Anomaly detection

→ DBSCAN

→ Elliptic Envelope

→ RANSAC

Declare no of cluster

← Kmeans → dist from each cluster

Hien → Connect 2 closest cluster

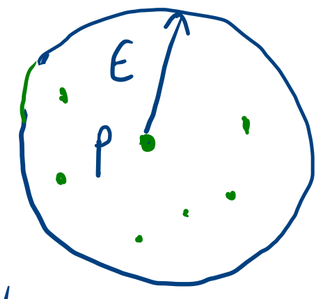
GMM → Probab from each cluster

DBSCAN

↳ If a point is surrounded by many other points → it is inside a cluster.

1) Min points → 4

2) Epsilon (Radius)

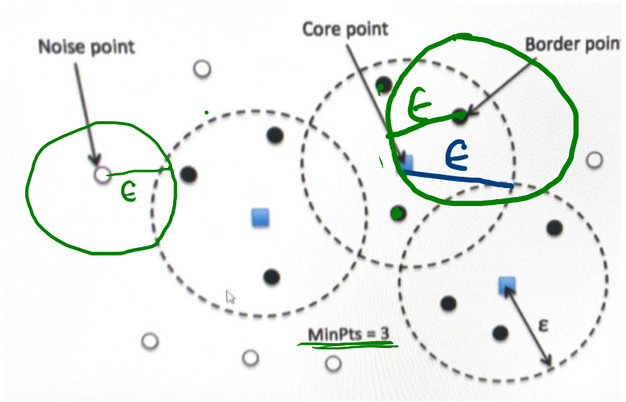


- Core point → If point p has > min point within epsilon radius
- Border point
- Noise point

Border point → It does not have minpts

↳ Atleast one core point in epsilon radius

Noise point → Neither core nor border point



Core point

Border

→ Decide ϵ & min point

→ Go to each point

↳ draw a circle

↳ Count no of point inside circle

> minpoint

< min point

Core point

Non - core

Atleast one core point

Boundary

Noise point

Algorithm

1. Start with any arbitrary point which is not visited

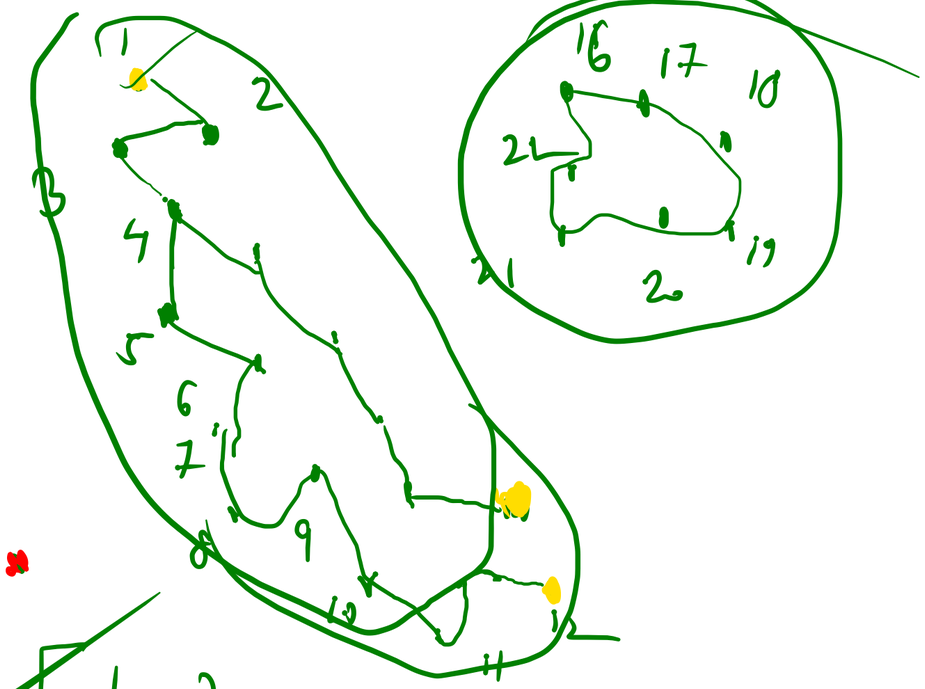
2. Check for core point, border noise

3. If point is core point then cluster formation start. So if the neighbour of this core point are also core point then these neighbour point will be added to the cluster and cluster grows

13

14

15



[1, 2, 3, 4 — — 12]

[14]

[15]

[16, 17, — — 21]

Pros

→ Don't need to select K

→ Can form non-spherical cluster

→ Not sensitive to outlier.

→ Require only 2 para : epsilon minpoint

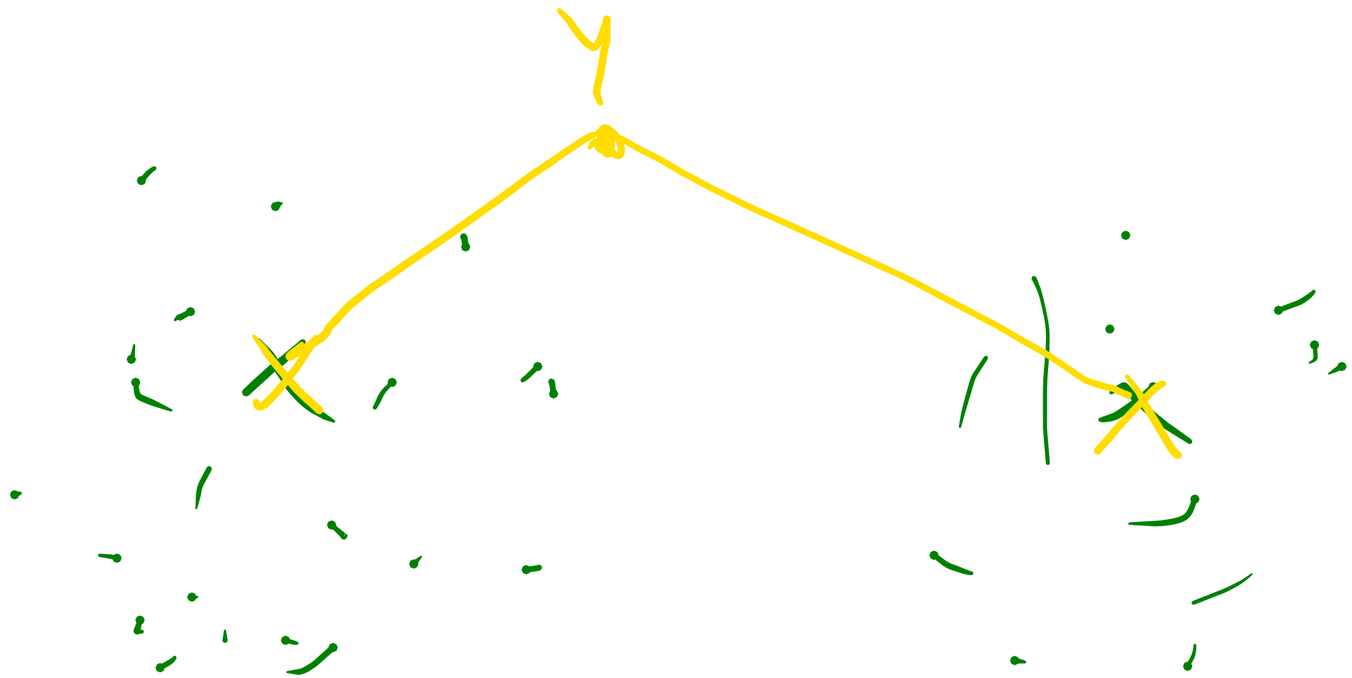
→ Can handle cluster of diff shape and size

Cons

→ Cluster is very sensitive to epsilon & min point

→ Offline algorithm

↳ No predict method


Y
X
X

DBSCAN	K-mean
1. Cluster can be any shape	Spherical
2. No need to specify no of cluster	Specify
3. Insensitive to outlier	Sensitive

:) Break until 10:35pm

Anomaly detection

→ Anomaly (outlier)
↳ Not normal behaviour

→ Why outlier exist?

1) Human error / machine error

2) Real unusual data
   - ↳ extremities
   - ↳ New

Novelty : Something which hasn't been seen before

Edtech

Car

Milage

Combustion — mileage / per litre

Electric — mileage / battery

Techniques for outlier detection

1 → IQR

Outlier $< Q_1 - 1.5\,IQR$

$> Q_3 + 1.5\,IQR$

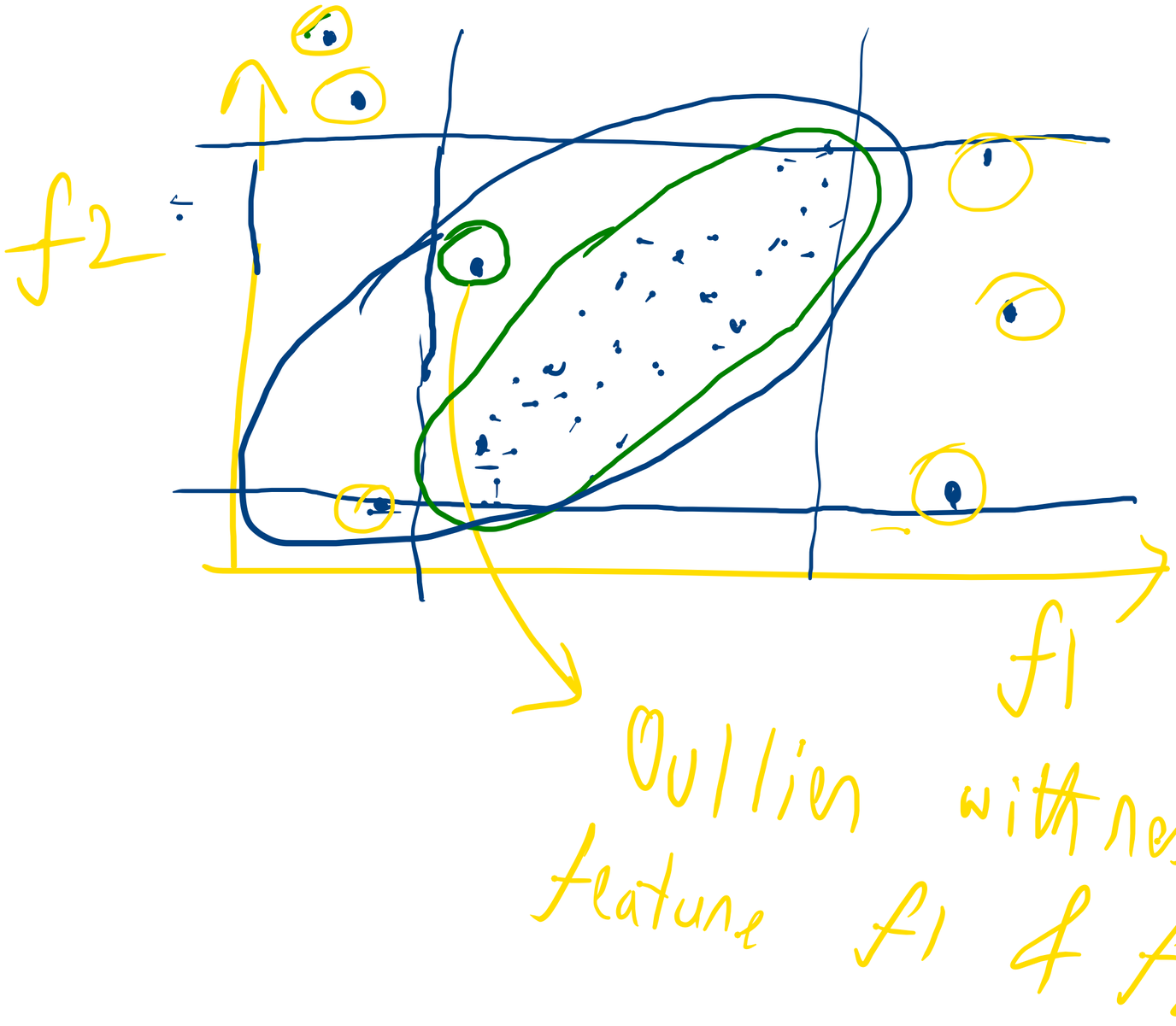
② Z-score $Z = \frac{X - \mu}{\sigma}$

Outlier $\begin{cases} Z < -3 \\ Z > 3 \end{cases}$ 3σ

③ Percentile

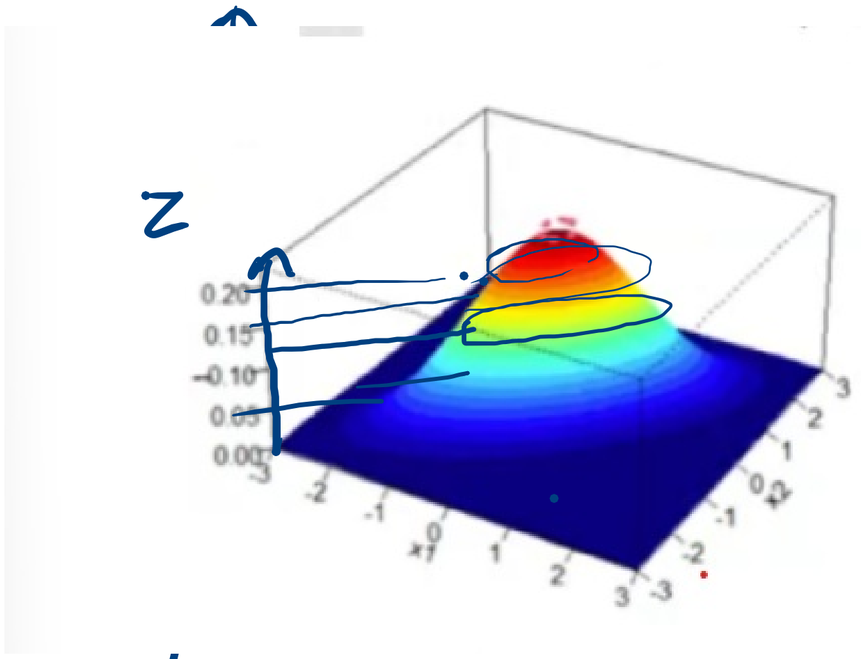
Outlier $\begin{cases} < 5\% & 2\% \\ > 95\% & 98\% \end{cases}$

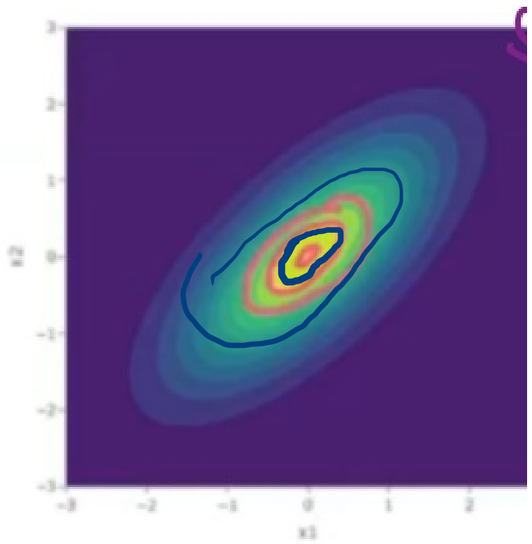
10%
f1
f2
f3
f4
f5
2%
2%
2%
2%
2%



Outlier with respect to both feature f1 & f2.

→ Circular, elliptical





X X Y

threshold

2%.