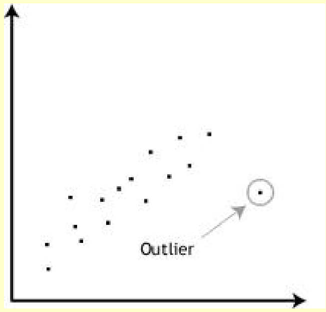
Webinar Series, 7 Nopember 2020
PROGRAM PASCASARJANA TERAPAN
POLITEKNIK ELEKTRONIKA NEGERI SURABAYA

Workshop & Tutorial
Data Mining with Python



Outlier Detection

Ali Ridho Barakbah

Knowledge Engineering Laboratory
Department of Information and Computer Engineering
Politeknik Elektronika Negeri Surabaya

What is Outlier?

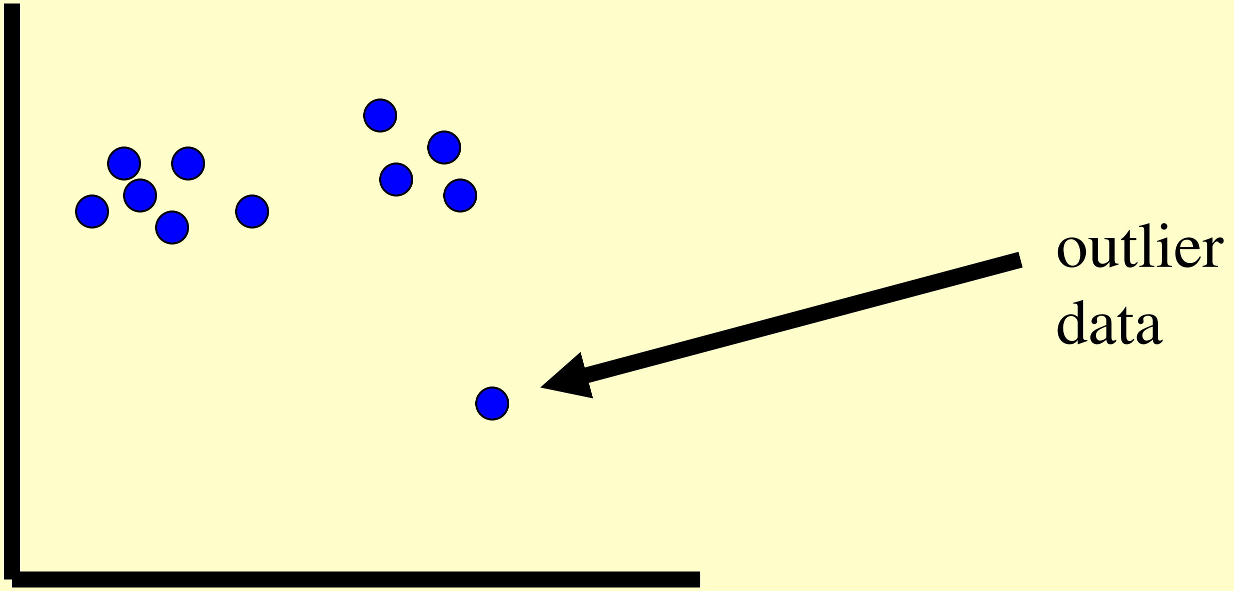
- Definition of Hawkins [Hawkins 1980]:

  “An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism”

- Statistics-based intuition

– Normal data objects follow a “generating mechanism”, e.g. some given statistical process

– Abnormal objects deviate from this generating mechanism


outlier
data

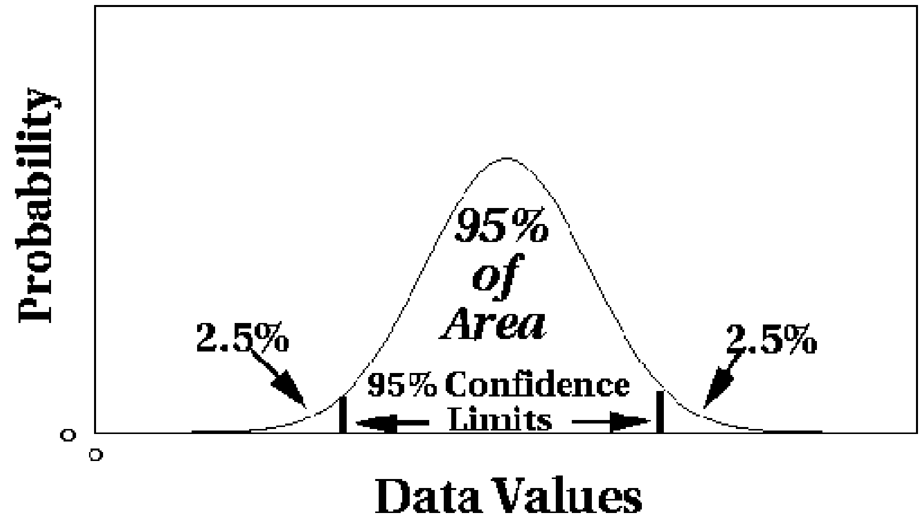
Applications

- Sample applications of outlier detection
  - Fraud detection
    - Purchasing behavior of a credit card owner usually changes when the card is stolen
    - Abnormal buying patterns can characterize credit card abuse
  - Medicine
    - Unusual symptoms or test results may indicate potential health problems of a patient
    - Whether a particular test result is abnormal may depend on other characteristics of the patients (e.g. gender, age, …)
  - Public health
    - The occurrence of a particular disease, e.g. tetanus, scattered across various hospitals of a city indicate problems with the corresponding vaccination program in that city
    - Whether an occurrence is abnormal depends on different aspects like frequency, spatial correlation, etc.

Applications

- Sample applications of outlier detection (cont.)
  - Sports statistics
    - In many sports, various parameters are recorded for players in order to evaluate the players' performances
    - Outstanding (in a positive as well as a negative sense) players may be identified as having abnormal parameter values
    - Sometimes, players show abnormal values only on a subset or a special combination of the recorded parameters
  - Detecting measurement errors
    - Data derived from sensors (e.g. in a given scientific experiment) may contain measurement errors
    - Abnormal values could provide an indication of a measurement error
    - Removing such errors can be important in other data mining and data analysis tasks
    - "One person 's noise could be another person 's signal."
  - …

Outlier Discovery: Statistical Approaches



Assume a model underlying distribution that generates data set (e.g. normal distribution)

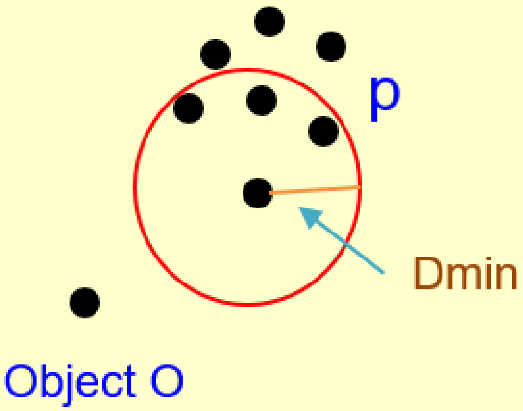
- Use discordancy tests depending on
  - data distribution
  - distribution parameter (e.g., mean, variance)
  - number of expected outliers
- Drawbacks
  - most tests are for single attribute
  - In many cases, data distribution may not be known

Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber, and Jian Pei, University of Illinois at Urbana-Champaign & Simon Fraser University

Outlier Discovery: Distance-Based Approach

- Introduced to counter the main limitations imposed by statistical methods
  - We need multi-dimensional analysis without knowing data distribution
- Distance-based outlier:

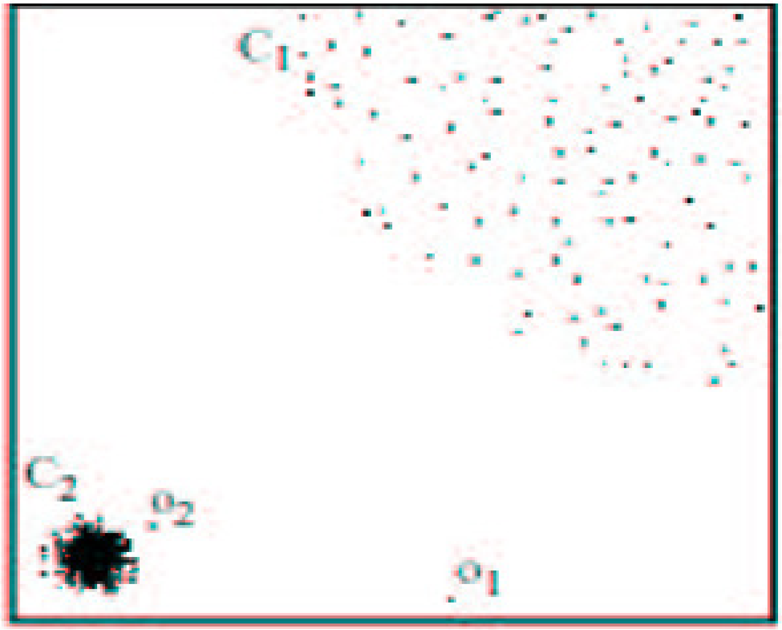
  A DB(p, Dmin)-outlier is an object O in a dataset T such that at least a fraction p of the objects in T lies at a distance greater than Dmin from O



Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber, and Jian Pei, University of Illinois at Urbana-Champaign & Simon Fraser University

Density-Based Local Outlier Detection

- M. M. Breunig, H.-P. Kriegel, R. Ng, J. Sander. LOF: Identifying Density-Based Local Outliers. SIGMOD 2000.
- Distance-based outlier detection is based on global distance distribution
- It encounters difficulties to identify outliers if data is not uniformly distributed
- Ex. C_1 contains 400 loosely distributed points, C_2 has 100 tightly condensed points, 2 outlier points o_1, o_2
- Distance-based method cannot identify o_2 as an outlier



- Need the concept of local outlier
- Local outlier factor (LOF)
  - Assume outlier is not crisp
  - Each point has a LOF

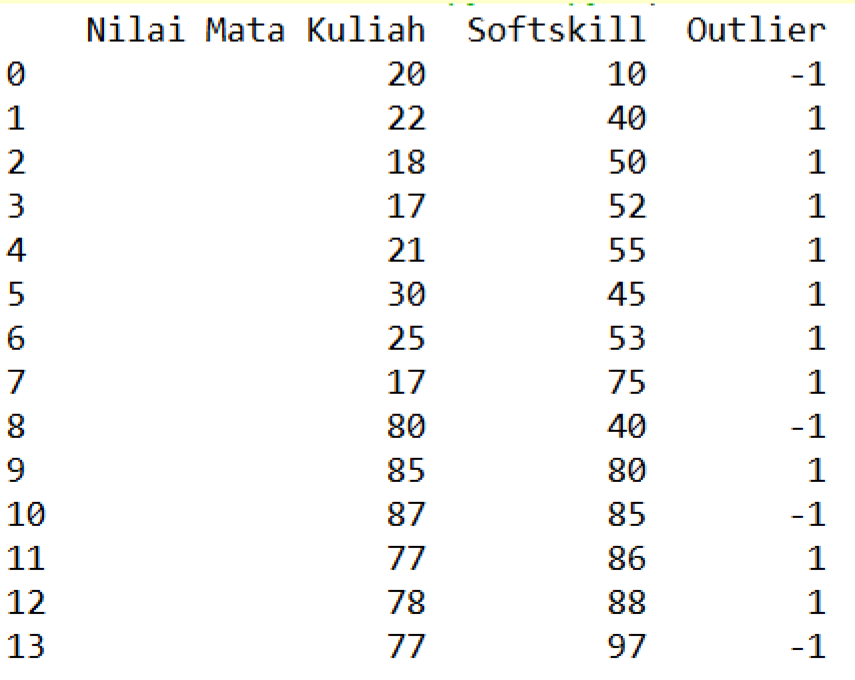
Data Mining: Concepts and Techniques, Jiawei Han, Micheline Kamber, and Jian Pei, University of Illinois at Urbana-Champaign & Simon Fraser University

Contoh Studi Kasus

Manakah dari data berikut yang termasuk outlier?

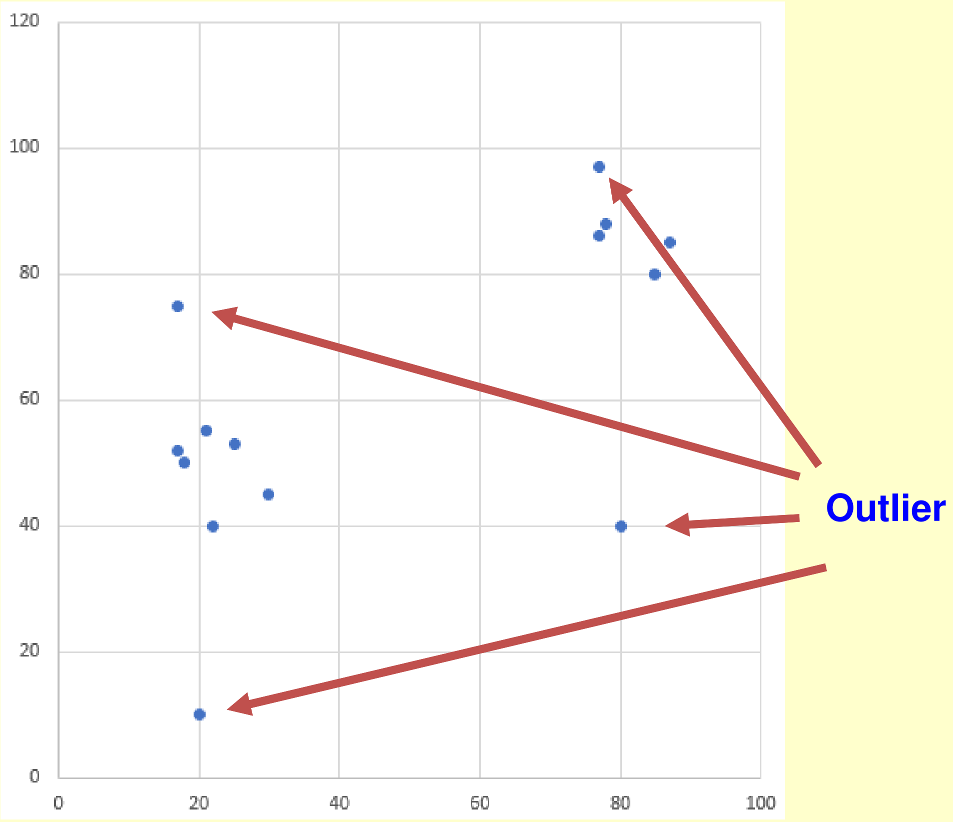
	Nilai Mata Kuliah	Softskill
data1	20	10
data2	22	40
data3	18	50
data4	17	52
data5	21	55
data6	30	45
data7	25	53
data8	17	75
data9	80	40
data10	85	80
data11	87	85
data12	77	86
data13	78	88
data14	77	97

```
import pandas as pd
import numpy as np
from sklearn.ensemble import IsolationForest

data={'Nilai': [20,22,18,17,21,30,25,17,80,85,87,77,78,77],
        'Keaktifan': [10,40,50,52,55,45,53,75,40,80,85,86,88,97]}

df=pd.DataFrame(data, columns=['Nilai','Keaktifan'])

clf = IsolationForest(contamination=0.3)
pred = clf.fit_predict(df)

df['Outlier']=pred.reshape(-1,1)

print(df)
```



```
    Nilai Mata Kuliah  Softskill  Outlier
0                  20         10       -1
1                  22         40        1
2                  18         50        1
3                  17         52        1
4                  21         55        1
5                  30         45        1
6                  25         53        1
7                  17         75        1
8                  80         40       -1
9                  85         80        1
10                 87         85       -1
11                 77         86        1
12                 78         88        1
13                 77         97       -1
```

	Nilai Mata Kuliah	Softskill
data1	20	10
data2	22	40
data3	18	50
data4	17	52
data5	21	55
data6	30	45
data7	25	53
data8	17	75
data9	80	40
data10	85	80
data11	87	85
data12	77	86
data13	78	88
data14	77	97


Outlier