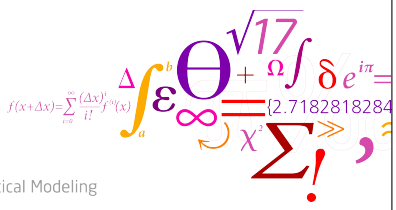
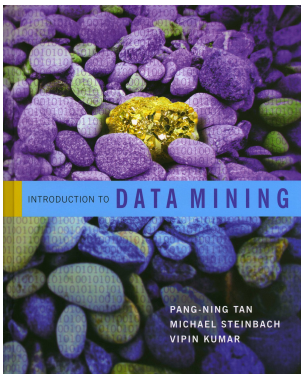
02450 Introduction to machine learning and data mining



DTU Informatics
Department of Informatics and Mathematical Modeling

Reading material

Tan, Steinbach and Kumar
"Introduction to Data Mining"
Section 10.1-10.4



Feedback Groups of the day:

Wen Hsin Li, Robert Lyck, Viivi Halla-aho

Oliver Naaby, Rasmus Olsen

Simon Hemmingsen, Thomas Hammerbak

Henrik Thirstrup, Mette Marie Nielsen Thomas Kristensen

Nicolai Pedersen, Jeppe Thagaard, Mathias Brade, Christian Kragh

Oscar Petersen, Søren Vørnle Nielsen

Andreas Jacobsen, Asger Anker Sørensen, Phong Le Trung

If possible, please (1) stay to give me feedback after the second lecture today (ca. 15:00) and (2) send an email or contact me at the exercises next week with feedback/suggestions on the exercises for today.

Lecture schedule

1. Introduction
   (Tan 1.1-1.4)

Data: Feature extraction and visualization

2. Data and feature extraction
   (Tan 2.1-2.3 + B1 (+ A)
3. Measures of similarity and summary statistics
   (Tan 2.4 + 3.1-3.2 + C1-C2)
4. Data visualization
   (Tan 3.3)

Supervised learning: Classification and regression

5. Decision trees and linear regression
   (Tan 4.1-4.3 + D)
6. Overfitting and performance evaluation
   (Tan 4.4-4.6)
7. Nearest neighbor, naive Bayes, and artificial neural networks
   (Tan 5.2-5.4)
8. Ensemble methods and multi class classifiers
   (Tan 5.6-5.8)

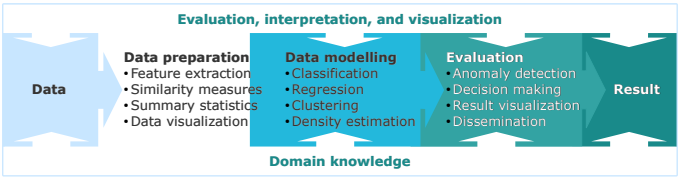
Unsupervised learning: Clustering and density est.

9. K-means and hierarchical clustering
   (Tan 8.1-8.3+8.5.7)
10. Mixture models and association mining
    (Tan 9.2.2 + 6.1-6.3)
11. **Density estimation and anomaly detection**
    (Tan 10.1-10.4)

Machine learning and data modelling in practice

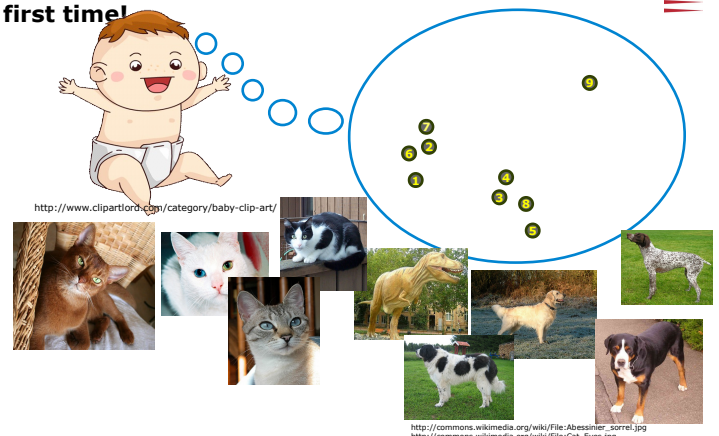
12. Putting it all together: Summary and overview
13. Mini project

Data modeling framework



After today you should be able to:
Understand and apply a variety of outlier/anomaly detection approaches, including density estimation approaches and proximity-based techniques.

Imagine (yet again) you observe the world for the first time!



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

How do we detect anomalous objects (i.e., the dinosaur in the world of cats and dogs)

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:Cat002.jpg
http://commons.wikimedia.org/wiki/File:Saurier2.jpg
http://commons.wikimedia.org/wiki/File:Golden_Retriever_Dukedestiny01.jpg
http://commons.wikimedia.org/wiki/File:MasPiri-Astro-SVE.jpg
http://commons.wikimedia.org/wiki/File:GroC3%9Fer_Schweizer_Sennenhund.jpg
http://commons.wikimedia.org/wiki/File:GermanShorthPtr_wb.jpg

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
- **Healt 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

Group exercise

- Come up with **your own definition** of an outlier / anomaly
- How can we detect outliers using some of the methods you have already learned in the course?

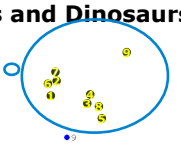
Group exercise

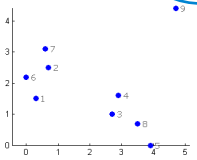
- Come up with **your own definition** of an outlier / anomaly
- How can we detect outliers using some of the methods you have already learned in the course?

Hawkins' definition of an outlier: *An outlier is an observation that differs so much from other observations as to arouse suspicion that it was generated by a different mechanism*

Probabilistic definition of an outlier: *An outlier is an object that has a low probability with respect to a probability distribution model of the data.*

Data example I: Cats, Dogs and Dinosaurs

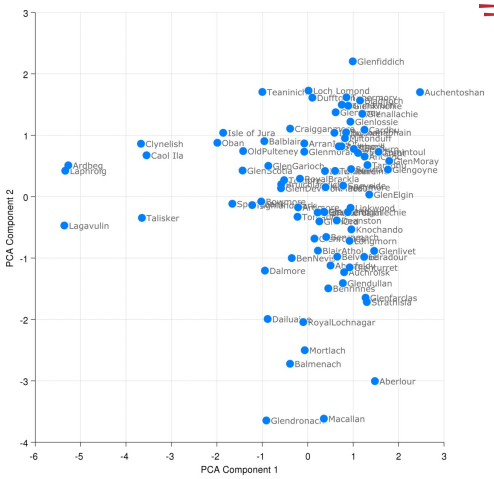




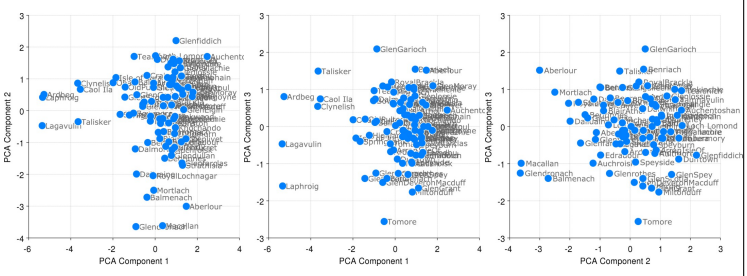
Data example II: Whisky

- 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



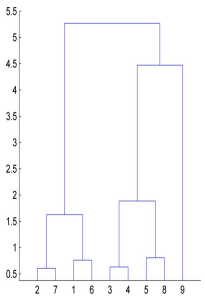
PCA plot



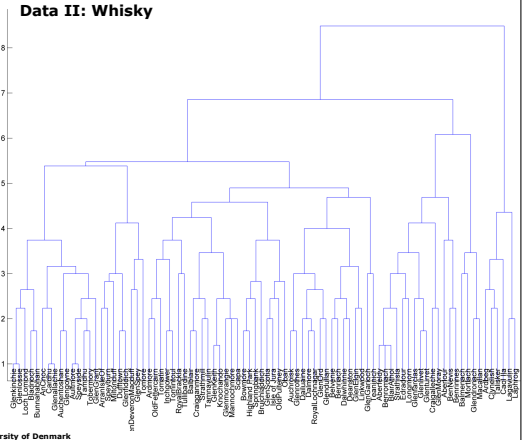
Dendrogram

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

Data I: Cats, Dogs and Dinosaurs



Data II: Whisky



Approaches to anomaly detection

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

  Approaches we will consider:
  - Univariate normal distribution
  - Kernel density estimation

- **Proximity-based techniques**
  - Measure the distance between data objects
  - Outliers are
    - Data objects far from the other data objects

  Approaches we will consider:
  - Mahanalobis distance to center of data
  - Distance to K^{th} nearest neigbour
  - Inverse average distance to K nearest neighbours (KNN density)
  - Average relative KNN density

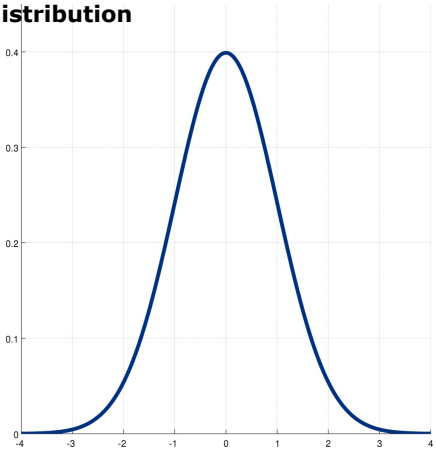
Density based techniques: Univariate normal distribution

- 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

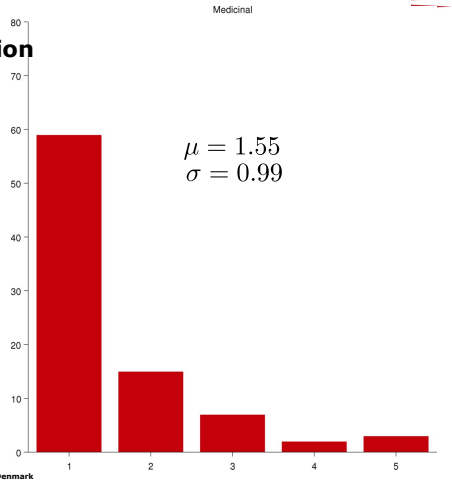
$$z = \frac{x - \mu}{\sigma}$$

- 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

Medicinal

$$\mu = 1.55$$
$$\sigma = 0.99$$



Normal distribution

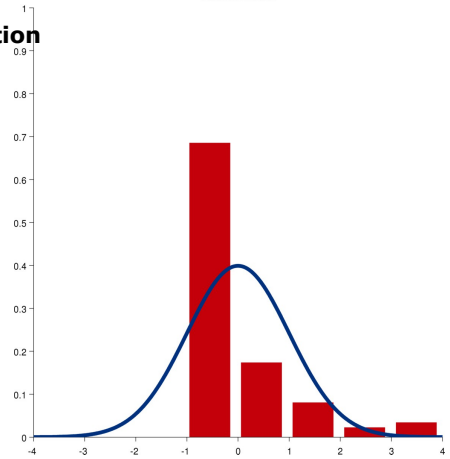
- Map attribute to standard Normal variable

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

- 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

Medicinal: z-score



Normal distribution

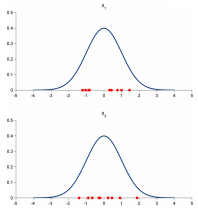
- Map attribute to standard Normal variable

- Choose a threshold

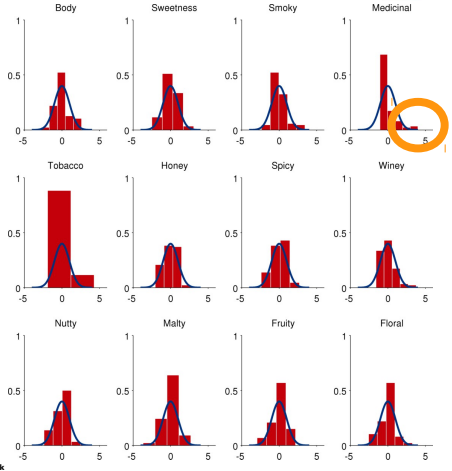
$$z = \frac{x - \mu}{\sigma}$$

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

Data I: Cats, Dogs and Dinosaurs



Data II: Whisky



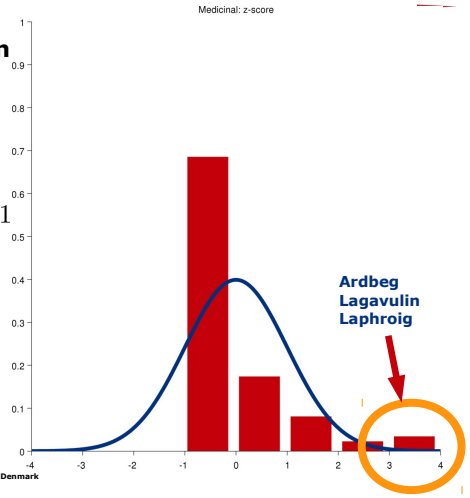
Normal distribution

- Map attribute to standard Normal variable

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

- Choose a threshold

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

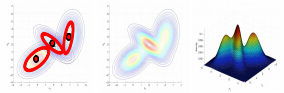


Density based techniques: Kernel Density Estimator

Rememer from Last week:
Gaussian Mixture Model (GMM)

Data density — Sum of cluster specific densities assumed normal distributed

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

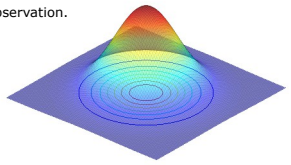


$\boldsymbol{\mu}_{(k)}$: Cluster center (prototypical example in cluster)

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

w_k : Relative density of the cluster

Kernel Density estimation based on Gaussian Kernel: Consider the GMM and define a Gaussian with mean x_n and co-variance $\sigma^2 \boldsymbol{I}$ 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!

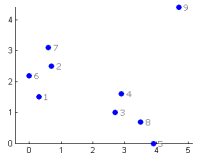
There is nothing special about the normal distribution. For a general mixture distribution p the general form of kernel density estimator is:

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

This may be useful if x is discrete or non-negative.

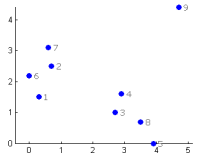
How do we determine σ?

Data I: Cats, Dogs and Dinosaurs

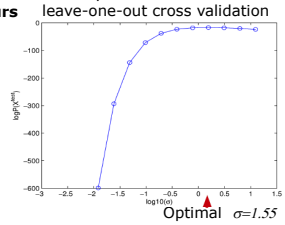


How do we determine σ?

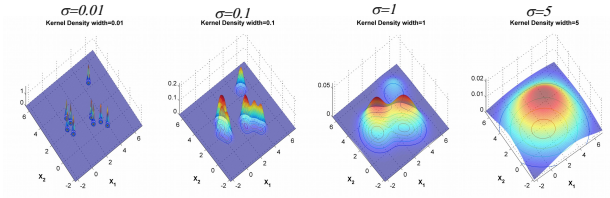
Data I: Cats, Dogs and Dinosaurs



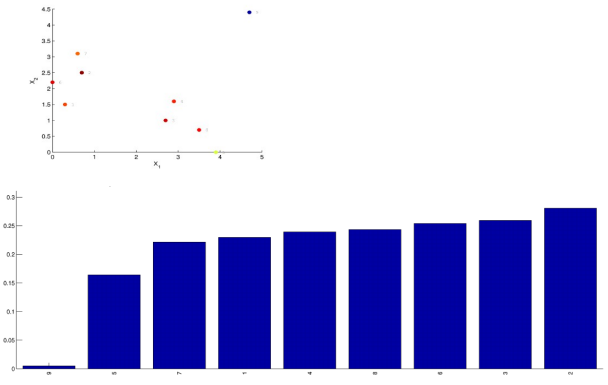
Density of test set based on leave-one-out cross validation



Optimal σ=1.55

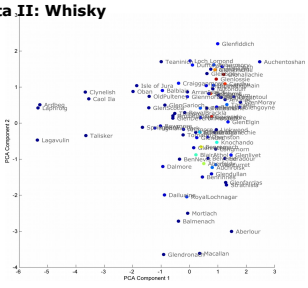


Etimated leave-one-out density evaluated at each observation

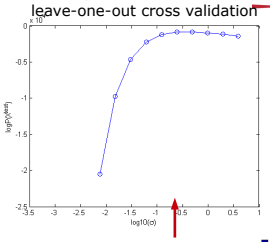


Etimated density evaluated at each observation

Data II: Whisky

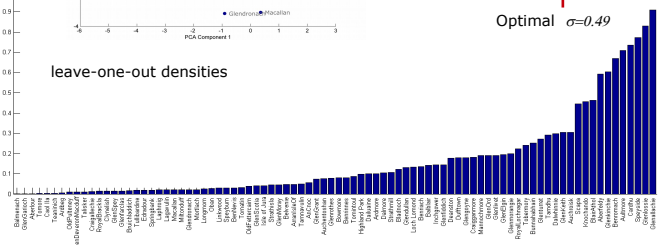


Density of test set based on leave-one-out cross validation

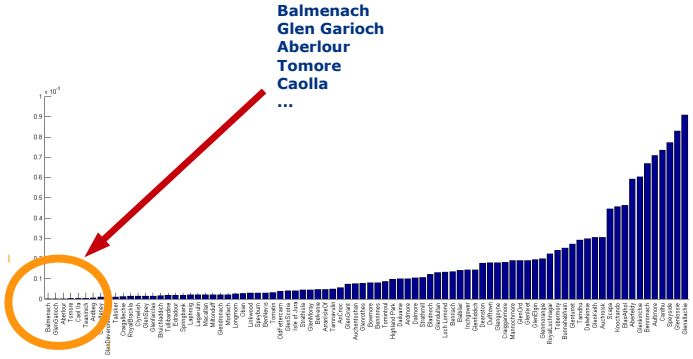


Optimal σ=0.49

leave-one-out densities



Data II: Whisky



Proximity-based techniques

- Mahanalobis distance to center of data
- Distance to Kth nearest neigbour
- Inverse average distance to KNN

Mahalanobis distance

- Distance to the mean of data, taking covariance into account

$$d_{\text{mahalanobis}}(\boldsymbol{x}, \bar{\boldsymbol{x}}) = (\boldsymbol{x} - \bar{\boldsymbol{x}})^\top \boldsymbol{S}^{-1} (\boldsymbol{x} - \bar{\boldsymbol{x}})$$

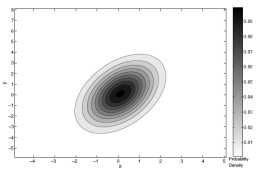


Figure 10.2. Probability density of points for the Gaussian distribution used to generate the points of Figure 10.3.

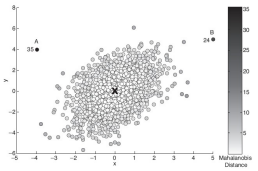


Figure 10.3. Mahalanobis distance of points from the center of a two-dimensional set of 2002 points.

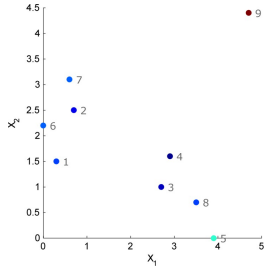
$\bar{\mathbf{x}}$ Mean vector

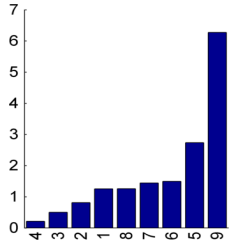
$\mathbf{S}$ Covariance matrix

Mahalanobis distance

- Distance to the mean of data, taking covariance into account

Data I: Cats , dogs and dinosaurs

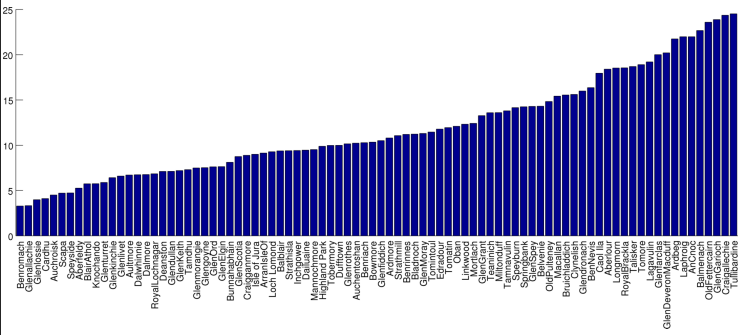




Mahalanobis distance

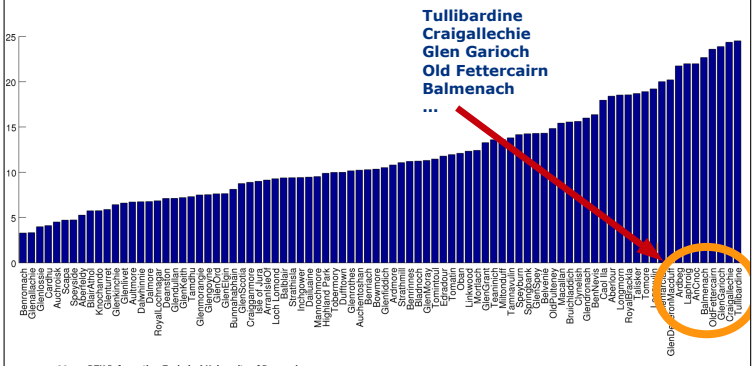
- Distance to the mean of data, taking covariance into account

Data II: Whisky



Mahalanobis distance

- Distance to the mean of data, taking covariance into account



Distance to k-nearest neighbor

- Measure the distance to the k'th nearest neighbor

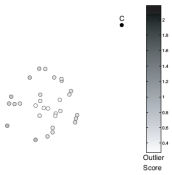


Figure 10.4. Outlier score based on the distance to fifth nearest neighbor.

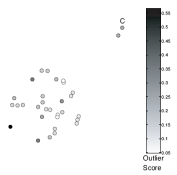


Figure 10.5. Outlier score based on the distance to the first nearest neighbor. Nearby outliers have low outlier scores.

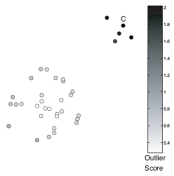


Figure 10.6. Outlier score based on distance to the fifth nearest neighbor. A small cluster becomes an outlier.

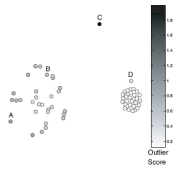
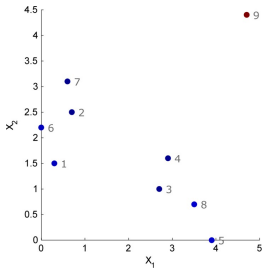


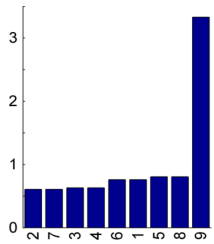
Figure 10.7. Outlier score based on the distance to the fifth nearest neighbor. Clusters of differing density.

Distance to k-nearest neighbor

- Measure the distance to the 1st nearest neighbor

Data I: Cats , dogs and dinosaurs

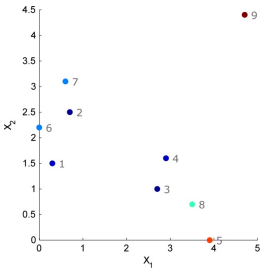


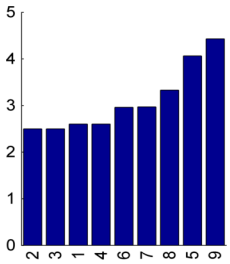


Distance to kth Nearest neighbour

- Measure the distance to the 5th nearest neighbor

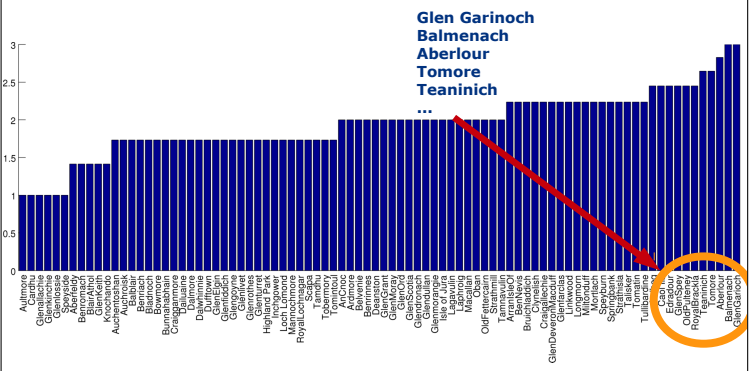
Data I: Cats , dogs and dinosaurs





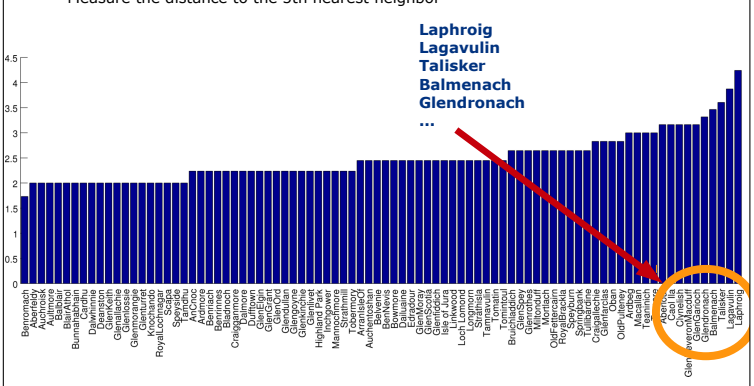
Distance to k-nearest neighbor

- Measure the distance to the 1st nearest neighbor



Distance to k-nearest neighbor

- Measure the distance to the 5th nearest neighbor



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

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

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

$$\text{average relative density}(\boldsymbol{x}, k) = \frac{\text{density}(\boldsymbol{x}, k)}{\frac{1}{k} \sum_{\boldsymbol{y} \in N(\boldsymbol{x},k)} \text{density}(\boldsymbol{y}, k)}$$

$N(\boldsymbol{x}, k)$ The set of k nearest neighbors

Consider the pairwise distance matrix given to the left. What is the density and average relative density of the first observation for k=3?

$d(\mathbf{x}_i,\mathbf{x}_j)$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$
$\mathbf{x}_1$	0	2.5	2.4	4.0	0.8	0.6	3.3
$\mathbf{x}_2$	2.5	0	0.6	1.6	2.9	3.0	1.1
$\mathbf{x}_3$	2.4	0.6	0	1.9	3.0	2.7	1.0
$\mathbf{x}_4$	4.0	1.6	1.9	0	4.5	4.6	3.8
$\mathbf{x}_5$	0.8	2.9	3.0	4.5	0	1.1	3.9
$\mathbf{x}_6$	0.6	3.0	2.7	4.6	1.1	0	3.8
$\mathbf{x}_7$	3.3	1.1	1.0	3.8	3.9	3.8	0

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

$$\text{average relative density}(\boldsymbol{x},k)=\frac{\text{density}(\boldsymbol{x},k)}{\frac{1}{k}\sum_{\boldsymbol{y}\in N(\boldsymbol{x},k)}\text{density}(\boldsymbol{y},k)}$$

Consider the pairwise distance matrix given to the left. What is the density and average relative density of the first observation for k=3?

$d(\mathbf{x}_i,\mathbf{x}_j)$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$
$\mathbf{x}_1$	0	2.5	2.4	4.0	0.8	0.6	3.3
$\mathbf{x}_2$	2.5	0	0.6	1.6	2.9	3.0	1.1
$\mathbf{x}_3$	2.4	0.6	0	1.9	3.0	2.7	1.0
$\mathbf{x}_4$	4.0	1.6	1.9	0	4.5	4.6	3.8
$\mathbf{x}_5$	0.8	2.9	3.0	4.5	0	1.1	3.9
$\mathbf{x}_6$	0.6	3.0	2.7	4.6	1.1	0	3.8
$\mathbf{x}_7$	3.3	1.1	1.0	3.8	3.9	3.8	0

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

$$\text{average relative density}(\boldsymbol{x},k)=\frac{\text{density}(\boldsymbol{x},k)}{\frac{1}{k}\sum_{\boldsymbol{y}\in N(\boldsymbol{x},k)}\text{density}(\boldsymbol{y},k)}$$

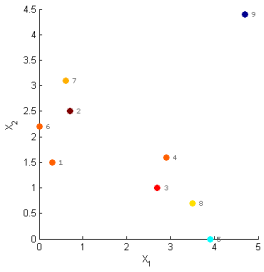
density($\mathbf{x}_1$,3)=[1/3(0.6+0.8+2.4)]$^{-1}$=3/3.8
density($\mathbf{x}_6$,3)=[1/3(0.6+1.1+2.7)]$^{-1}$=3/4.4
density($\mathbf{x}_5$,3)=[1/3(0.8+1.1+2.9)]$^{-1}$=3/4.8
density($\mathbf{x}_3$,3)=[1/3(0.6+1.0+1.9)]$^{-1}$=3/3.5

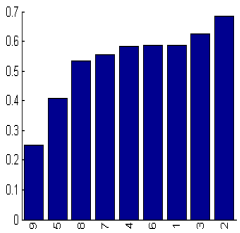
Av. Rel. Density($\mathbf{x}_1$,3)=[3/3.8]/[1/3 (3/4.4+3/4.8+3/3.5)]=1.0945

Inverse distance density estimation

- KNN density (5 nearest neighbors)

Data I: Cats , dogs and dinosaurs

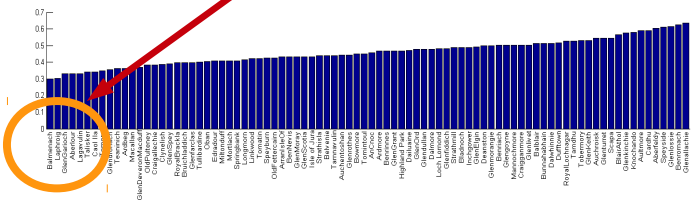




Inverse distance density estimation

- KNN density (5 nearest neighbors)

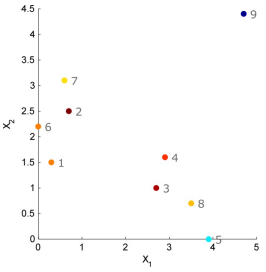
Balmenach
Laphroig
GlenGarioch
Aberlour
Lagevulin
...

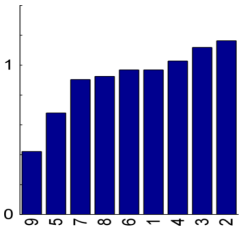


Average Relative density

- Average Relative KNN density (5 nearest neighbors)

Data I: Cats , dogs and dinosaurs

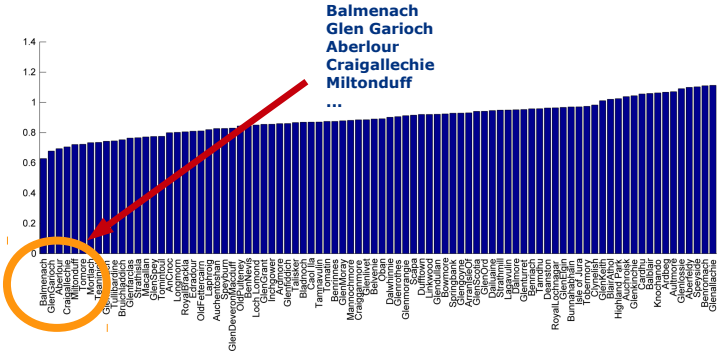




Average relative density

- Average relative KNN density (5 nearest neighbors)

Balmenach
Glen Garioch
Aberlour
Craigallechie
Miltonduff
...



Results using different methods

- **Univariate Normal distribution**
  - Ardbeg
  - Lagavulin
  - Laphroig
- **Kernel Density Estimation**
  - Balmenach
  - Glen Garioch
  - Aberlour
  - Tomore
  - Caolla
- **Mahalanobis distance**
  - Tullibardine
  - Craigallechie
  - Glen Garioch
  - Old Fettercairn
  - Balmenach
- **Distance to nearest neighbor**
  - Glen Garioch
  - Balmenach
  - Aberlour
  - Tomore
  - Teaninich
- **Distance to 5th nearest neighbor**
  - Laphroig
  - Lagavulin
  - Talisker
  - Balmenach
  - Glendronach
- **KNN density**
  - Balmenach
  - Laphroig
  - Glen Garioch
  - Aberlour
  - Lagavulin
- **KNN average relative density**
  - Balmenach
  - Glen Garioch
  - Aberlour
  - Craigallechie
  - Miltonduff

Common: Balmenach, Glen Garioch, Laphroig, Aberlour, Tomore, Lagavulin, Craigallechie

Example of exam questions

QI: What is the average relative density for observation 2 (i.e. $\mathbf{x}_2$) for k=2 nearest neighbours?

A:1/5
B:3/10
C:7/10
D:1

$d(\mathbf{x}_i,\mathbf{x}_j)$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$
$\mathbf{x}_1$	0	2.0	0.2	0.9	0.2
$\mathbf{x}_2$	2.0	0	1.5	0.5	2.0
$\mathbf{x}_3$	0.2	1.5	0	1.2	1.4
$\mathbf{x}_4$	0.9	0.5	1.2	0	1.0
$\mathbf{x}_5$	0.2	2.0	1.4	1.0	0

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

$$\text{avg. rel. den.}(\boldsymbol{x}, k) = \frac{\text{density}(\boldsymbol{x}, k)}{\frac{1}{k} \sum_{\boldsymbol{y} \in N(\boldsymbol{x},k)} \text{density}(\boldsymbol{y}, k)}$$

QI: C