University of Southampton

COMP6237 Data Mining
Lecture 12: Outlier Detection

Zhiwu Huang

Zhiwu.Huang@soton.ac.uk

Lecturer (Assistant Professor) @ VLC of ECS

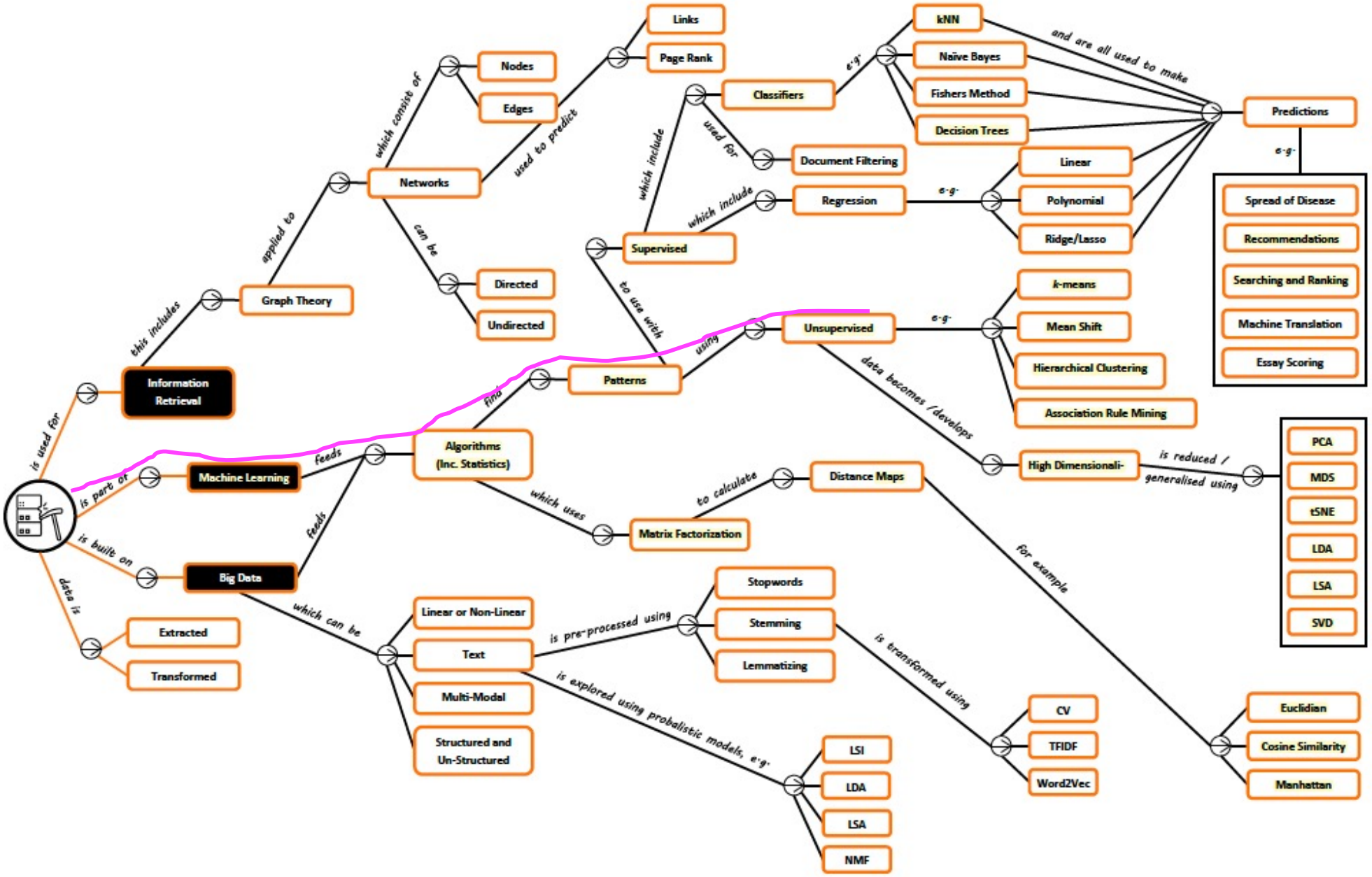
University of Southampton

Lecture slides available here:

http://comp6237.ecs.soton.ac.uk/zh.html

(Thanks to Prof. Jonathon Hare and Dr. Jo Grundy for providing the lecture materials used to develop the slides.)

Outlier Detection – **Roadmap**



Outlier Detection – **Textbook**

9 Anomaly Detection

*In anomaly detection, the goal is to find objects that do not conform to normal patterns or behavior. Often, anomalous objects are known as **outliers**, since, on a scatter plot of the data, they lie far away from other data points. Anomaly detection is also known as **deviation detection**, because anomalous objects have attribute values that deviate significantly from the expected or typical attribute values, or as **exception mining**, because anomalies are exceptional in some sense. In this chapter, we will mostly use the terms anomaly or outlier. There are a variety of anomaly detection approaches from several areas, including statistics, machine learning, and data mining. All try to capture the idea that an anomalous data object is unusual or in some way inconsistent with other objects.*

- Introduction to Data Mining, *P. Tan et al*
  https://www-users.cse.umn.edu/~kumar001/dmbook/index.php

Outlier Detection – **Overview (1/3)**

Do all of these look right?

TRANSACTION DETAIL

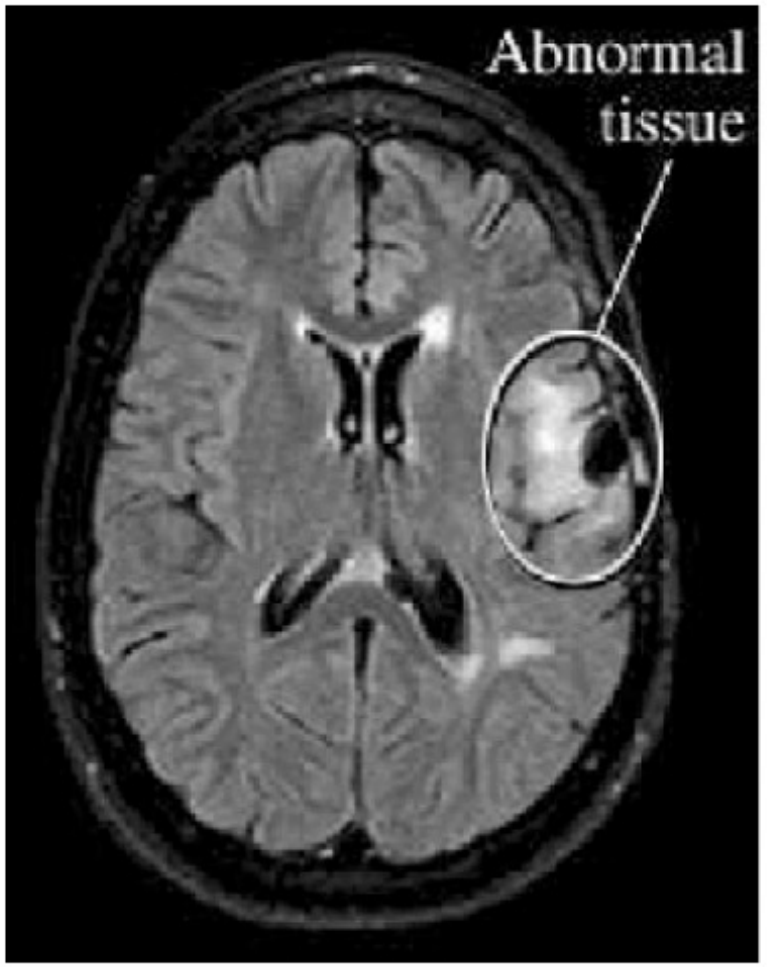
DATE	DESCRIPTION	AMOUNT	BALANCE
	Beginning Balance		**$771.41**
07/16	Sgc Live Learnin Dirdepdd PPD ID:	**81.00**	852.41
07/16	Tutors L Dirdepdd PPD ID:	**60.00**	912.41
07/16	Card Purchase 07/14 Wines And Sp Los Angeles CA Card 8240	-28.96	883.45
07/16	Card Purchase 07/15 Store # 144 Glendale CA Card 8240	-6.55	876.90
07/16	Card Purchase 07/15 House Glendale CA Card 8240	-22.54	854.36
07/16	Card Purchase With Pin 07/16 Metrolink 900 Wilshire Los Angeles CA Card 8240	-9.75	844.61
07/19	Card Purchase 07/15 Coffee Glendale CA Card 8240	-15.18	829.43
07/19	Card Purchase 07/16 Donuts #21 Los Angeles CA Card 8240	-3.50	825.93
07/19	Card Purchase 07/16 Store 22525 Los Angeles CA Card 8240	-5.25	820.68
07/19	Card Purchase 07/16 Lax Airp Los Angeles CA Card 8240	-10.77	809.91
07/19	07/17 Online Transfer To Sav ...3692 Transaction#:	-312.00	497.91
07/23	Online Transfer From Sav ...3692 Transaction#:	**1,100.00**	1,597.91
07/26	Card Purchase 07/23 Air00121903508 Fort Worth TX Card 8240	-1,100.95	496.96
07/27	Card Purchase 07/26 Bucktown (655) Chicago IL Card 8240	-17.53	479.43
07/28	Payment Received 07/28 Direct CA Card 8240	**470.00**	949.43
07/29	Card Purchase Return 06/29 Airway00002394707 Washington DC Card 8240	**548.91**	1,498.34
07/29	07/29 Online Transfer To Sav ...3692 Transaction#:	-449.00	1,049.34

https://www.xcover.com/en/help/documents-bank-statement

Outlier Detection – Overview (2/3)

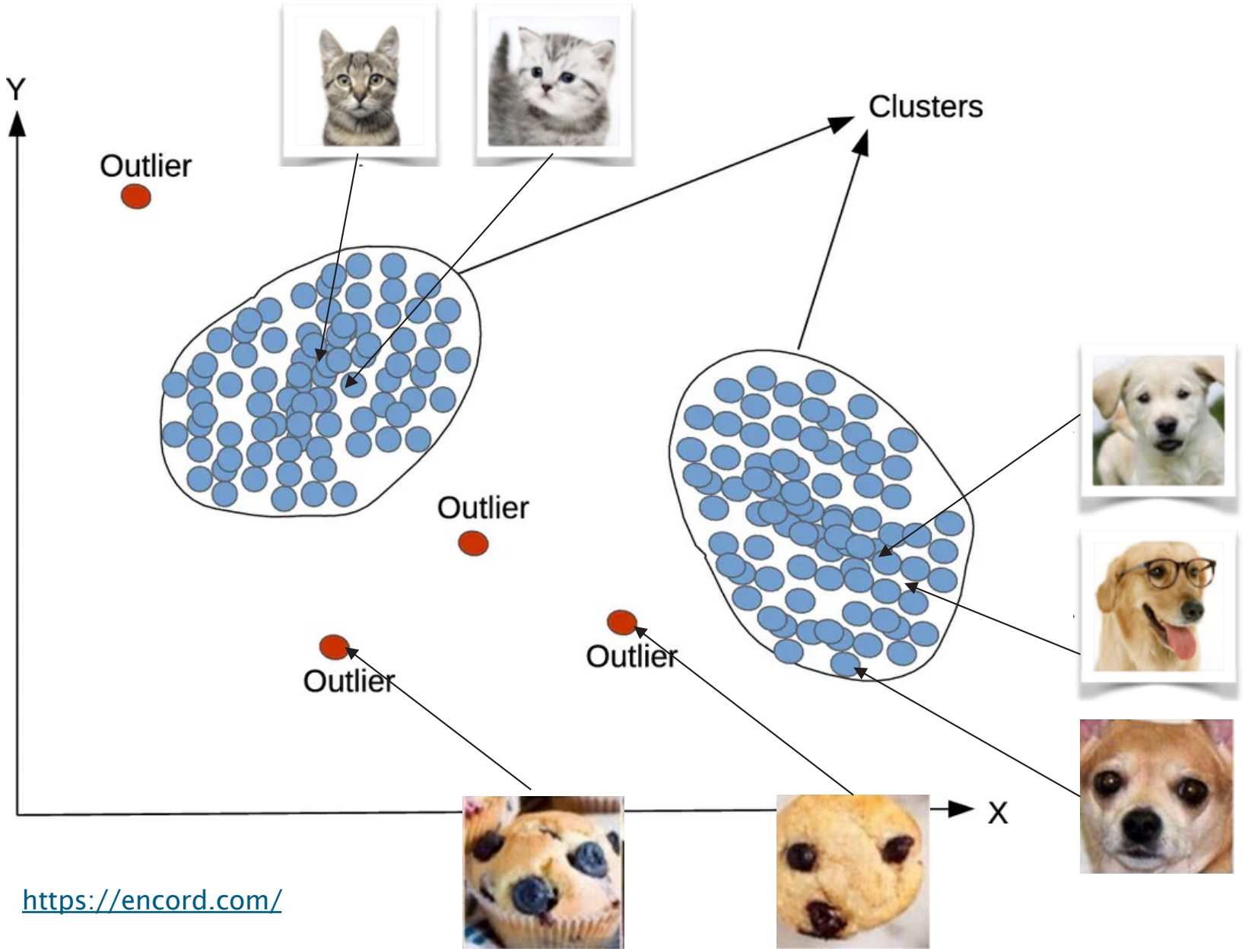
If you see lots of scans that look like this:

Then it is easier to see that there is something wrong here



Credit: Jo Grundy

Outlier Detection – **Overview (3/3)**



https://encord.com/

Outlier Detection – **Learning Outcomes**

- **LO1**: Demonstrate an understanding of techniques for outlier detection, such as: (exam)
  - Applying extreme value analysis method
  - Understanding the key idea and steps of the learned clustering methods for outlier detection
  - Discussing the advantages and disadvantages of the learned outlier detection approaches
- **LO2**: Implement the learned algorithms for outlier detection (coursework)

Assessment hints: Multi-choice Questions (single answer: concepts, calculation etc)

- *Textbook Exercises: textbooks (Programming + Mining)*
- *Other Exercises: https://www-users.cse.umn.edu/~kumar001/dmbook/sol.pdf*
- *ChatGPT or other AI-based techs*

Outlier Detection – Approach

We can model data using a Gaussian distribution:
Univariate:

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

Estimate mean:

- $\mu = \frac{1}{N} \sum_{i=1}^{N} x_i$

Estimate standard deviation:

- $\sigma = \frac{1}{N} \sum_{i=1}^{N} (x_i - \mu)$

Credit: Jo Grundy

Outlier Detection – Approach

A Data mining approach:

- Model the data
- What does not fit is outlier

Can use many different models
Need:

- a measure of fit

Demo: Jupyter notebook

Credit: Jo Grundy

Outlier Detection – Extreme Value Statistics

University of Southampton

Extreme Value Statistics

A way to characterise extreme values using a rule similar to the central limit theorem.

Also known as the Fisher-Tippet theorem

$$f(x) \approx \frac{1}{\beta} e^{\frac{x-\mu}{\beta} - e^{\frac{x-\mu}{\beta}}}$$

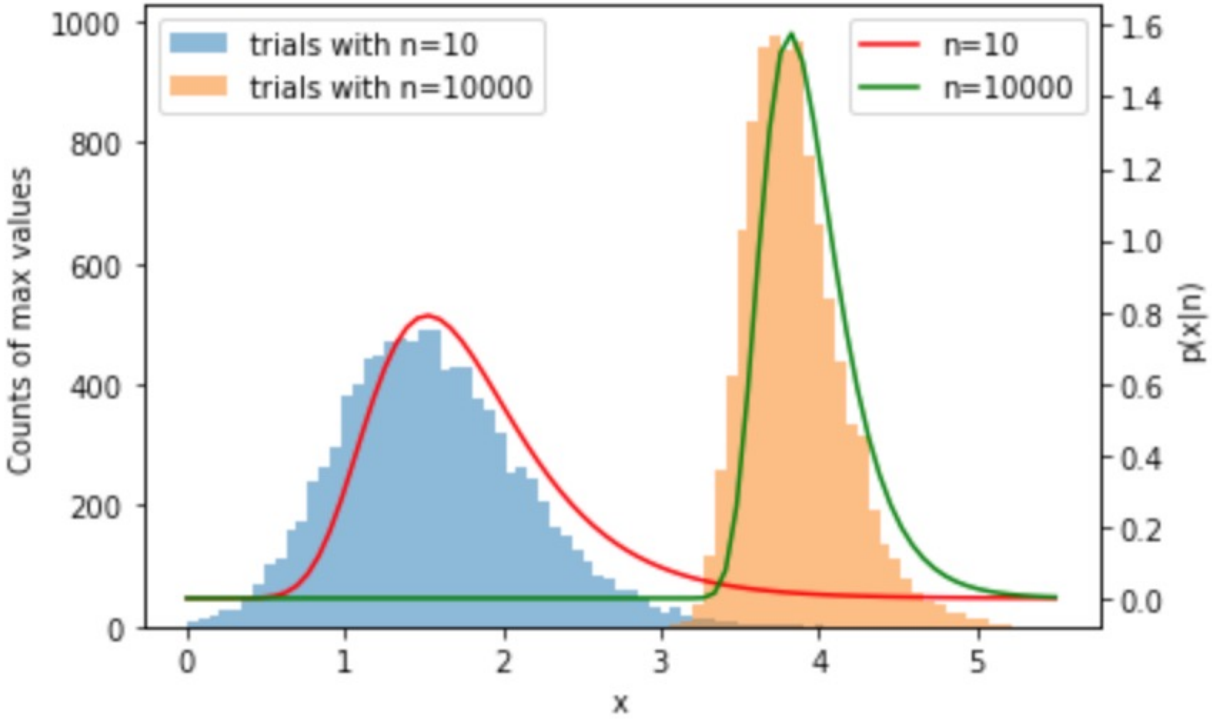
- The limiting distribution of these extremes must be one of three types of generalized extreme value distributions - Gumbel, Fréchet or Weibull
- It makes a distinction between real anomalies versus values that are simply extremes of the primary data distribution.

http://www.nematrian.com/Pages/ExtremeValueTheoryCombined.pdf

Demo: Jupyter notebook

Credit: Jo Grundy

Outlier Detection – Extreme Value Statistics




```python
def mc_max(n, m):
    """
    generates n numbers, records max value,
    m times,
    uses Gausian with mean 0 and stddev 1
    returns max value for each one in array
    """
    max_arr = []
    for i in range(m):
        max_arr.append(max(np.random.randn(n)))
    return np.array(max_arr)
```

The Weibull distribution is used here to give a probability that a value is an maximal value from a normal distribution. With more samples, the distribution is more clearly defined.

See e.g. S.J.Roberts IEE Proceedings 2000, 147,6,363-367

https://ieeexplore.ieee.org/document/889967

Credit: Jo Grundy

Outlier Detection – Gaussian Distribution

We can model the data using a multivariate Gaussian distribution:

$$p(x) = \frac{1}{2\pi^{\frac{p}{2}}\sqrt{|C|}} \exp\{-\frac{1}{2}(\boldsymbol{x} - \boldsymbol{m})^T C^{-1}(\boldsymbol{x} - \boldsymbol{m})\}$$

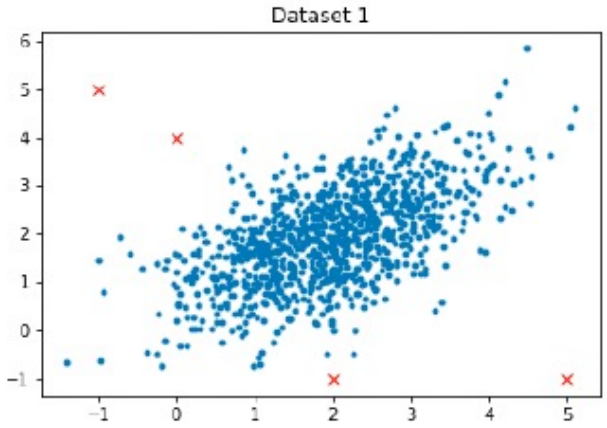
Covariance and mean can be estimated from the data.. how?
mean $= \boldsymbol{m} = \frac{1}{N}\sum_i^N x_i$
covariance is proportional to the inner product of the mean centred data
or

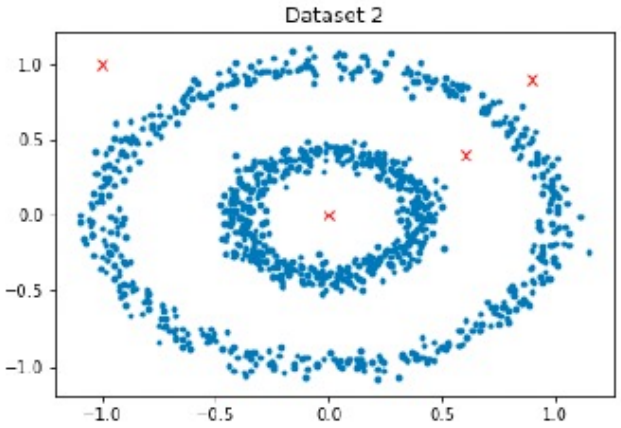
$$C = \frac{1}{N}\sum_i^N (\boldsymbol{x_i} - \boldsymbol{m})(\boldsymbol{x_i} - \boldsymbol{m})^T$$

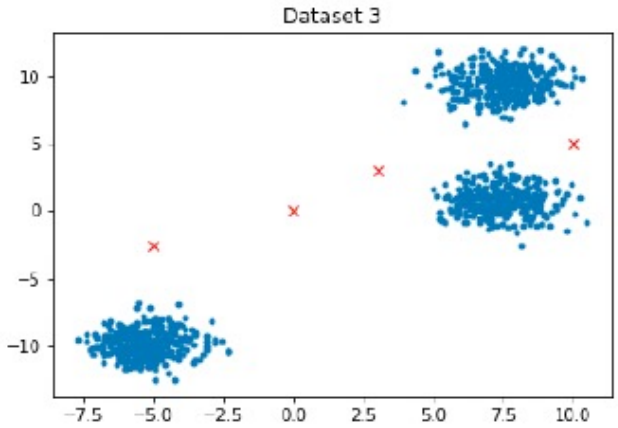
Credit: Jo Grundy

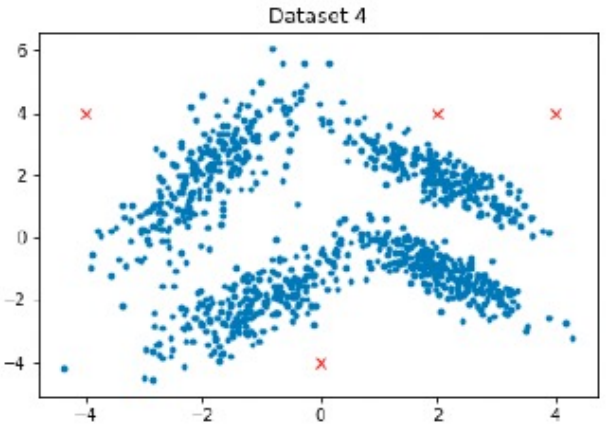
Outlier Detection – Gaussian Distribution

For example:





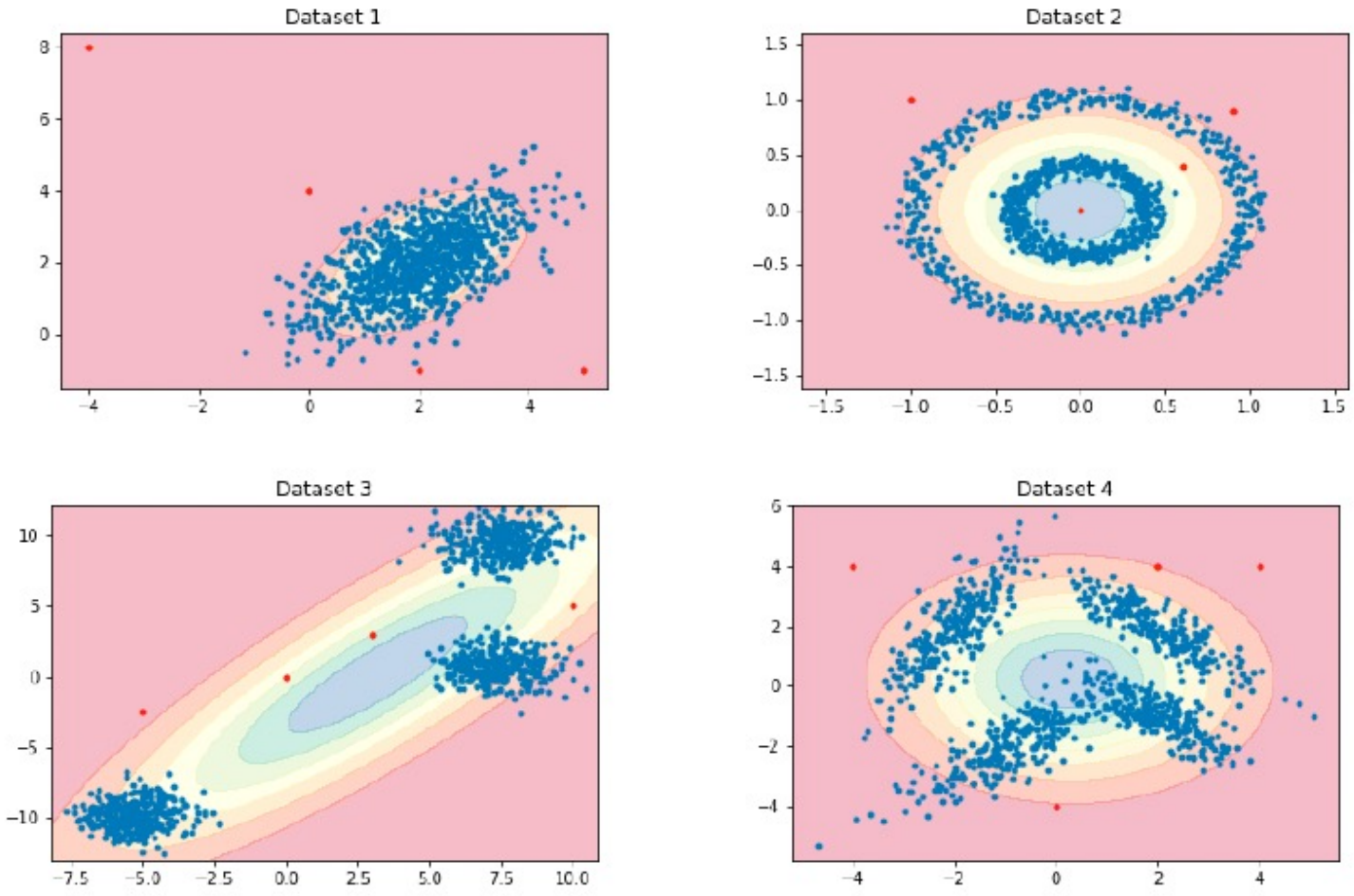




Credit: Jo Grundy

Outlier Detection – Gaussian Distribution

Also.. Does not fit multimodal or oddly shaped distributions



Credit: Jo Grundy

Outlier Detection – Gaussian Mixture Model

Try using more than one Gaussian: **Gaussian Mixture Model**

$$\sum_{k}^{K} \pi_k p(x|\mu, C)$$

Estimate weighting π, mean μ and covariance C?
If we knew the weights, mean and covariance, we could calculate the probability
if we knew the probabilities, we could calculate the weights, mean and covariance
Expectation maximisation: generalisation of K Means

Credit: Jo Grundy

Outlier Detection – Gaussian Mixture Model

$$L(\theta|X_1,\ldots,X_n) = \prod_{i=1}^{n}\sum_{k=1}^{K}\pi_k N(x_i;\mu_k,\sigma_k^2)$$

$$\ell(\theta) = \sum_{i=1}^{n}\log\left(\sum_{k=1}^{K}\pi_k N(x_i;\mu_k,\sigma_k^2)\right)$$

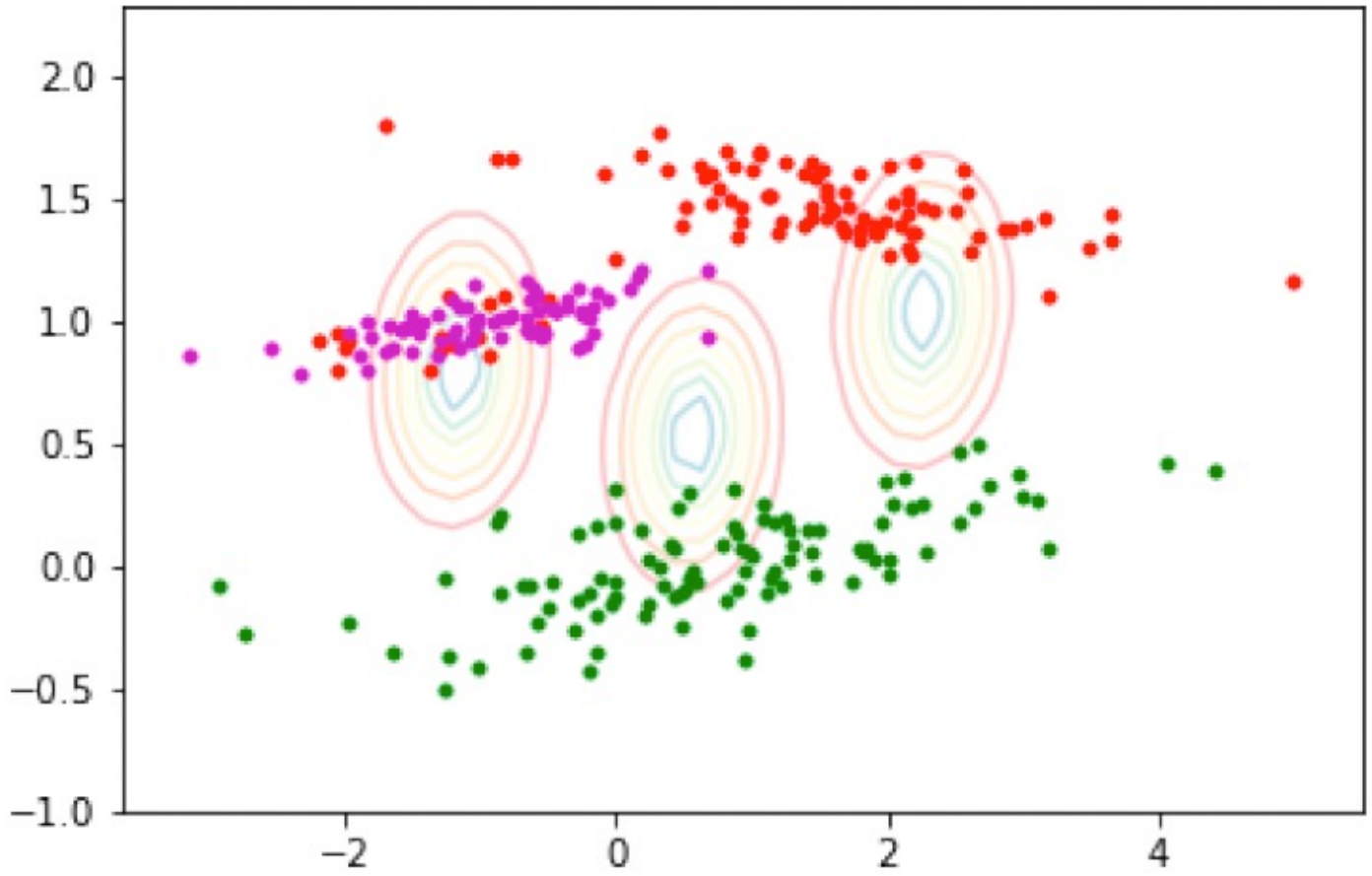
Algorithm 1: GMM

Data: X ($n \times p$ data), k Gaussians to use
Initialise π_k, μ_k and C_k ;
while *not converged* **do** ← *Evaluate the log-likelihood with the new parameter*
  for $x_i \in X$ **do** ← *E-step: Evaluate the posterior probabilities*
    for $j \in 1,\ldots,k$ **do**
      responsibilities $r_{i,j} = p(x_i|\mu_j, C_j)$;
    end
  end
  for $j \in 1,\ldots,k$ **do** ← *M-step: Estimate new parameters*
    $N_j = \sum_{i=0}^{n} r_{i,j}$;
    $\pi_j = \frac{N_j}{N}$;
    $\mu_j = \frac{1}{N_j}\sum_{i=0}^{n} r_{i,j}x_i$;
    $C_j = \frac{1}{N_j}\sum_{i=0}^{n} r_{i,j}(x_i - \mu_j)(x_i - \mu_j)^T$;
  end
end

Credit: Jo Grundy

Outlier Detection – Gaussian Mixture Model

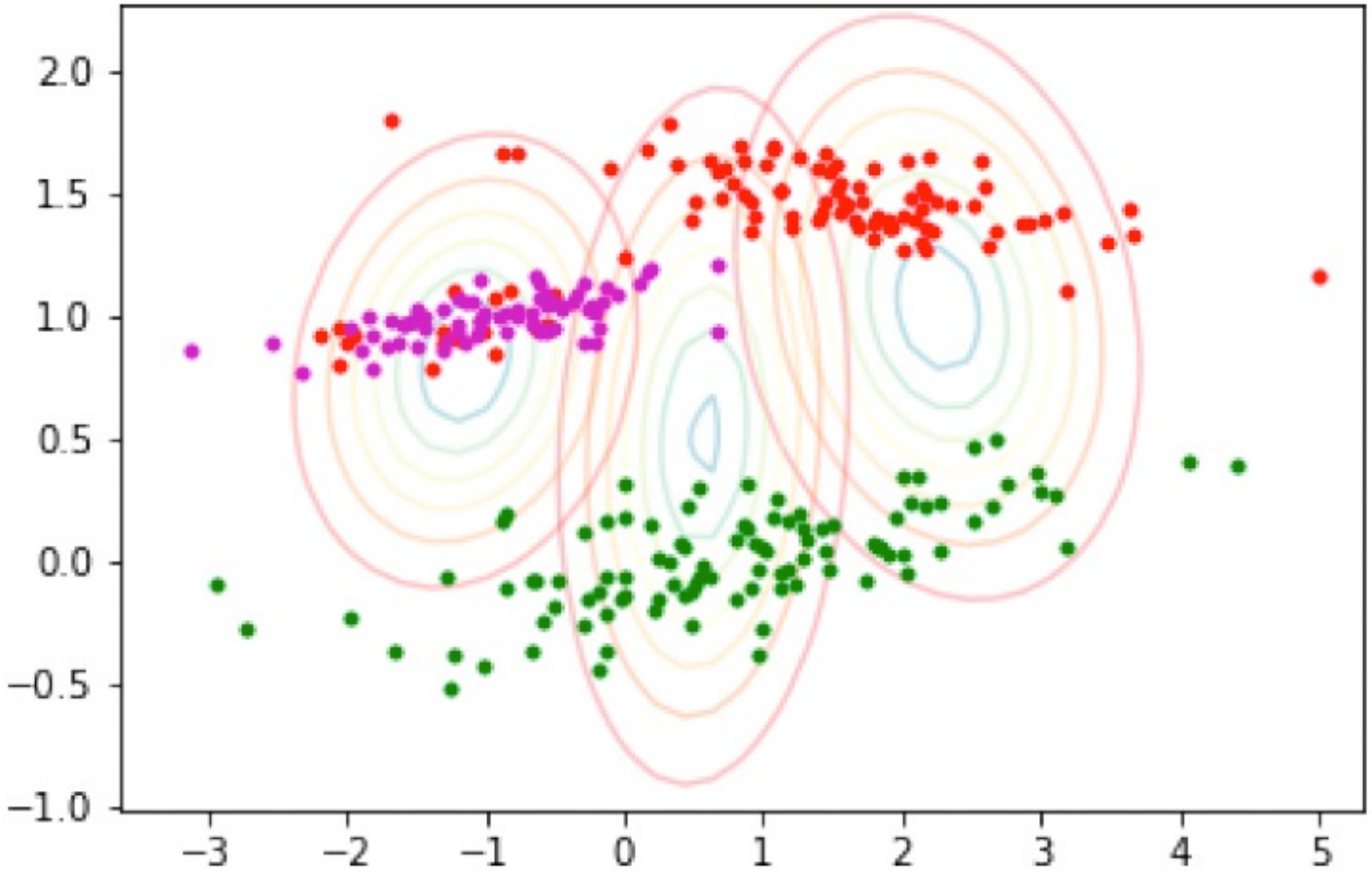
Step by step:



Credit: Jo Grundy

Outlier Detection – Gaussian Mixture Model

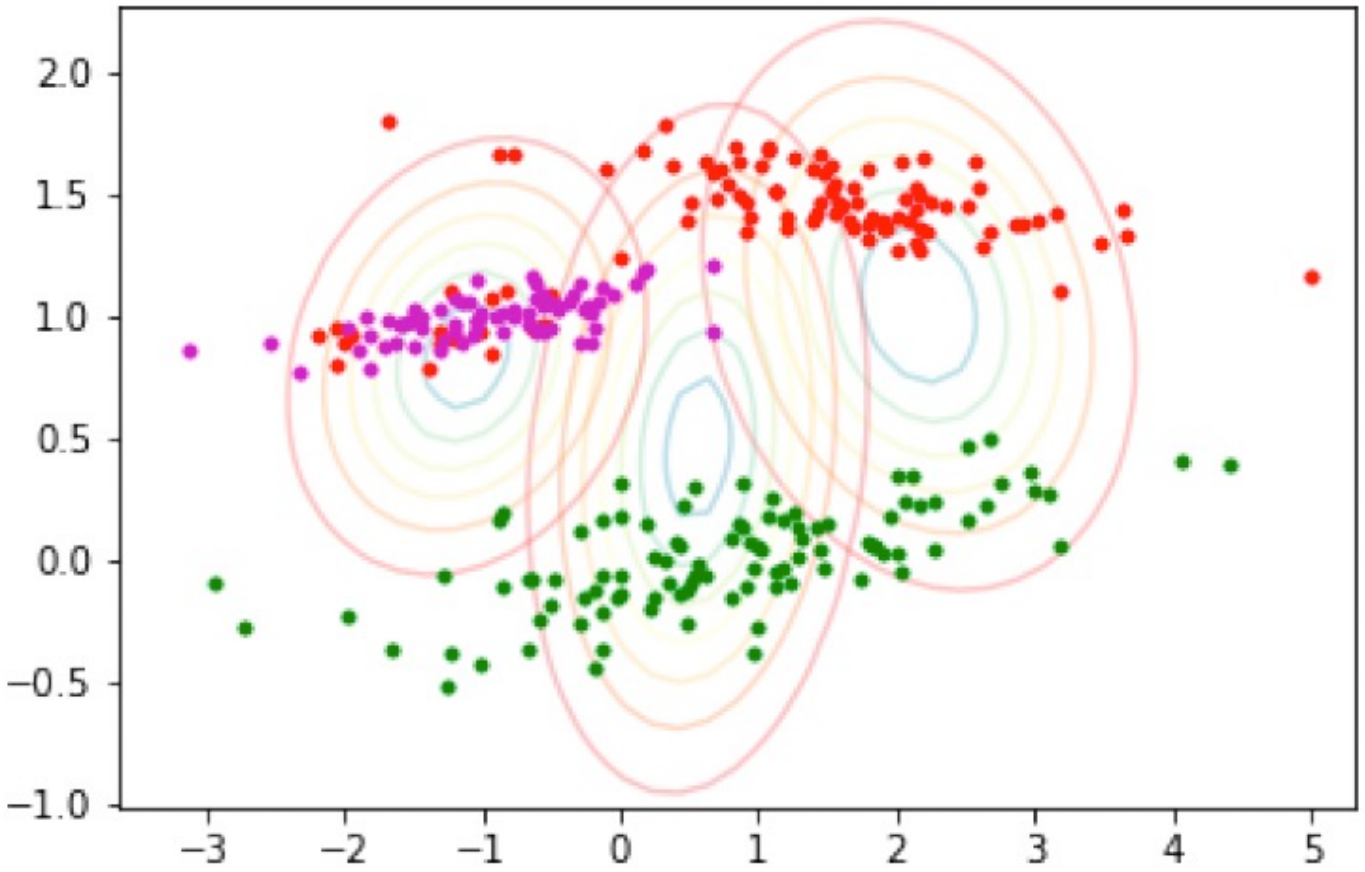
Step by step:



Credit: Jo Grundy

Outlier Detection – Gaussian Mixture Model

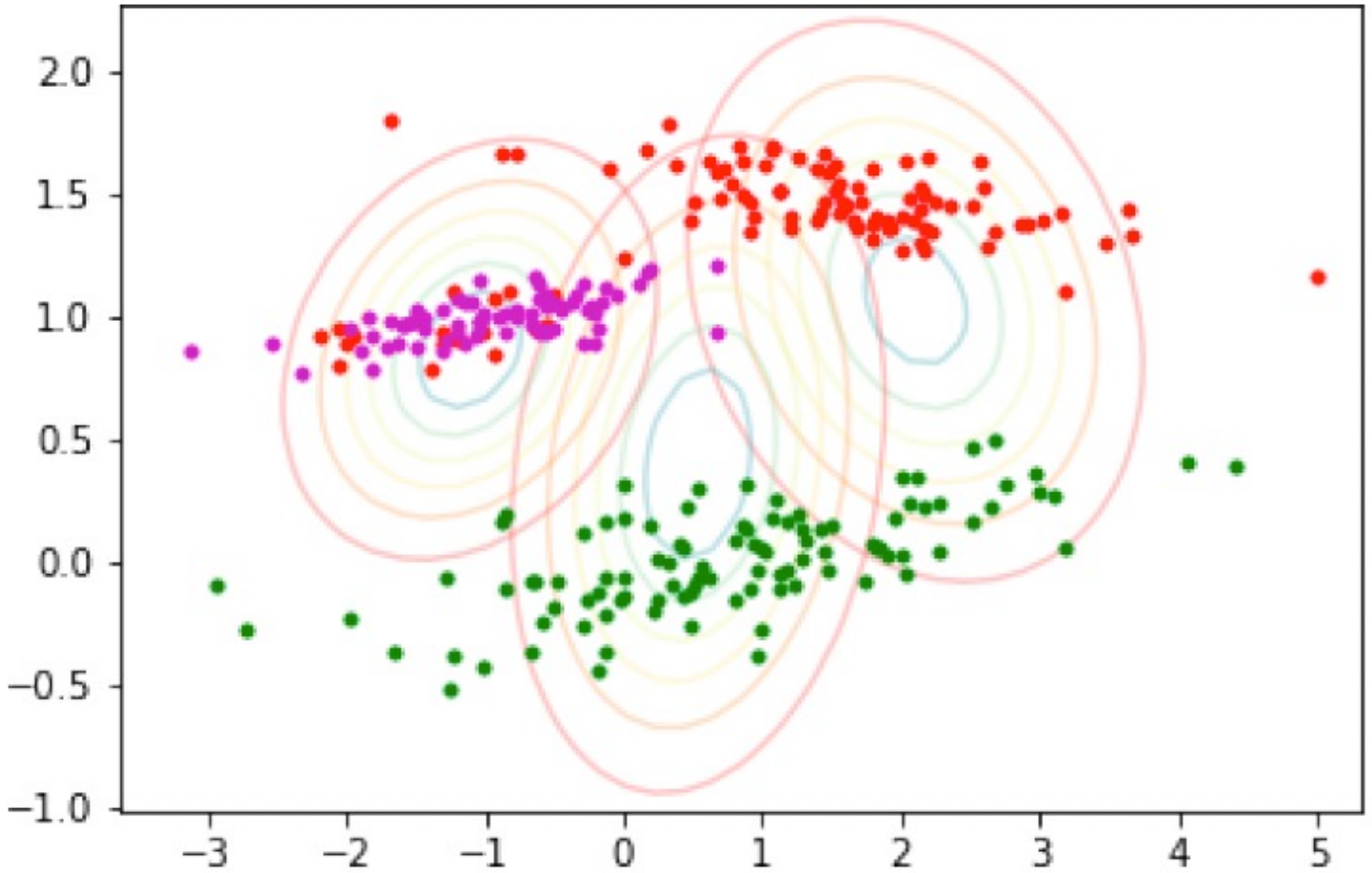
Step by step:



Credit: Jo Grundy

Outlier Detection – Gaussian Mixture Model

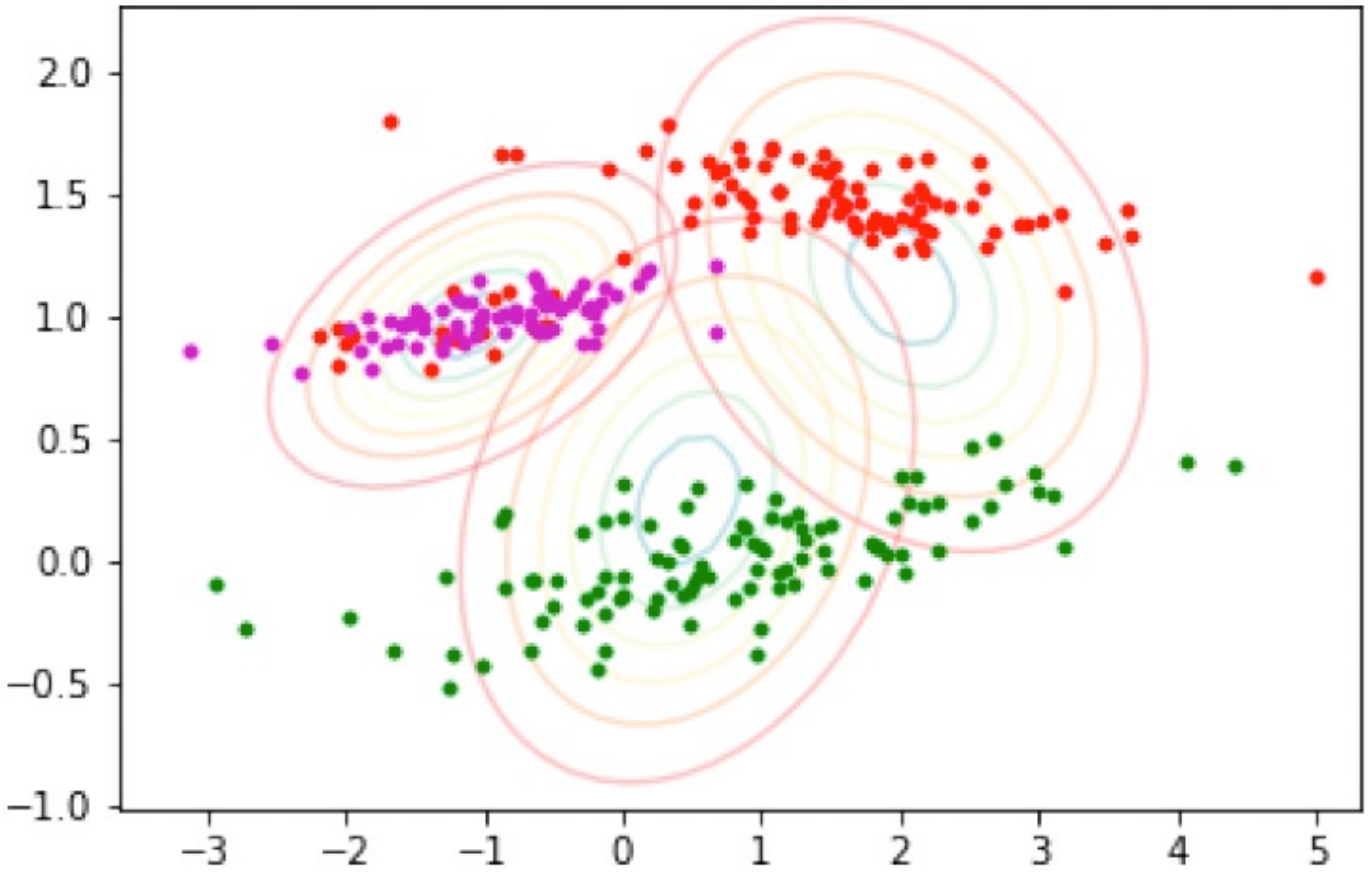
Step by step:



Credit: Jo Grundy

Outlier Detection – Gaussian Mixture Model

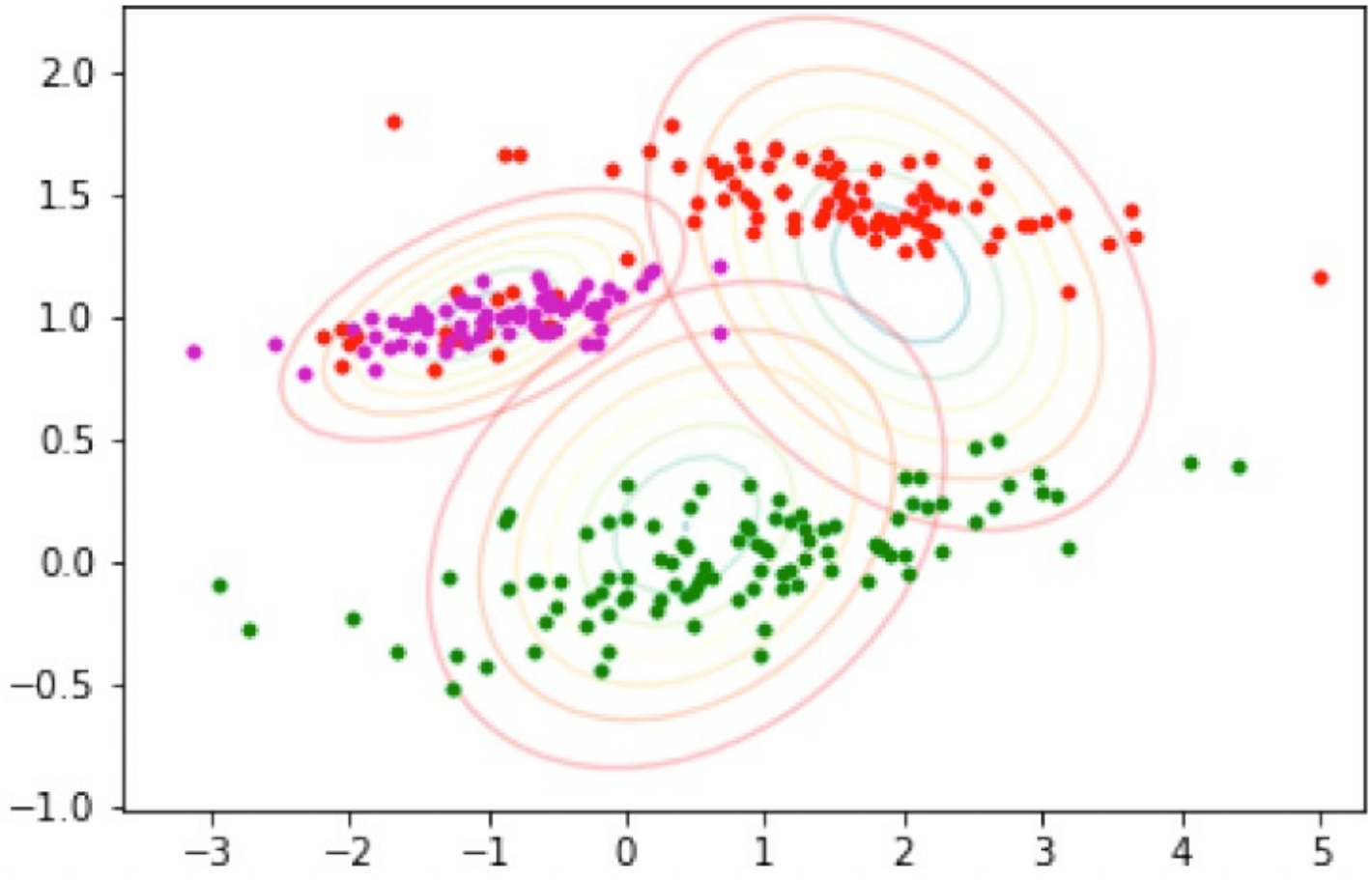
Step by step:



Credit: Jo Grundy

Outlier Detection – Gaussian Mixture Model

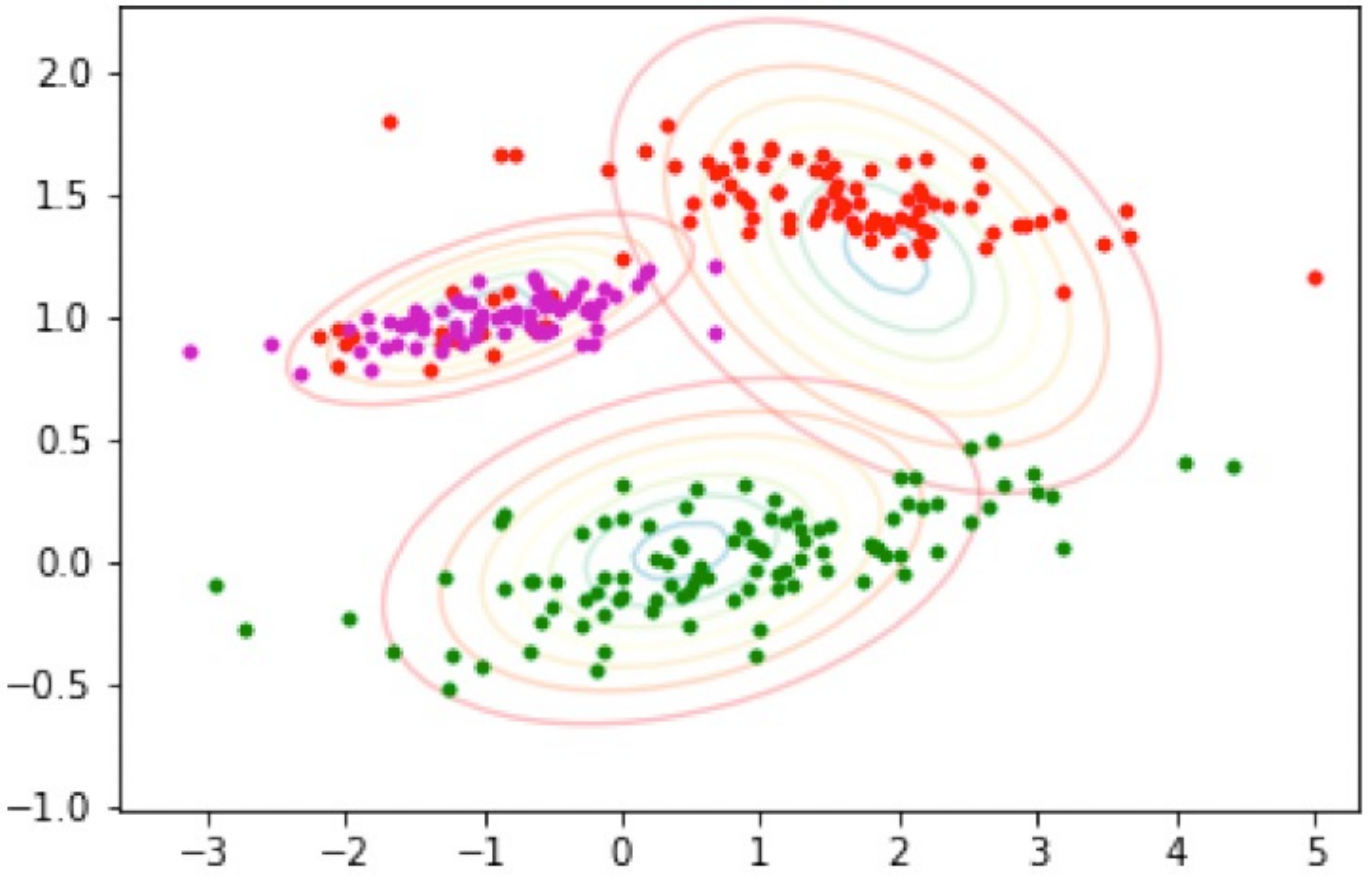
Step by step:



Credit: Jo Grundy

Outlier Detection – Gaussian Mixture Model

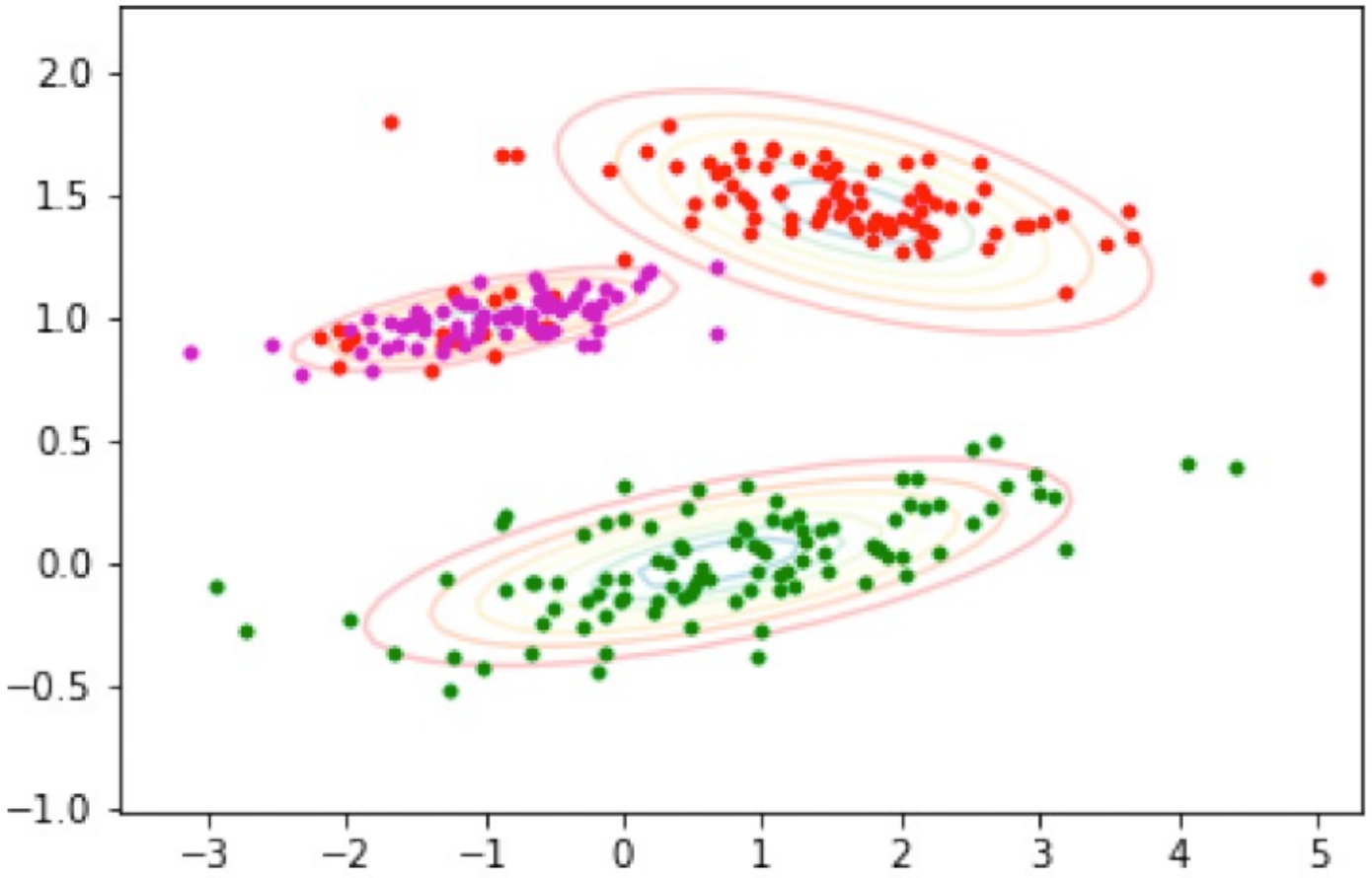
Step by step:



Credit: Jo Grundy

Outlier Detection – Gaussian Mixture Model

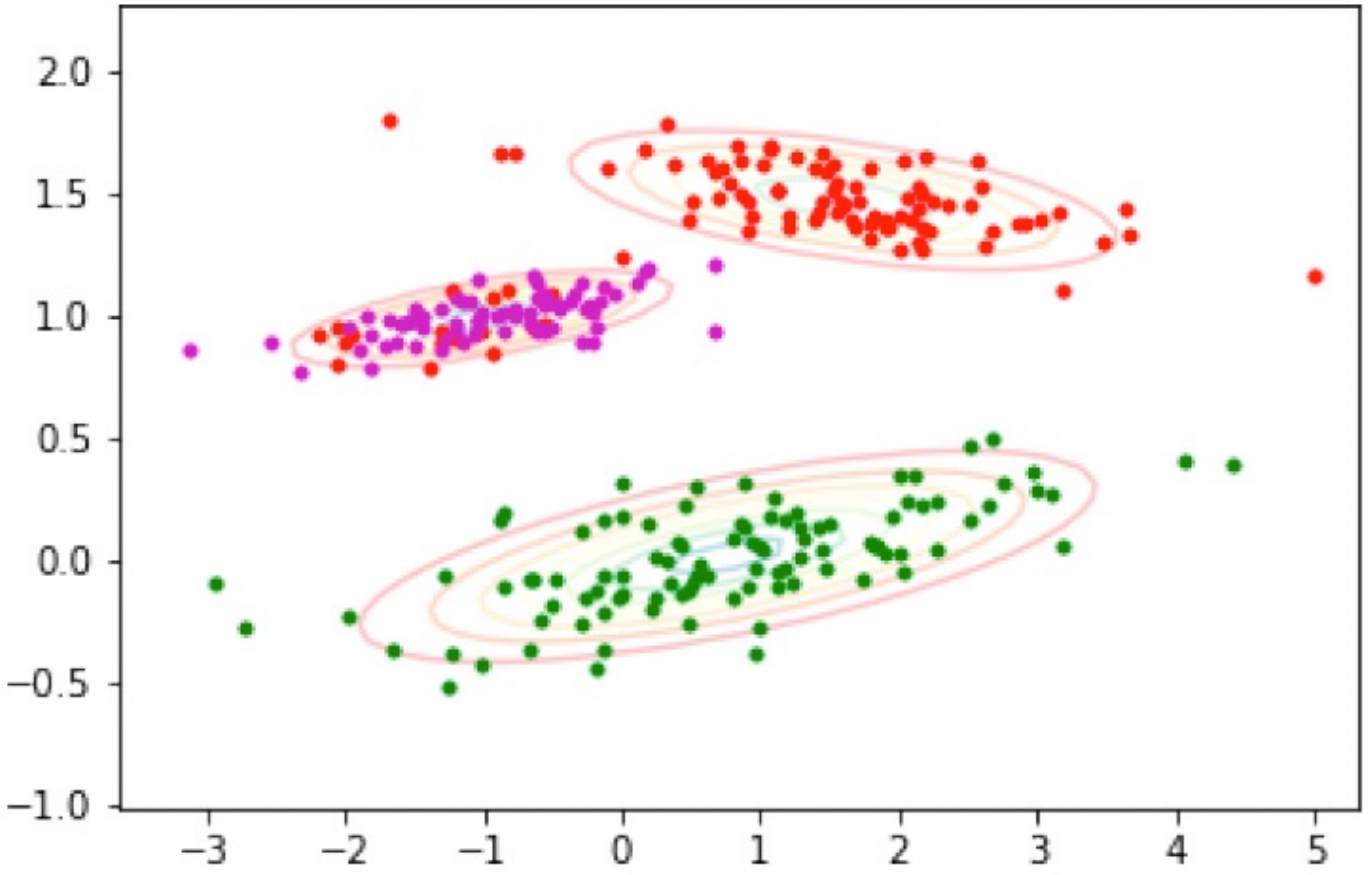
Step by step:



Credit: Jo Grundy

Outlier Detection – Gaussian Mixture Model

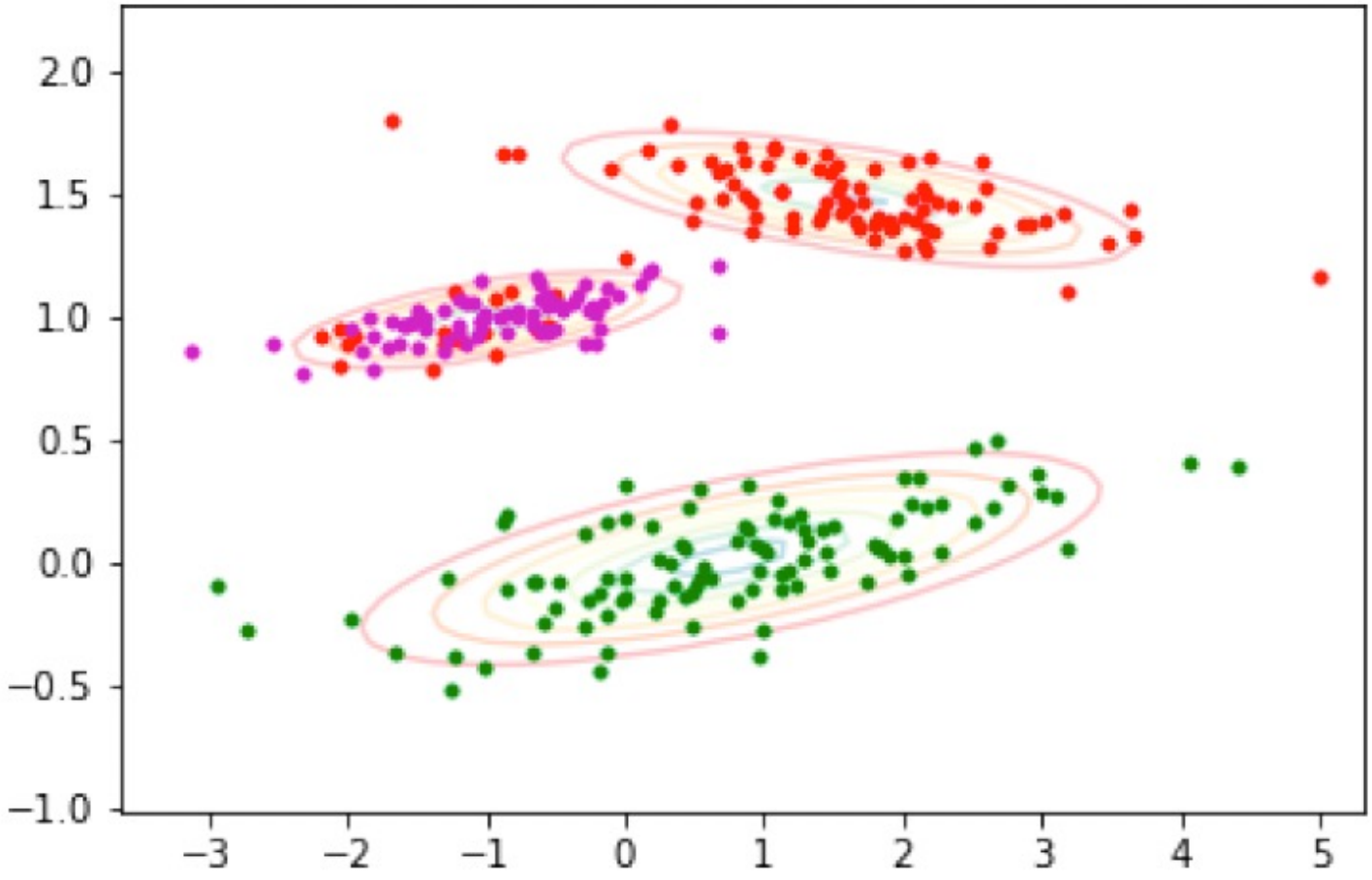
Step by step:



Credit: Jo Grundy

Outlier Detection – Gaussian Mixture Model

Step by step:



Credit: Jo Grundy

Outlier Detection – Gaussian Mixture Model

Initialisation:

- randomly - can cause issues
- use K Means - works quite well

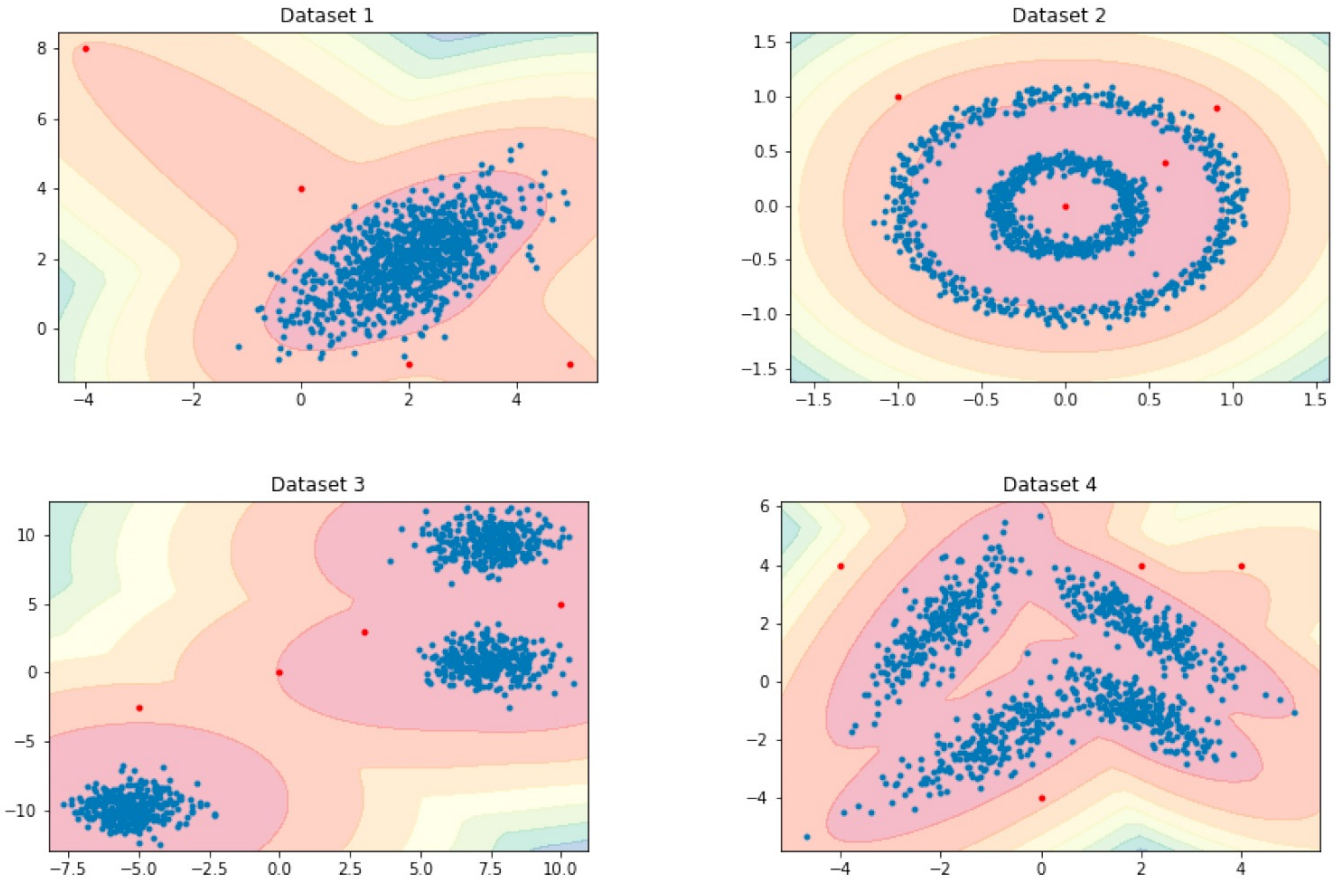
Convergence:

- Can check for an increase in the total probability
- $\sum_{i=0}^{k} \sum_{j=1}^{n} r_{i,j}$ $\quad \ell(\theta) = \sum_{i=1}^{n} \log\left(\sum_{k=1}^{K} \pi_k N(x_i; \mu_k, \sigma_k^2)\right)$
- best to use logs

Credit: Jo Grundy

Outlier Detection – Gaussian Mixture Model

Test on datasets:



Works reasonably well for the three Gaussian distributions. Note sensitivity to outliers. What about the circular data set?

Credit: Jo Grundy

Outlier Detection – DBSCAN

DBSCAN - good for outlier detection as well as clustering
Recap: Density Based Spatial Clustering and Noise
Needs:

- maximum radius $N_\epsilon(\mathbf{x}) = B_d(\mathbf{x}, \epsilon) = \{\mathbf{y} \mid \delta(\mathbf{x}, \mathbf{y}) \leq \epsilon\}$
- minimum number $|N_\epsilon(\mathbf{x})| \geq minpts$

Max radius is the limit on which to look for neighbours
Min number is the lower limit on what can be in a cluster

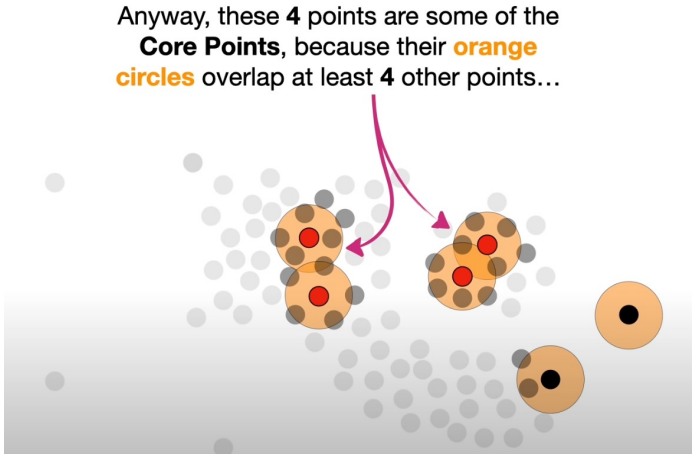
Credit: Jo Grundy

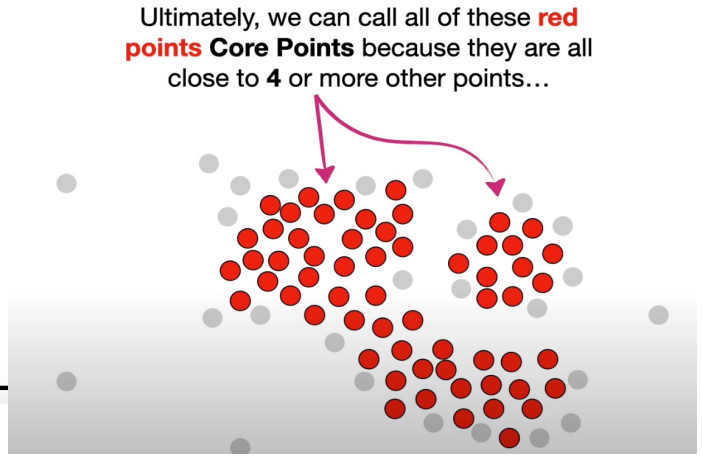
Outlier Detection – DBSCAN

Algorithm 2: DBSCAN

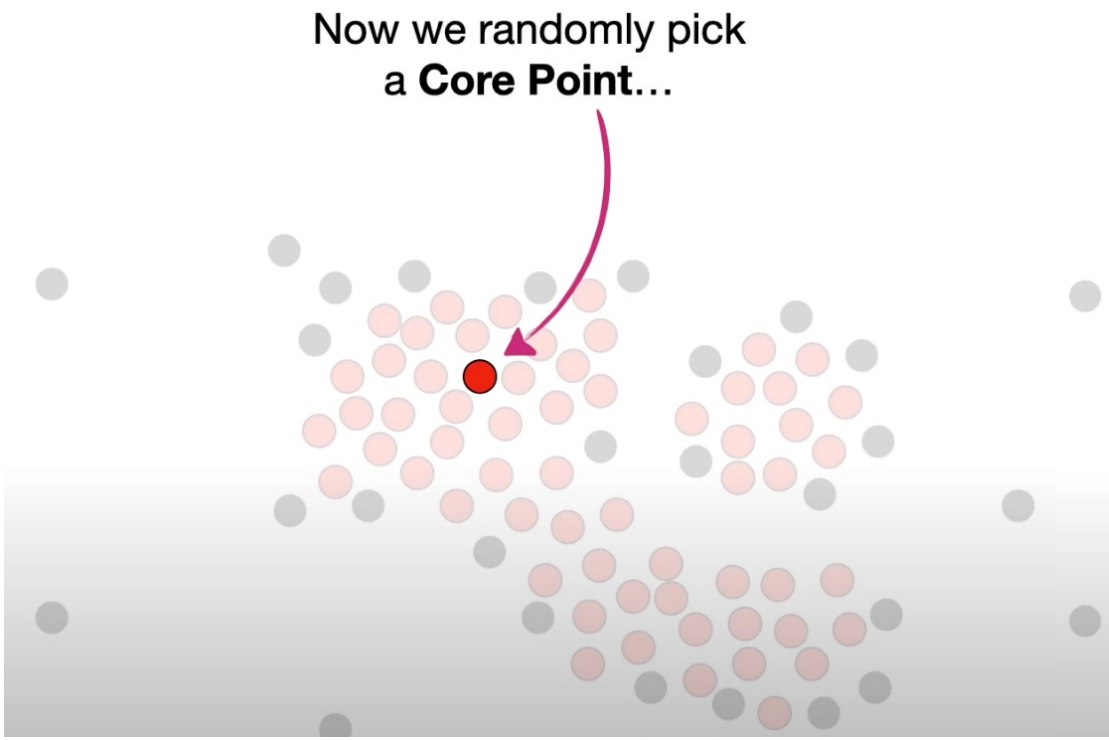
Data: X, *eps*, *min_pts*
initialse *labels* list as zeros, *count* list, *core* list;
Find neighbours for each point, Find core points;
class $= 1$;
for *each core point p* **do**
    add neighbours(p) to queue;
    while *queue not empty* **do**
        neighbours = next(queue);
        for *q in neighbours* **do**
            set label(q = *class*;
            if *label(q) is 'core'* **then**
                add neighbours(q) to queue
            end
        end
    end
    class = *class* + 1
end
return labels;

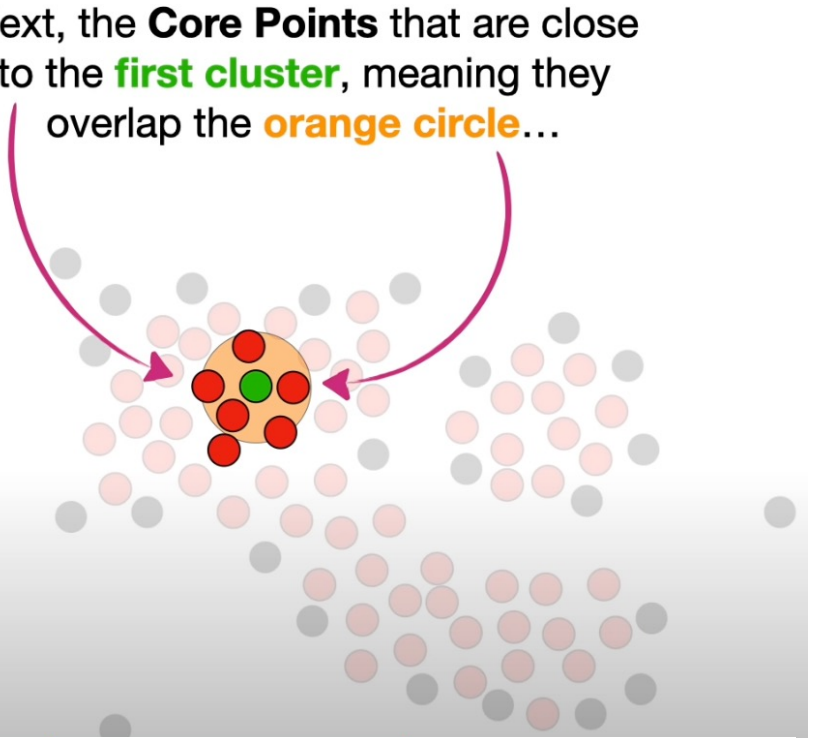
$$|N_\epsilon(\mathbf{x})| \geq minpts$$

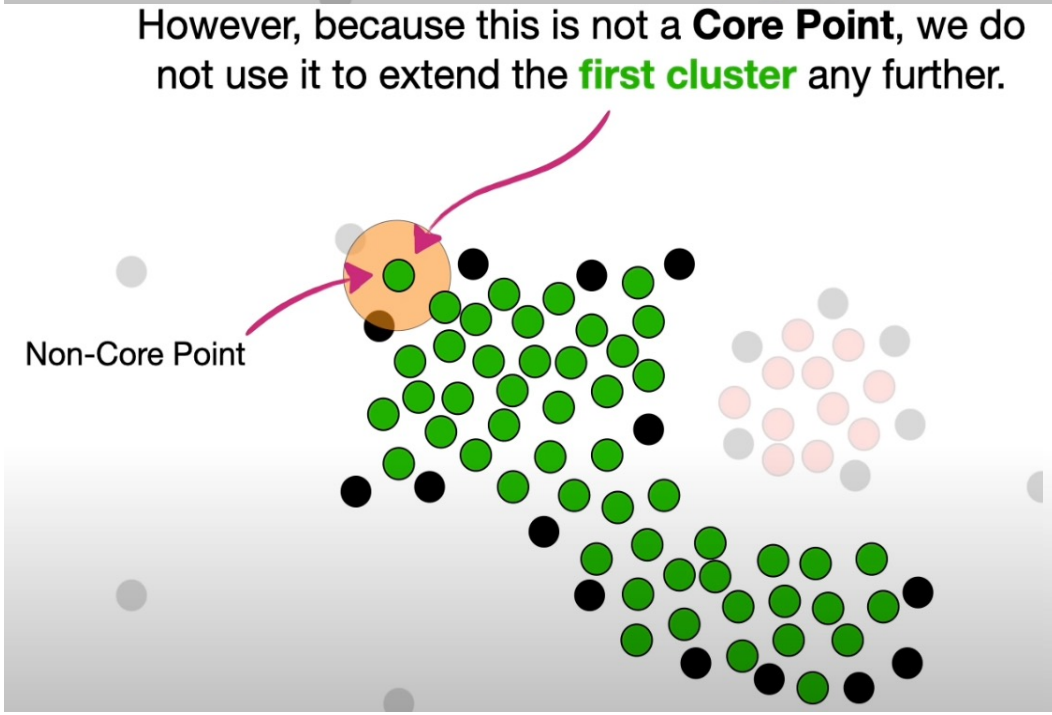


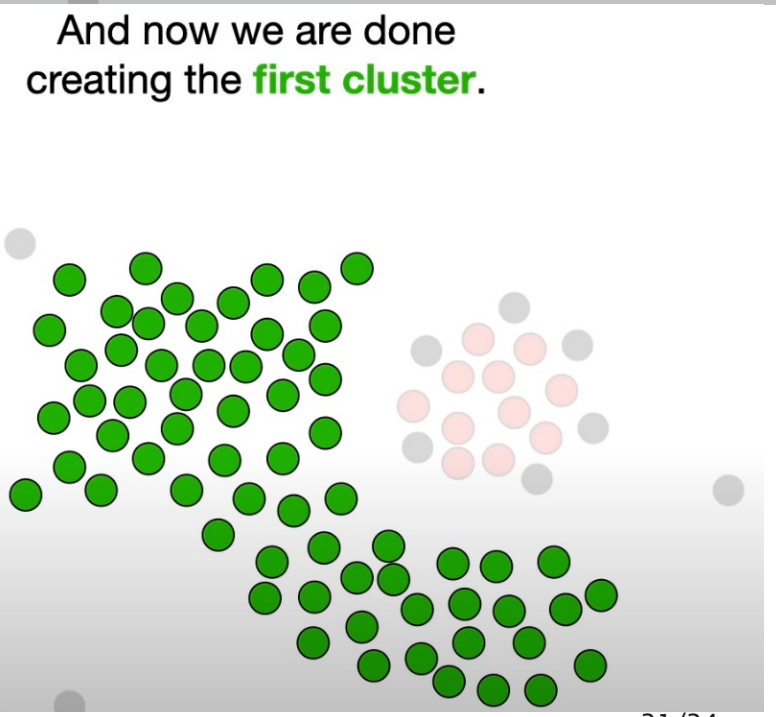


Credit: Jo Grundy



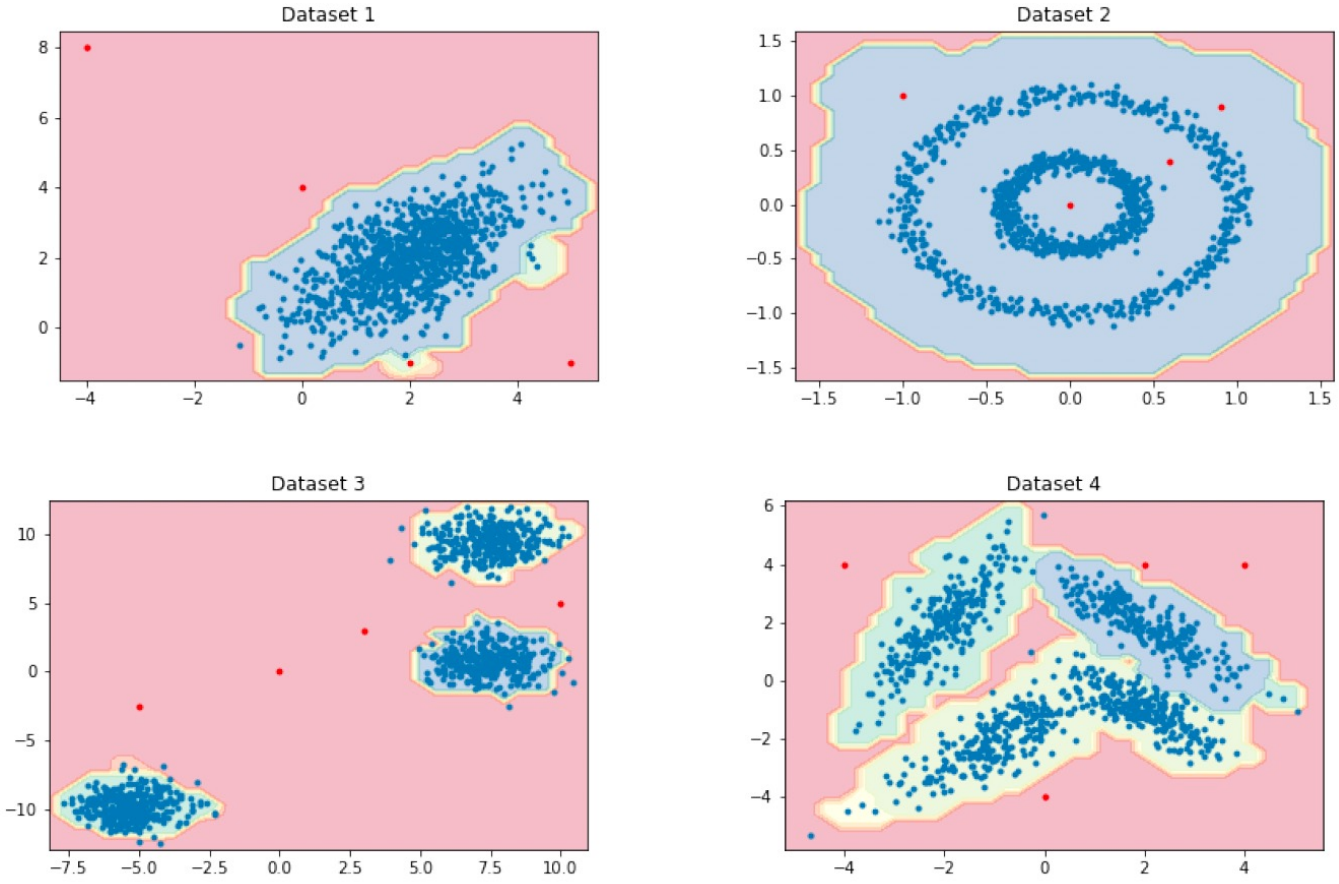






https://www.youtube.com/watch?v=RDZUdRSDOok

Outlier Detection – DBSCAN

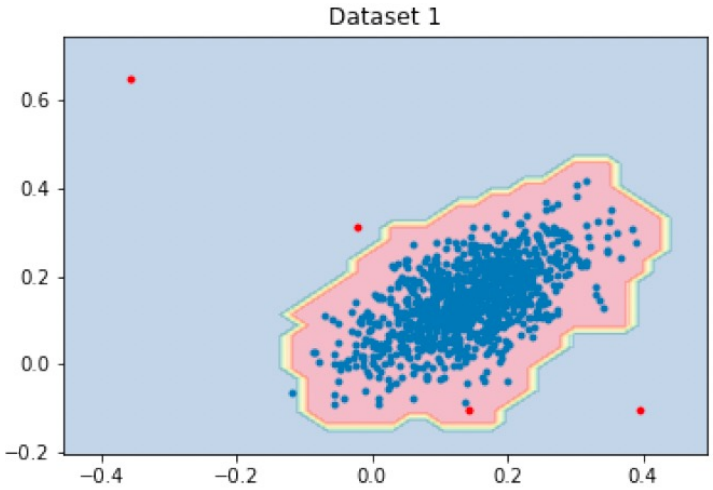


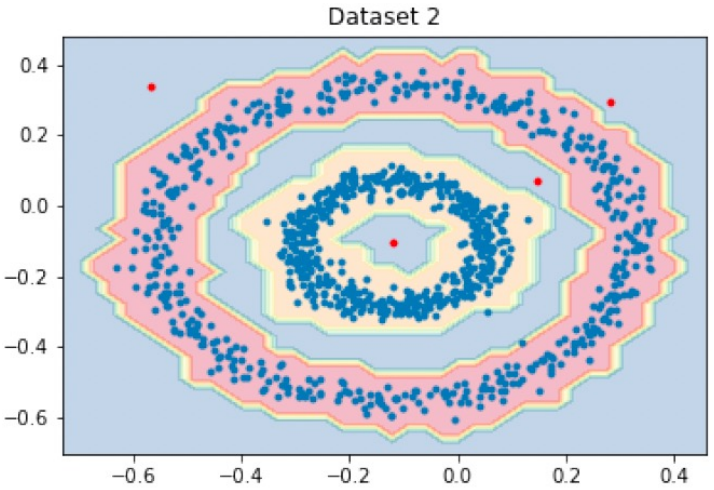
What is going on here? works well (ish) on the Gaussian datasets, but not on the oddly shaped one..

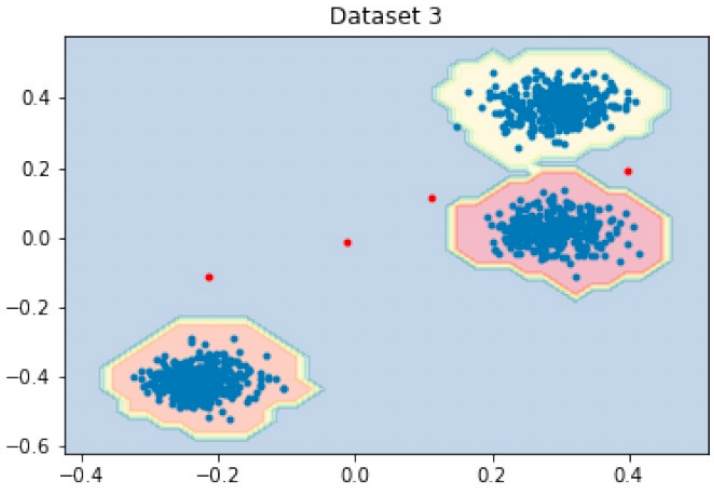
Credit: Jo Grundy

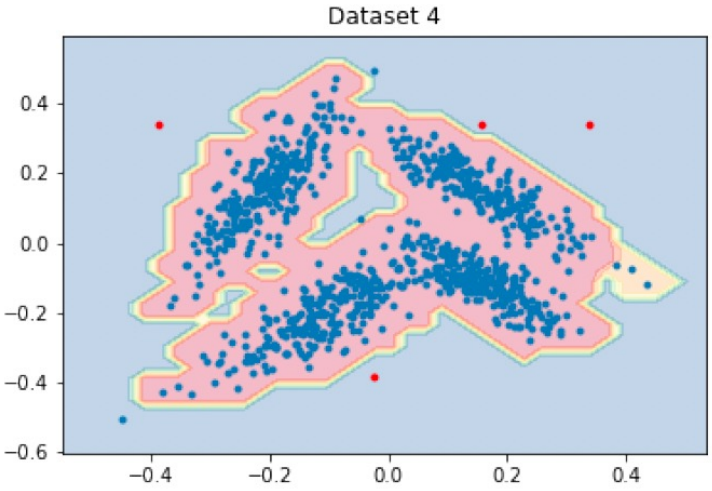
Outlier Detection – DBSCAN

Normalisation! - and adjusting *eps*









Credit: Jo Grundy

University of Southampton

Outlier Detection – Summary

Outlier detection is explored as a data mining problem:.
Extreme value statistics:

- to help tell the difference between an anomaly and an extreme member of a distribution

Gaussian Mixture Models:

- Models the system as a mixture of Gaussian distributions
- uses Expectation Maximisation to find parameters
- can be distorted by outliers

DBSCAN:

- Used for outlier detection
- Robust to outliers
- can have issues with parameters eps

Credit: Jo Grundy