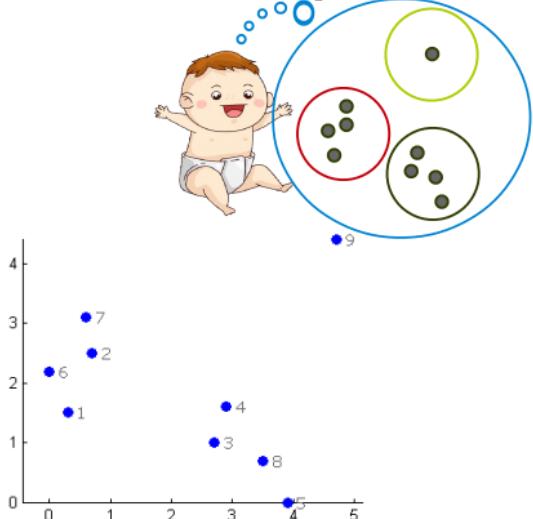
Anomaly detection: Definition

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

Anomaly detection: Example

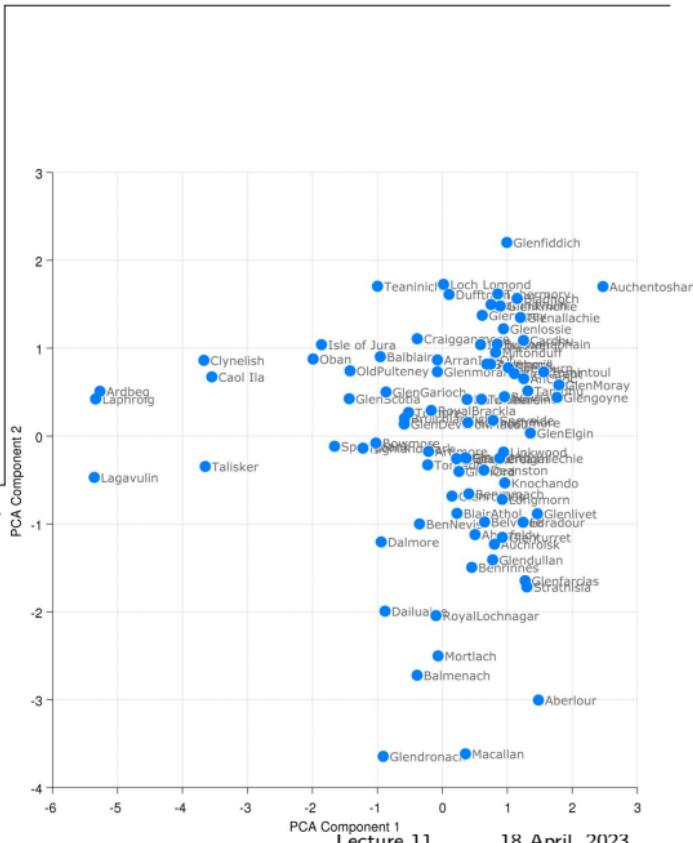
- Credit card **fraud detection**
 - Recognize dubious credit card transactions based on the transaction history of the card holder
- Network **intrusion detection**
 - Detect hacker attacks, web crawlers etc.
- **Ecosystem disturbances**
 - Detect hurricanes, floods droughts, heat waves and fires
- **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

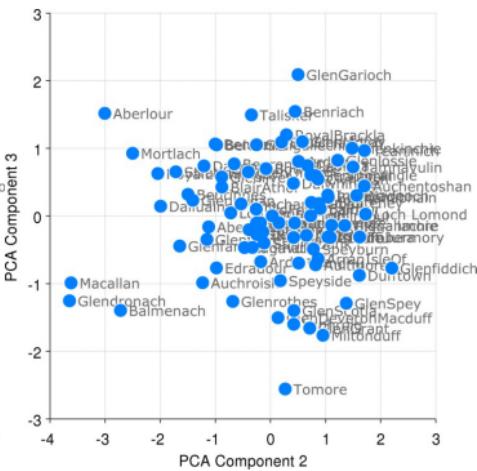
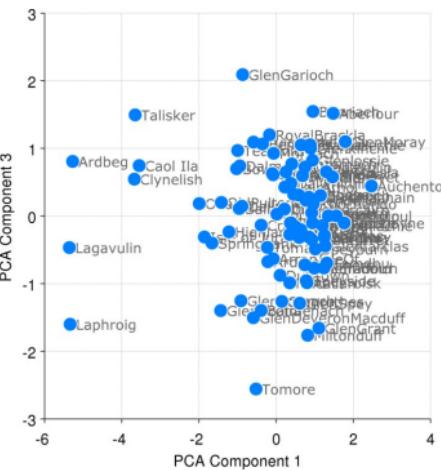
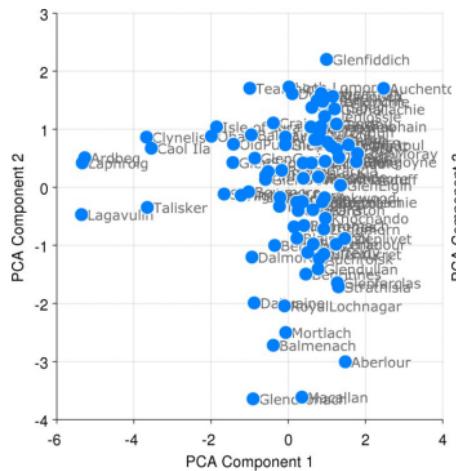


Data example II: Whisky

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



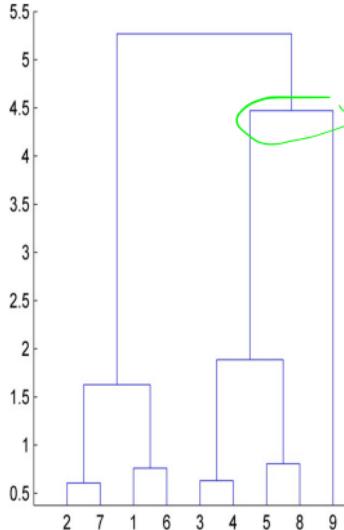
PCA plot



Dendrogram

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

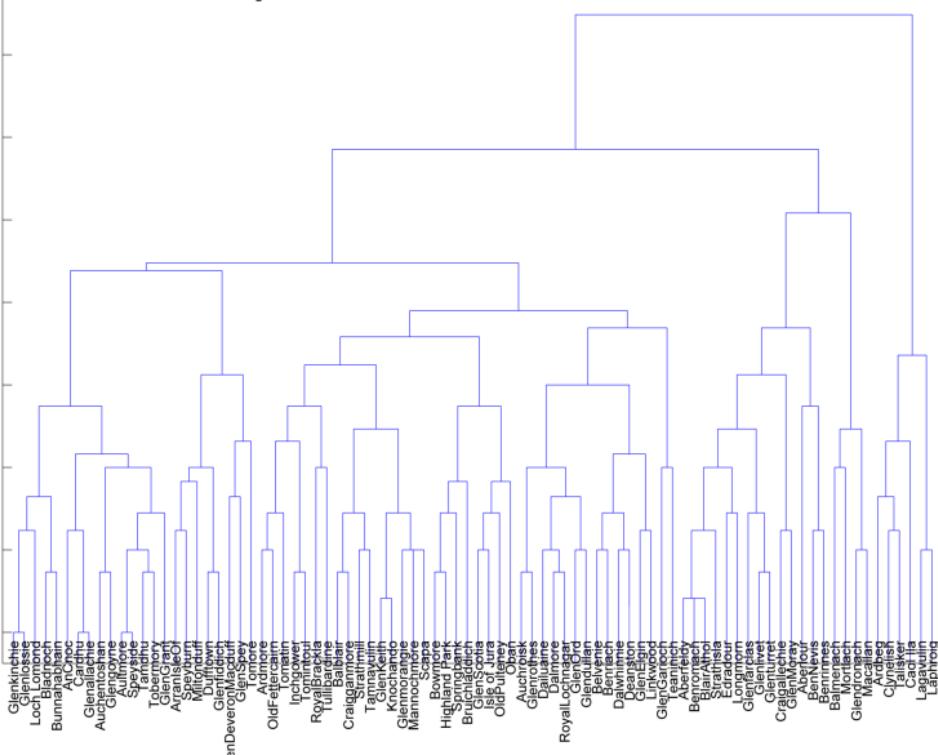
Data I: Cats, Dogs and Dinosaurs



Nr. 9 linker med høj værdi.

Dvs. den er 4.5 væk fra den nærmeste.
Den er lang fra de andre punkter.

Data II: Whisky



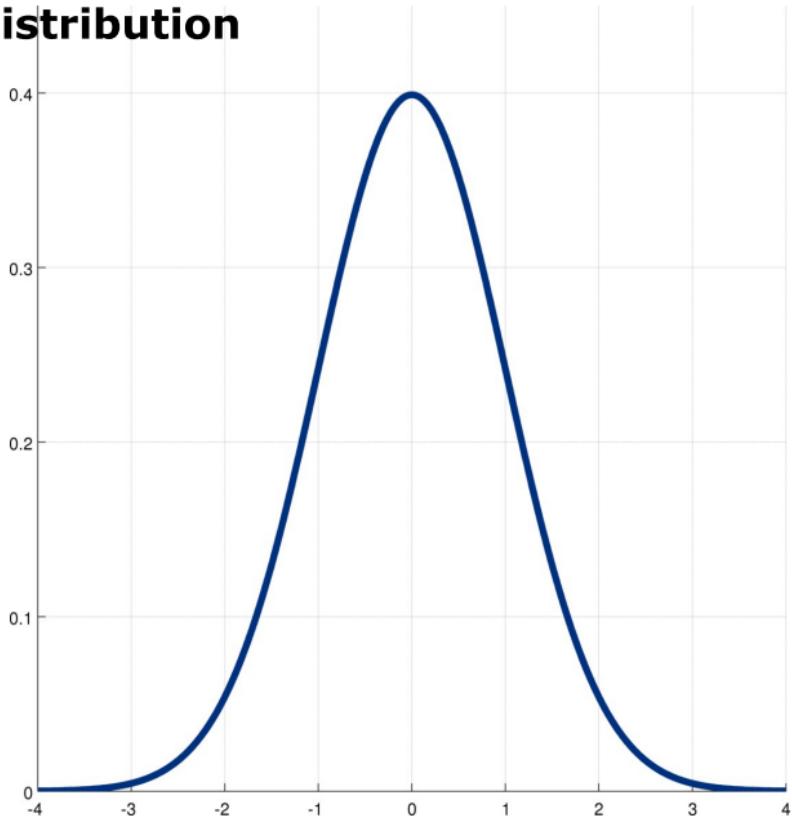
Density based techniques: Univariate normal distribution

- 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



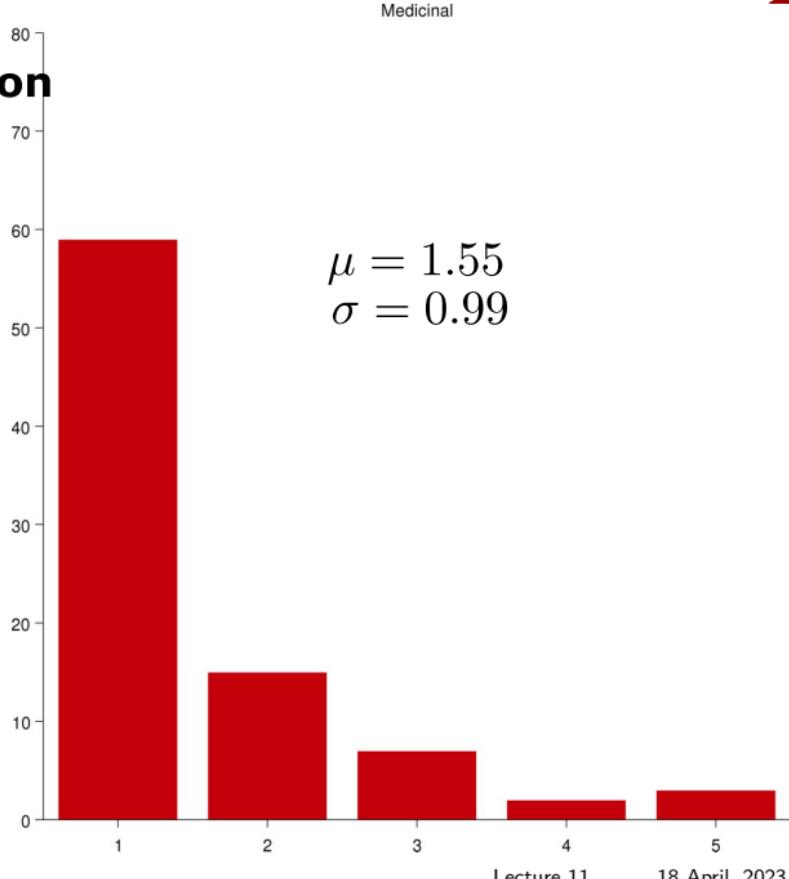
Normal distribution

- 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



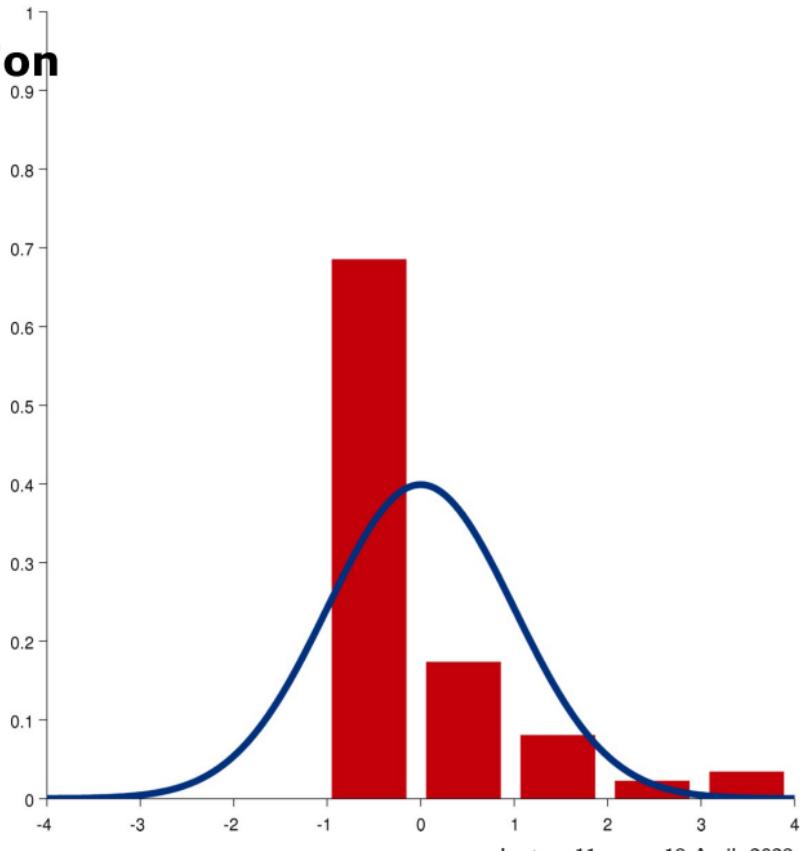
Normal distribution

- 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



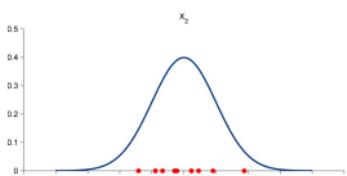
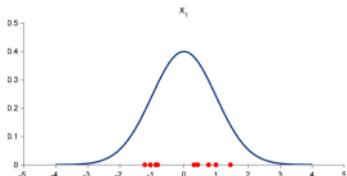
Normal distribution

- Map attribute to standard Normal variable
- Choose a threshold

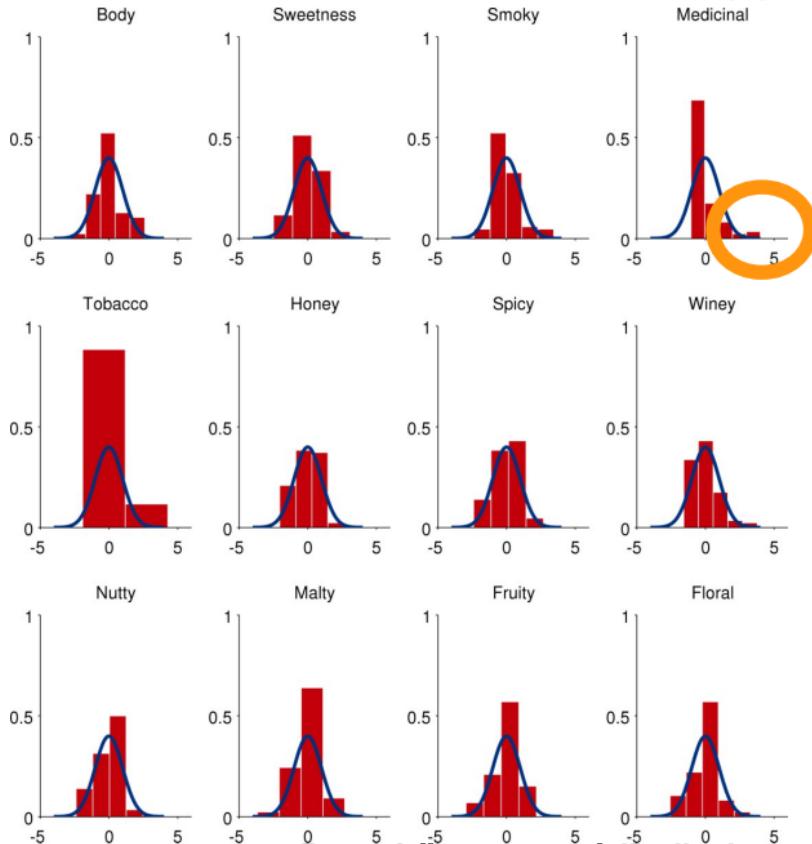
$$z = \frac{x - \mu}{\sigma}$$

$$p(|z| > c) = 0.001$$
$$c = 3.2905$$

Data I: Cats, Dogs and Dinosaurs



Data II: Whisky



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

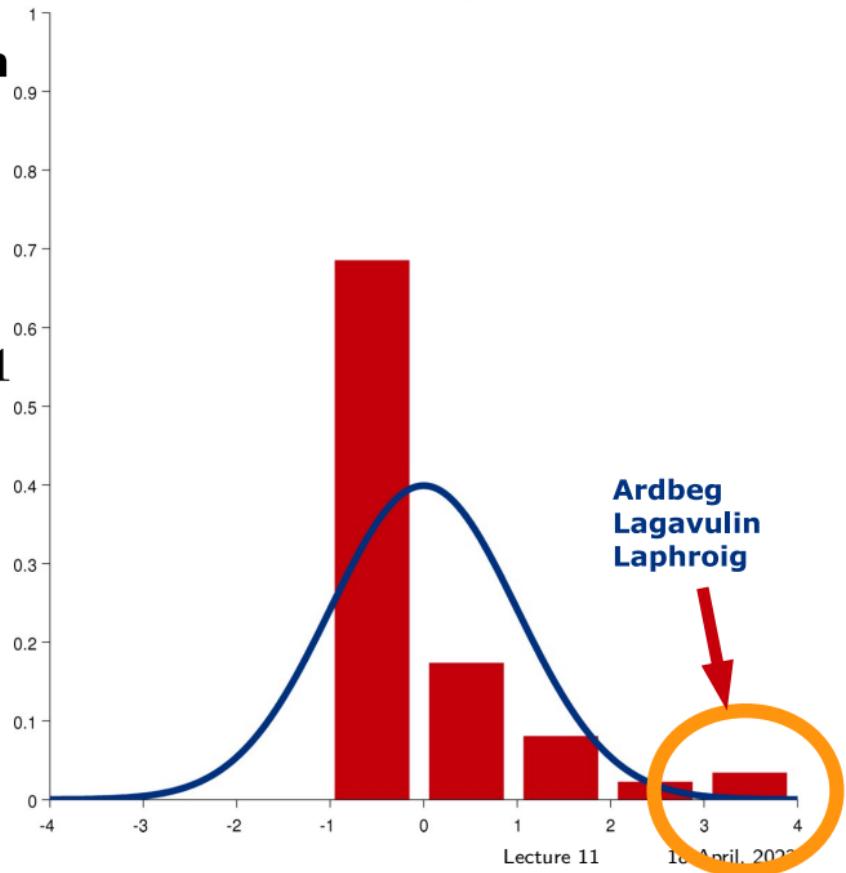
Normal distribution

- 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**
 - Estimate the density of data objects
 - Outliers are:
 - Data objects in low density area
- **We can of course use the GMM to evaluate the density of test data.**
 - why not on the training data?**
- **Approaches we will presently also consider:**
 - Kernel density estimation
 - Inverse average distance to K nearest neighbours (KNN density)
 - Average relative KNN density

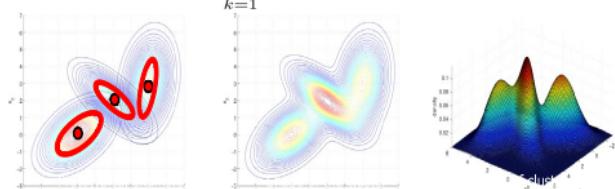
Density based techniques: Kernel Density Estimator

Recall the Gaussian Mixture Model (GMM)

Data density Sum of cluster specific densities assumed normal distributed

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

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

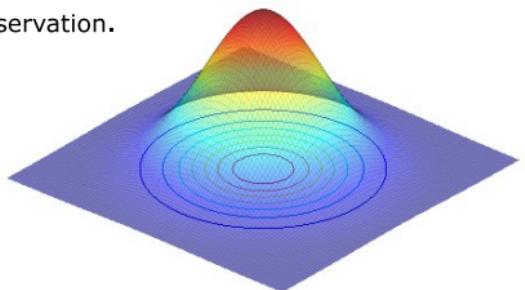


$\boldsymbol{\mu}_{(k)}$: Cluster center (prototypical example in cluster)

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

w_k : Relative density of the cluster

Kernel Density estimation based on Gaussian Kernel:
Consider the GMM and define a Gaussian with mean \mathbf{x}_n and co-variance $\sigma^2 \mathbf{I}$ around each Observation.



Let all observation weight the same, i.e. $w_n = 1/N$

$$p(\mathbf{x}) = \sum_{n=1}^N \frac{1}{N} \mathcal{N}(\mathbf{x} | \mathbf{x}_n, \sigma^2 \mathbf{I})$$

Only free parameter σ^2 !

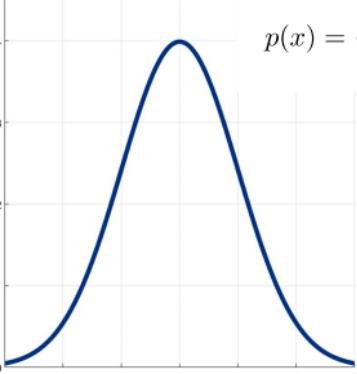
There is nothing special about the normal distribution. For a general mixture distribution p the general form of kernel density estimator is:

$$p(\mathbf{x}) = \sum_{n=1}^N \frac{1}{N} p(\mathbf{x} | \mathbf{x}_n, \theta)$$

This may be useful if \mathbf{x} is discrete or non-negative.

Quiz 03: Kernel density (Spring 2013)

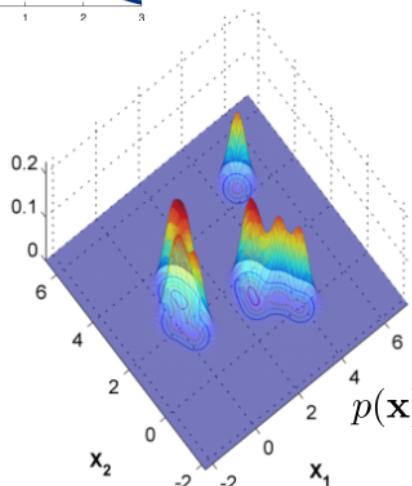
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$



Consider five observations of an attribute x given by

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

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

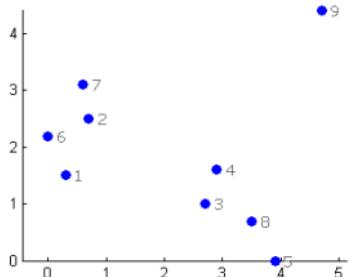


- A. $\frac{1}{\sqrt{8\pi}} \exp\left(-\frac{53}{4}\right)$
- B. $\frac{1}{5\sqrt{8\pi}} \exp\left(-\frac{53}{4}\right)$
- C. $\frac{1}{5\sqrt{8\pi}} \left(\exp\left(-\frac{1}{2}\right) + 2 \cdot \exp\left(-\frac{1}{8}\right) + \exp\left(-\frac{9}{2}\right) + \exp(-8)\right)$
- D. $\frac{1}{5\sqrt{8\pi}} \left(\exp(-1) + 2 \cdot \exp\left(-\frac{1}{4}\right) + \exp(-9) + \exp(-16)\right)$
- E. Don't know.

$$p(\mathbf{x}) = \sum_{n=1}^N \frac{1}{N} \mathcal{N}(\mathbf{x}|\mathbf{x}_n, \sigma^2 \mathbf{I})$$

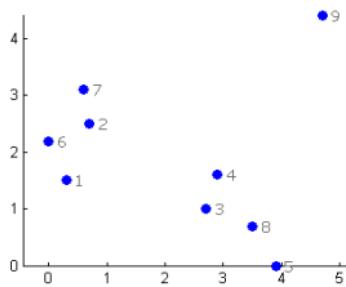
How do we determine σ^2 ?

Data I: Cats, Dogs and Dinosaurs

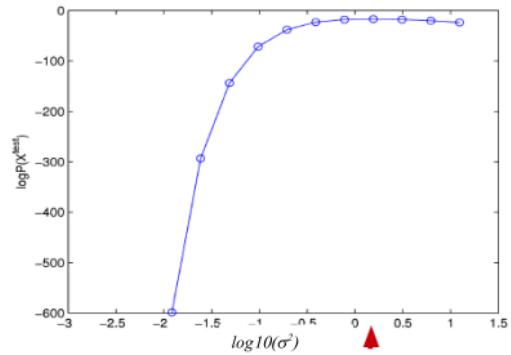


How do we determine σ^2 ? Crossvalidation!

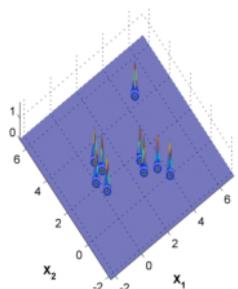
Data I: Cats, Dogs and Dinosaurs



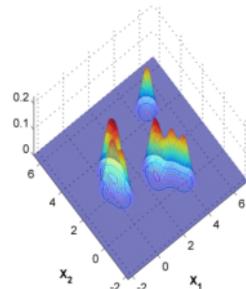
Density of test set based on leave-one-out cross validation



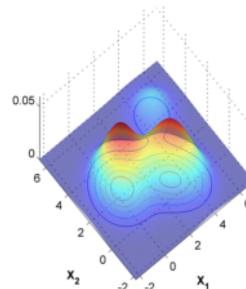
$$\sigma^2 = 0.01$$



$$\sigma^2 = 0.1$$

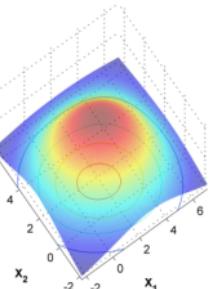


$$\sigma^2 = 1$$

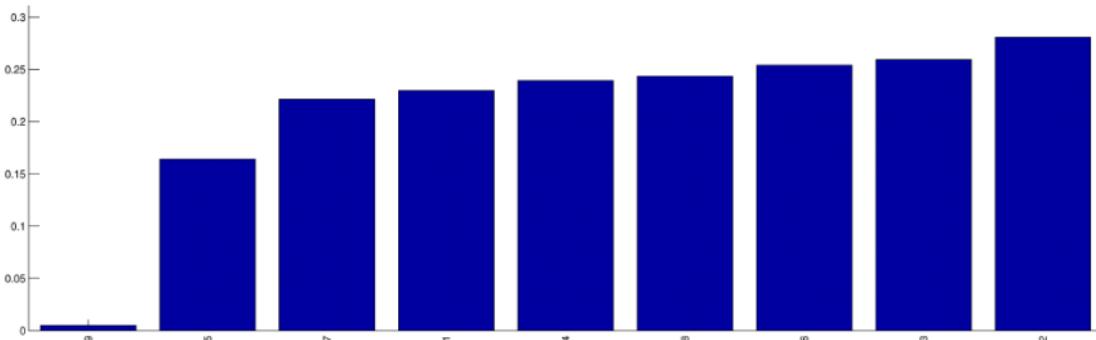
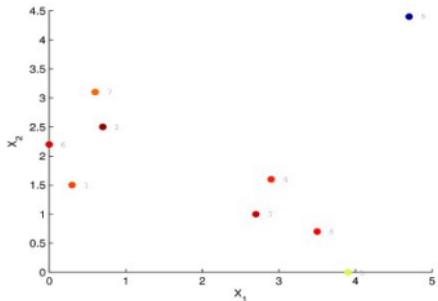


$$\sigma^2 = 1.55$$

$$\sigma^2 = 5$$

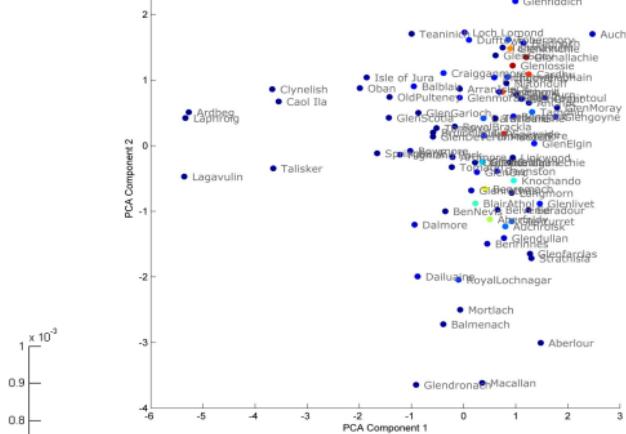


Estimated leave-one-out density evaluated at each observation

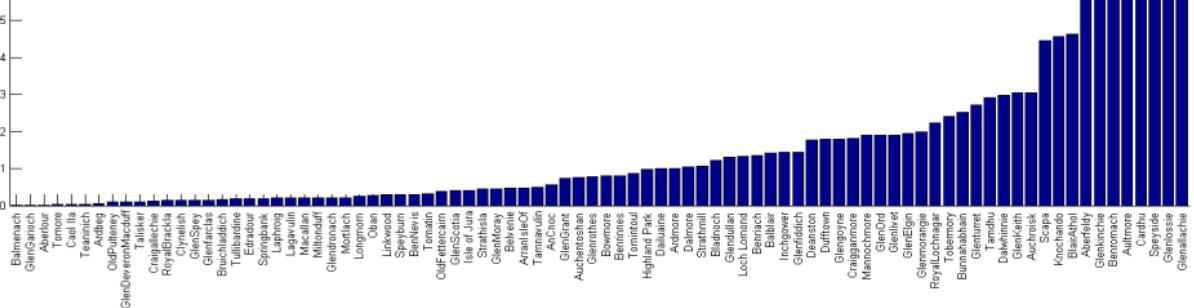


Estimated density evaluated at each observation

Data II: Whisky

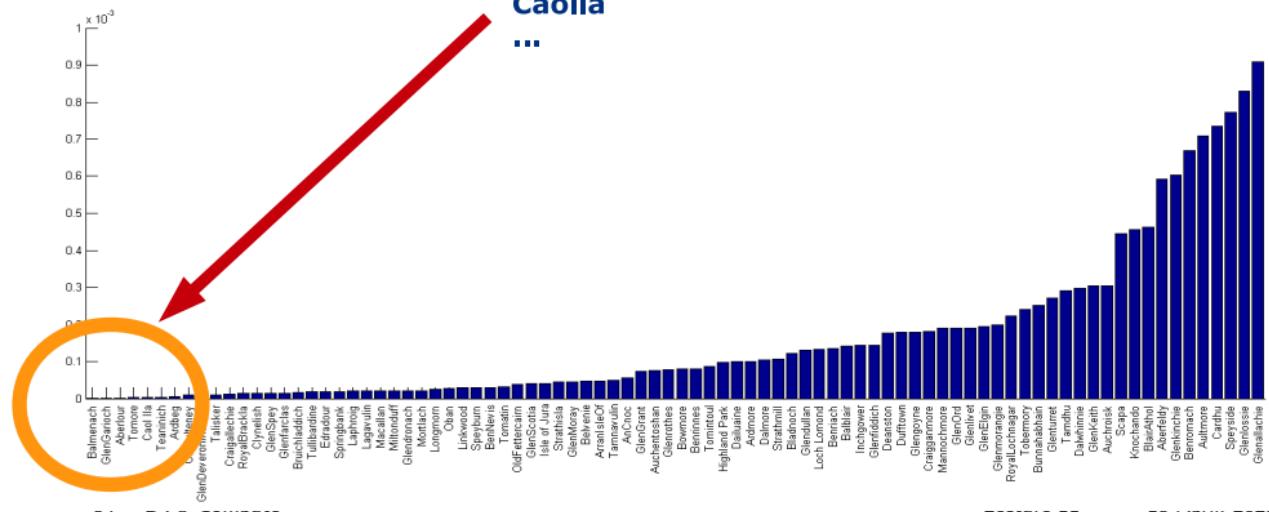


leave-one-out densities



Data II: Whisky

Balmenach
Glen Garioch
Aberlour
Tomore
Caolla
...



Inverse distance density estimation

- **Distance based measure of density**

- Density is inverse proportional to average distance to k nearest neighbors
- Density is low if nearest neighbors are far away

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

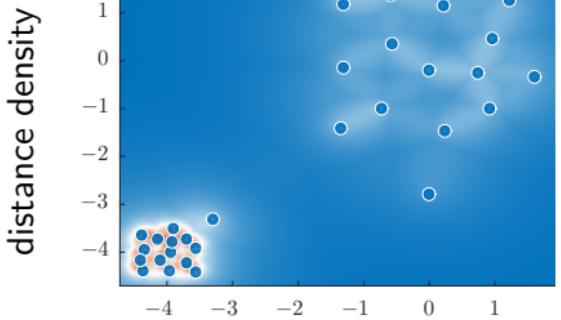
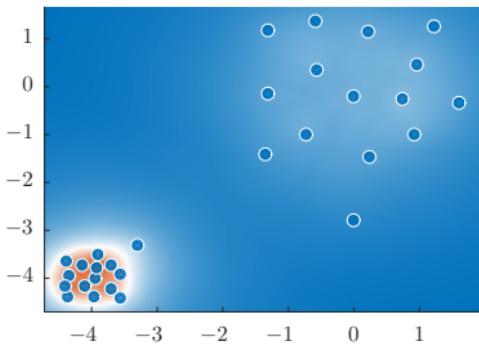
- **Relative density**

- Density compared to density at nearest neighbors

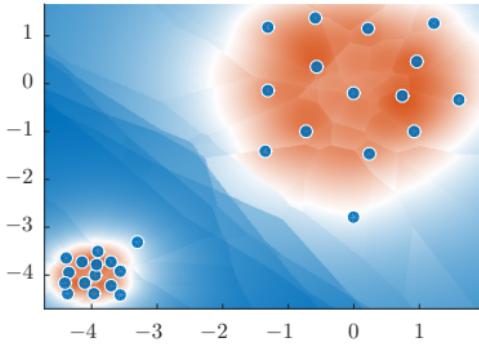
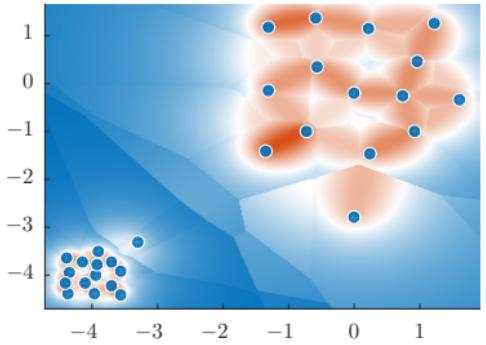
$$\text{ard}_{\mathbf{X}}(\mathbf{x}_i, K) = \frac{\text{density}_{\mathbf{X}_{\setminus i}}(\mathbf{x}_i, K)}{\frac{1}{K} \sum_{\mathbf{x}_j \in N_{\mathbf{X}_{\setminus i}}(\mathbf{x}_i, K)} \text{density}_{\mathbf{X}_{\setminus j}}(\mathbf{x}_j, K)}$$

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

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

$k = 2$  $k = 6$ 

ARD



Quiz 4: ARD (Spring 2015)

	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 pollution) 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:

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

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

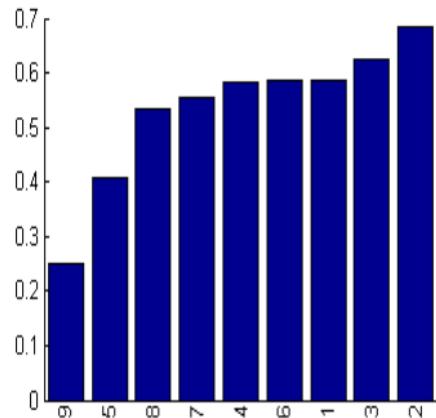
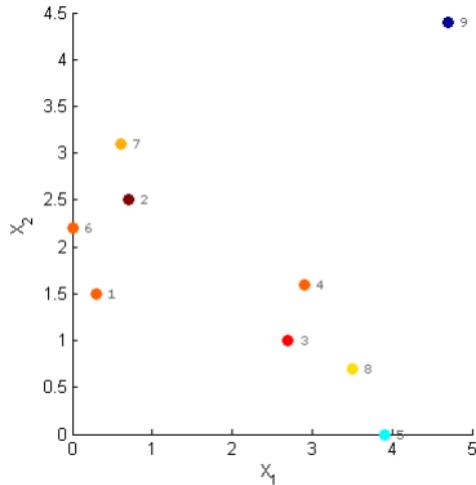
where $N(\mathbf{x}, K)$ is the set of K nearest neighbors of observation \mathbf{x} and $\text{a.r.d.}(\mathbf{x}, K)$ is the average relative density of \mathbf{x} using K nearest neighbors. What is ARD for observation O1 for $K = 2$ nearest neighbors?

- 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)

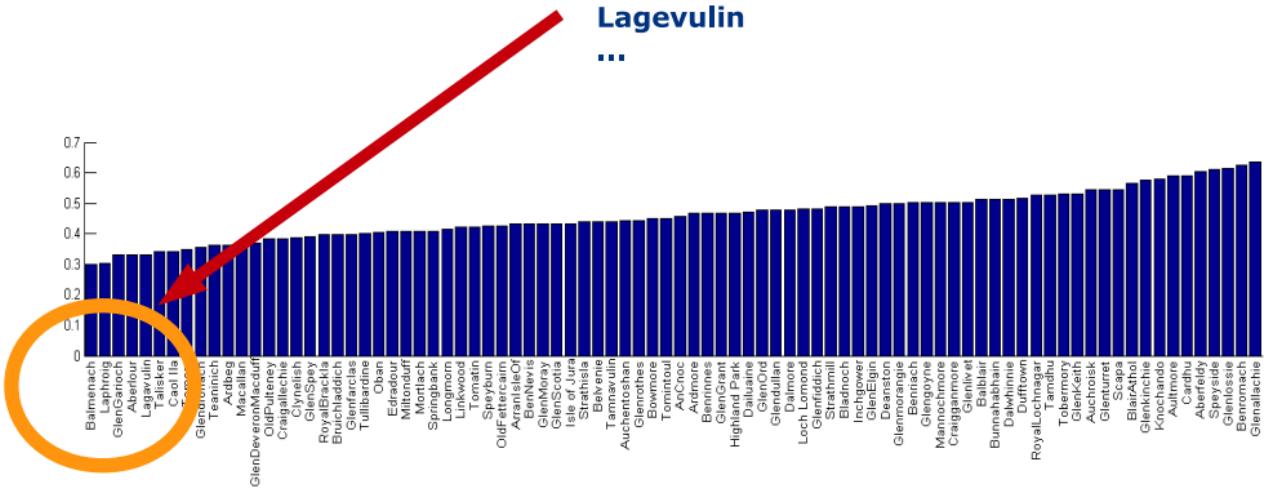
Data I: Cats , dogs and dinosaurs



Inverse distance density estimation

- KNN density (5 nearest neighbors)

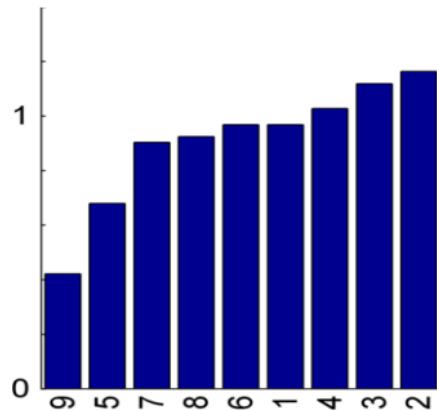
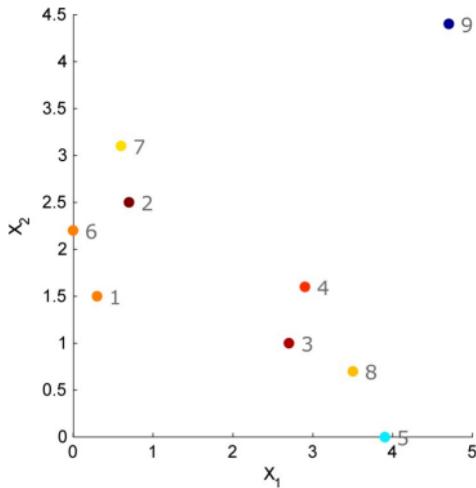
Balmenach
Laphroig
GlenGarioch
Aberlour
Lagevulin
...



Average Relative density

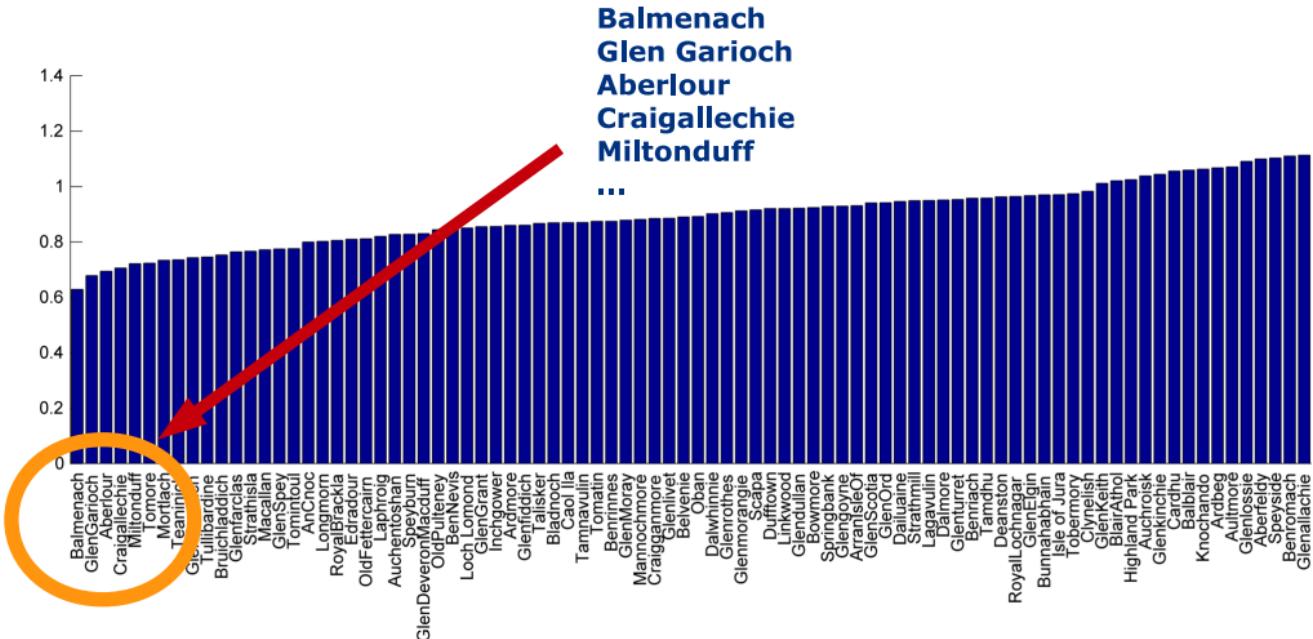
- Average Relative KNN density (5 nearest neighbors)

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

- **KNN density**

- Balmenach
- Laphroig
- Glen Garioch
- Aberlour
- Lagavulin

- **KNN average relative density**

- Balmenach
- Glen Garioch
- Aberlour
- Craigallechie
- Miltonduff

Common: Balmenach, Glen Garioch,
Aberlour

Resources

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