Seg 8

Outlier

- A outlier is a data point that significantly differ from others observation in a dataset.

identify and dealing with outliers is important in various field including statistics, ML, data analysis.

— যেমন CGPA (2.4) আছে, result 1.5 আসছে এটা outliers

How Occour in outliers:

reason.

- **Measurement Errors:** mistakes in record data on entry data

  → 1000 Centen হবে 100 বসে গেল

DATA entry errors: Human or system errors while inputting data.

— misplace decimal — numerical data.

Sampaling error: Data collected from a non respective sample

— টিচার donken না দিয়ে আমাদের [illegible].

Data processing issues: Errors in data transmission, cleaning

— Duplicate records, mismatched data.

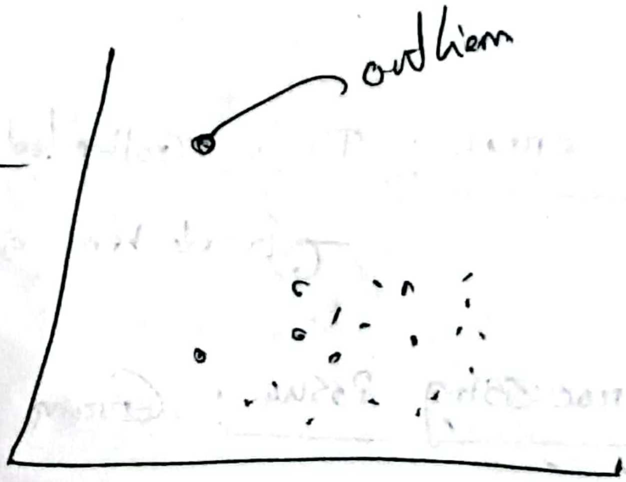
Why outlier important

- Affetive model performance, ALgorithm like, Linear Regression, k-mean, NN are especially sensitive to outliers
- Outliers heavily influence the mean, Standard divison summaries.
- Outliers may reveal unexpected pattern
- proper handaling of outliers ensure cleaner data

Types of outliers

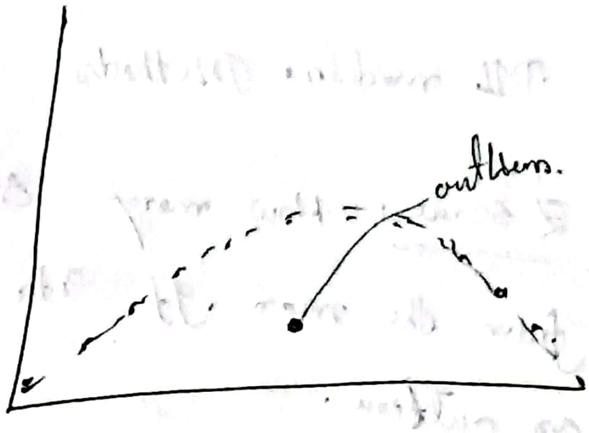
Glable outliers : It is a data points that are significantly different from all other dataset points in.
outliers easily identify.

(অনেক জনের থাকলে live
বাকির income 7000 একজনের
700.



Contextual / Conditional outliers : It is data points that are considered only in specific contexts on conditions.

- Require additional context to identify
- Often linked to time, Loc, other conditions.

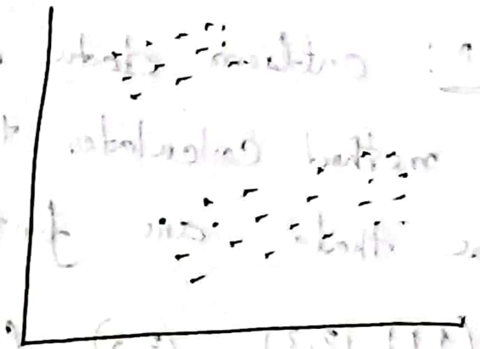


Collective outliers : It is a groups of data points that together differ from the overall pattern, even individual points might not be outliers.

- Indicate different behaviour on pattern

zone dataset in different zone

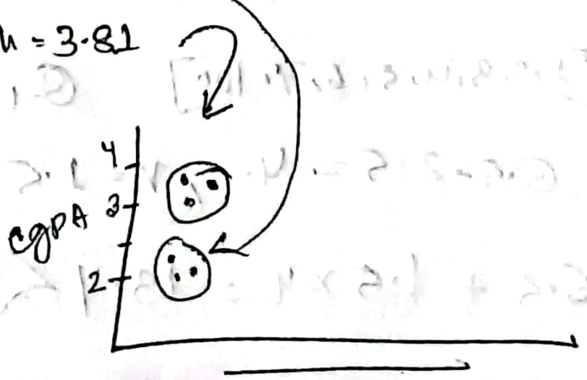
~~Cgpa 3rd Semester 2.86~~



Example CGPA

1st = 2.2   2nd = 2.3   3rd = 2.2

4th = 3.57   5th = 3.83   6th = 3.81



Outliers Detection

Outliers Detection ensured the Quality and accuracy of ML models. Methods.

Z score: How many standard division a data points is from the mean if its greater than threshold (3, -3) considered as outliers.

Ex → (80, 85, 88, 90, 92, 100, 250) mean = 96.42.
Standard division = 59.79.

$$Z = \frac{250 - 96.42}{58.79} = 2.61 < 3, \text{ so } 250 \text{ is outlier.}$$

KNN: outliers data points that have few neighbors: the method calculates the distance from each point, identify those that are farther away from other.

Ex: (1,1) (2,2), (3,3) (100,100) ← (100,100) is outlier.

IQR: InterQuartile Range

Calculate the range 1st quartile (Q_1) and 3rd quartile (Q_3). Ex → [1, 2, 3, 4, 5, 6, 7, 100] $Q_1 = 2.5$, $Q_3 = 6.5$.

IQR = 6.5 - 2.5 = 4. IV = 1.5

upper bound = 6.5 + 1.5 × 4 = 13.4 < 100 so 100 is outlier

Local outlier factor (LOF): Measure density of data points.

- যদি কোনোটির -এর Lower density হয় then outliers.

(1,1), (2,2), (3,3) (50,50) ← outlier

Clustering → DBSCAN: It Groups point that are close to each other and identifies outliers as point don't belong to any other cluster

(1,1) (2,2) (10,60) (100,100) → high density.

↳ outliers bcz অন্যদের সাথে fit না হয়

K means Cluster: point that are far from center of their cluster considered outliers.

Outliers Detection

- Supervised
  - Z-score
  - modified Z score
  - inter Quartile range.
  - SVM
  - KNN.
- Unsupervised
  - Elliptic Envelop
  - Isolation forest.
  - K - mean.
  - DBSCAN.
  - Local outlier factor