Data Preprocessing

Outline

- Types of data
- Data Preprocessing Techniques
- Missing data
- Noisy data
- Outliers
- Duplicates

Part I

What is Data?

- Collection of records and their attributes
- An attribute is a characteristic of an object
- A collection of attributes describe an object

Attributes

Objects

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Types of Data

- Record Data
  - Transactional Data
- Temporal Data
  - Time Series Data
  - Sequence Data
- Spatial & Spatial-Temporal Data
  - Spatial Data
  - Spatial-Temporal Data
- Graph Data
  - Transactional Data
- UnStructured Data
  - Twitter Status Message
  - Review, news article
- Semi-Structured Data
  - Paper Publications Data
  - XML format

Record Data

- Transaction Data

TID	*Items*
1	Bread, Coke, Milk
2	Beer, Bread
3	Beer, Coke, Diaper, Milk
4	Beer, Bread, Diaper, Milk
5	Coke, Diaper, Milk

Market-Basket Dataset

Data Matrix Example for Documents

- Each document becomes a `term' vector,
  - each term is a component (attribute) of the vector,
  - the value of each component is the number of times the corresponding term occurs in the document.

	team	coach	play	ball	score	game	win	lost	timeout	season
Document 1	3	0	5	0	2	6	0	2	0	2
Document 2	0	7	0	2	1	0	0	3	0	0
Document 3	0	1	0	0	1	2	2	0	3	0

Temporal Data

- Time Series Data

Yahoo Finance Website

Biological Sequence Data

Are Rhesus Monkeys or Gibbons More Closely Related to Humans? DNA and polypeptide sequences from closely related species are more similar to each other than are sequences from more distantly related species. In this exercise, you will look at amino acid sequence data for the β polypeptide chain of hemoglobin, often called β-globin. You will then interpret the data to hypothesize whether the monkey or the gibbon is more closely related to humans.

How Such Experiments Are Done Researchers can isolate the polypeptide of interest from an organism and then determine the amino acid sequence. More frequently, the DNA of the relevant gene is sequenced, and the amino acid sequence of the polypeptide is deduced from the DNA sequence of its gene.

Data from the Experiments In the data below, the letters give the sequence of the 146 amino acids in β-globin from humans, rhesus monkeys, and gibbons. Because a complete sequence would not fit on one line here, the sequences are broken into three segments. The sequences for the three different species are aligned so that you can compare them easily. For example, you can see that for all three species, the first amino acid is V (valine) and the 146th amino acid is H (histidine).

Interpret the Data

1. Scan the monkey and gibbon sequences, letter by letter, circling any amino acids that do not match the human sequence. (a) How many amino acids differ between the monkey and the human sequences? (b) Between the gibbon and human?
2. For each nonhuman species, what percent of its amino acids are identical to the human sequence of β-globin?
3. Based on these data alone, state a hypothesis for which of these two species is more closely related to humans. What is your reasoning?

Species	Alignment of Amino Acid Sequences of β-globin
Human	1 VHLTPEEKSA VTALWGKVNV DEVGGEALGR LLVVYPWTQR FFESFGDLST
Monkey	1 VHLTPEEKNA VTTLWGKVNV DEVGGEALGR LLLVYPWTQR FFESFGDLSS
Gibbon	1 VHLTPEEKSA VTALWGKVNV DEVGGEALGR LLVVYPWTQR FFESFGDLST
Human	51 PDAVMGNPKV KAHGKKVLGA FSDGLAHLDN LKGTFATLSE LHCDKLHVDP
Monkey	51 PDAVMGNPKV KAHGKKVLGA FSDGLNHLDN LKGTFAQLSE LHCDKLHVDP
Gibbon	51 PDAVMGNPKV KAHGKKVLGA FSDGLAHLDN LKGTFAQLSE LHCDKLHVDP
Human	101 ENFRLLGNVL VCVLAHHFGK EFTPPVQAAY QKVVAGVANA LAHKYH
Monkey	101 ENFKLLGNVL VCVLAHHFGK EFTPQVQAAY QKVVAGVANA LAHKYH
Gibbon	101 ENFRLLGNVL VCVLAHHFGK EFTPQVQAAY QKVVAGVANA LAHKYH

4. What other evidence could you use to suppor your hypothesis?

MB A version of this Scientific Skills Exercise can be assigned in MasteringBiology.

Data from Human: http://www.ncbi.nlm.nih.gov/protein/AAA21113.1; rhesus monkey: http://www.ncbi.nlm.nih.gov/protein/122634; gibbon: http://www.ncbi.nlm.nih.gov/protein/122616

Interval Data

EL= { (A, 1, 5),(C, 3, 12), (B, 4, 9), (D, 9, 15) }

measured along a scale, in which each point is placed at equal distance from one another. Interval data always appears in the form of numbers or numerical values where the distance between the two points is standardized and equal

Spatial & Spatial-Temporal Data

- Spatial Data

data is collected across both space and time

Spatial & Spatial-Temporal Data

- Spatial Data

Average Monthly Temperature of land and ocean

Spatial & Spatial-Temporal Data

- Spatial Data

Dengue Disease Dataset (Singapore)

Spatial & Spatial-Temporal Data

- Trajectory Data: Set of Harricans

http://csc.noaa.gov/hurricanes

Spatial & Spatial-Temporal Data

- Trajectory Data: (of 87 users obtained using RFID)

Vast 2008 Challenge – RFID Dataset

User Movement Data

- *Trajectory*
  - *Movement trail of a user*
  - *Sampling Points: <latitude, longitude, time>*

Semistructured Data: Example

Semi-structured data does not follow the format of a tabular data model

Graph Data

Unstructured Data

Jason Y.
Reviews Written: 69

09/22/2006

Excellent Vietnamese restaurant. Some of my favorites are the sour salmon soup, vietnamese steak, summer rolls, and orange chicken. The tofu w/ginger and scallion special is good. I prefer the pho at pho-specific places. they don't give you basil or chilies with the pho here.

Bookmark Send to a Friend

Aaron E.
Reviews Written: 75

01/18/2006

They have great fried spring rolls and fresh garden rolls. However, for $3.50 2 smallish rolls doesn't seem that great. They make up for this with a large bowl of Pho soup for around $6.00. This place is great for food just before you see a movie at the Uptown theater. However, leave plenty of time. When the place is busy, the waitstaff or kitchen starts to suck and get real slow. You will miss your movie. Give yourself 1.5 hours before a movie start time. This PHO place is not as ghetto as other PHO places (like PHO 75).

Bookmark Send to a Friend

Sean G.
Reviews Written: 126

03/06/2006

The pho is good -- esp for DC -- but not great compared to some of the Pho joints in Seattle who can drop a bowl of broth and noodles that will blow your mind for under $4 bucks. Travelling... the downside is you learn what youre missing..... That said, this is a decent spot for a late lunch over a hot bowl of the the good stuff.

Bookmark Send to a Friend

Nils J.
Reviews Written: 76

04/14/2006

Good pho, reasonable prices, and great for eating dinner before you go to the Uptown for a flick. The wait staff can be very unresponsive, though, which is annoying. Unfortunately, they don't give you basil and cilantro along with the sprouts to dump into your pho.

Bookmark Send to a Friend

Nathalie L.
Reviews Written: 208

07/08/2006

I was missing home one day and decided to pop in for some comfort food and Nam Viet was exactly what I needed. The waitresses and waiters speaking Vietnamese and the aroma of the pho really lifted my spirits that day.

Bookmark Send to a Friend

What people do with the time series data?

Clustering

Classification

Motif Discovery

Rule Discovery

Query by Content

Visualization

Novelty Detection

Motif Association

What people do with the trajectory data?

Clustering

Frequent Travel Patterns

Motif Discovery

Prediction

Visualization

Classification

Activity 1

- Find top 3 recent research activities around the world that are analyzing data. You need to write short summary for each research activities. First three line must follow following format:
  - **Line 1**: Problem they are trying to sole along with dataset they are using
  - **Line 2**: How they are solving the problem
  - **Line 3**: Justify yourself why you rate this work as a top 5 activities

Attribute Values

- Attribute values are **numbers** or **symbols** assigned to an attribute

- Distinction between attributes and attribute values
  - Same attribute can be mapped to different attribute values
    - Example: height can be measured in feet or meters

  - Different attributes can be mapped to the same set of values
    - Example: Attribute values for ID and age are integers
    - But properties of attribute values can be different
      - ID has no limit but age has a maximum and minimum value

Attribute Values

Data is described using attribute values

Attributes

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	95K	Yes
6	No	Married	60K	No
7	Yes	Divorced	220K	No
8	No	Single	85K	Yes
9	No	Married	75K	No
10	No	Single	90K	Yes

Types of Attributes

- There are different types of attributes
  - Nominal
    - Examples: ID numbers, eye color, zip codes
  - Ordinal
    - Examples: rankings (e.g., taste of potato chips on a scale from 1-10), grades, height in {tall, medium, short}
  - Interval
    - Examples: calendar dates
  - Ratio
    - Examples: length, time, counts

Attribute Type		Description	Examples
Categorical (Qualitative)	Nominal	The values of a nominal attribute are just different names; i.e., nominal values provide only enough information to distinguish one object from another. $(=, \neq)$	zip codes, employee ID numbers, eye color, gender
	Ordinal	The values of an ordinal attribute provide enough information to order objects. $(<, >)$	hardness of minerals, $\{good, better, best\}$, grades, street numbers
Numeric (Quantitative)	Interval	For interval attributes, the differences between values are meaningful, i.e., a unit of measurement exists. $(+, -)$	calendar dates, temperature in Celsius or Fahrenheit
	Ratio	For ratio variables, both differences and ratios are meaningful. $(*, /)$	temperature in Kelvin, monetary quantities, counts, age, mass, length, electrical current

Discrete and Continuous Attributes

- Discrete Attribute
  - Has only a finite or countable infinite set of values
  - Examples: zip codes, counts, or the set of words in a collection of documents
  - Often represented as integer variables.

- Continuous Attribute
  - Has real numbers as attribute values
  - Examples: temperature, height, or weight.
  - Practically, real values can only be measured and represented using a finite number of digits.

Part II

Data Quality

- Examples of data quality problems:
  - Noise and outliers
  - Missing values
  - Duplicate data

A mistake or a millionaire?

Missing values

Inconsistent duplicate entries

Tid	Refund	Marital Status	Taxable Income	Cheat
1	Yes	Single	125K	No
2	No	Married	100K	No
3	No	Single	70K	No
4	Yes	Married	120K	No
5	No	Divorced	10000K	Yes
6	No	NULL	60K	No
7	Yes	Divorced	220K	NULL
8	No	Single	85K	Yes
9	No	Married	90K	No
9	No	Single	90K	No

Data Cleaning

- Data cleaning tasks
  - Fill in missing values
  - Identify outliers and smooth out noisy data
  - Correct inconsistent data
  - Resolve redundancy caused by data integration

1. Noisy Data

- Noise: random error or variance in a measured variable.
- Incorrect attribute values may due to
  - faulty data collection instruments
  - data entry problems
  - data transmission problems
  - etc
- Other data problems which requires data cleaning
  - duplicate records, incomplete data, inconsistent data

Data Quality: Noise

- Noise refers to modification of original values
  - Examples: distortion of a person's voice when talking on

Two Sine Waves

Two Sine Waves + Noise

Frequency Plot (FFT)

Data Quality: Handle Noise(Binning)

- Binning
  - sort data and partition into (equi-depth) bins
  - smooth by bin means, bin median, bin boundaries, etc.
- Regression
  - smooth by fitting a regression function
- Clustering
  - detect and remove outliers
- Combined computer and human inspection
  - detect suspicious values automatically and check by human

Data Quality: Handle Noise(Binning)

- Equal-width binning
  - Divides the range into *N* intervals of *equal size*
  - Width of intervals:
  - Simple
  - Outliers may dominate result
- Equal-depth binning
  - Divides the range into *N* intervals,

    each containing approximately *same number* of records
  - Skewed data is also handled well

Simple Methods: Binning

Example: customer ages

Binning Methods for Data Smoothing

- Sorted data for price (in dollars):

4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

- Partition into (equi-depth) bins:
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
- Smoothing by bin means:
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
- Smoothing by bin boundaries: closest boundary value
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

Data Quality: Handle Noise(*Regression*)

- Replace noisy or missing values by predicted values
- Requires model of attribute dependencies (maybe wrong!)
- Can be used for data smoothing or for handling missing data

Data Quality: Handle Noise(Binning)

Example: Sorted price values 4, 8, 9, 15, 21, 21, 24, 25, 26, 28, 29, 34

* Partition into three (equi-depth) bins
  - Bin 1: 4, 8, 9, 15
  - Bin 2: 21, 21, 24, 25
  - Bin 3: 26, 28, 29, 34
* Smoothing by bin means
  - Bin 1: 9, 9, 9, 9
  - Bin 2: 23, 23, 23, 23
  - Bin 3: 29, 29, 29, 29
* Smoothing by bin boundaries
  - Bin 1: 4, 4, 4, 15
  - Bin 2: 21, 21, 25, 25
  - Bin 3: 26, 26, 26, 34

2. Missing Data

- Data is not always available
  - E.g., many tuples have no recorded values for several attributes, such as customer income in sales data
- Missing data may be due to
  - equipment malfunction
  - inconsistent with other recorded data and thus deleted
  - data not entered due to misunderstanding
  - certain data may not be considered important at the time of entry
  - not register history or changes of the data

How to Handle Missing Data?

- Ignore the tuple
- Fill in missing values manually: tedious + infeasible?
- Fill in it automatically with
  - a global constant : e.g., "unknown", a new class?!
  - the attribute mean
  - the most probable value: inference-based such as Bayesian formula, decision tree, or EM algorithm

Data Quality: Missing Values

- Reasons for missing values
  - Information is not collected
    (e.g., people decline to give their age and weight)
  - Attributes may not be applicable to all cases
    (e.g., annual income is not applicable to children)

- Handling missing values
  - Eliminate Data Objects
  - Estimate Missing Values
  - Ignore the Missing Value During Analysis
  - Replace with all possible values (weighted by their probabilities)

https://www.datacamp.com/tutorial/techniques-to-handle-missing-data-values

Methods for handling missing values
Deletion of data instance
Imputation of missing values
Model-based procedures
Machine-learning methods for handling missing data
Statistical imputation methods
Imputation based on machine-learning
Maximum likelihood with expectation-maximization algorithm
• Mean imputation
• Regression imputation
• Multiple imputation
• Hot Deck imputation
• K-nn imputation
• SOM imputation
• ANN imputation
• MLP imputation
• Multi-Task networks
• Gaussian mixture models
• Ensemble-methods
• Fuzzy methods
• SVM methods
• Decision trees and inductive algorithms

Mean Imputation

	col1	col2	col3	col4	col5
0	2	5.0	3.0	6	NaN
1	9	NaN	9.0	0	7.0
2	19	17.0	NaN	9	NaN

mean()

	col1	col2	col3	col4	col5
0	2.0	5.0	3.0	6.0	7.0
1	9.0	11.0	9.0	0.0	7.0
2	19.0	17.0	6.0	9.0	7.0

Mean, median, mode imputation

Regression Imputation

If we know there is a correlation between the missing value and other variables, we can often get better guesses by regressing the missing variable on other variables.

	total_bill	size	tip
total_bill	1	0.6	0.67
size	0.6	1	0.5
tip	0.67	0.5	1

we can use a simple linear model regressing total_bill on tip to fill the missing values in total_bill

Multiple Imputation

1. **Imputation step**. M imputations (completed datasets) are generated under some chosen imputation model.
2. **Completed-data analysis (estimation) step**. The desired analysis is performed separately on each imputation (m = 1, … , M). This is called completed-data analysis and is the primary analysis to be performed once missing data have been imputed.
3. **Pooling step**. The results obtained from M completed-data analyses are combined into a single multiple-imputation result.

Hot deck imputation

each missing value is replaced with an observed response from a "similar" unit

PhysActive	Weight
TRUE	51
TRUE	63
FALSE	98
TRUE	NA
FALSE	81
FALSE	88

PhysActive	Weight
TRUE	51
TRUE	63
FALSE	98
TRUE	NA
FALSE	81
FALSE	88

K-nn imputation

use 'k' samples to estimate the value of the missing data points. Each sample's missing values are imputed using the mean value of the 'k'-neighbors found in the dataset.

Part III Outliers

- Types of Outliers
- Methods to find outliers
  - Box plots
  - IQR method
  - Z-score method
  - K-nn based method
  - LOF
  - Clustering

Data Quality: Outliers

- Outliers are data objects with characteristics that are considerably different than most of the other data objects in the data set

https://towardsdatascience.com/outlier-detection-methods-in-machine-learning-1c8b7cca6cb8

Types of Outliers

1. Type 1 – Global - “Point Anomalies,” are kind of outliers which deviate significantly from the rest of the data.
2. Type 2 – Contextual - anomalous in a specific context or condition (but not otherwise)
3. Type 3 – Collective- collection of data points is anomalous with respect to the entire data set

Outliers are different from the noise data.

- Noise is random error or variance in a measured variable
- Noise should be removed before outlier detection

An outlier may indicate an experimental error, or it may be due to variability in the measurement.

Collective anomalies

- A collection of related data instances is anomalous
- Requires a relationship among data instances
  - Sequential data
  - Spatial data
  - Graph data
- The individual instances within a collective anomaly are not anomalous by themselves

Methods to identify outliers in the data

1. Box plots

2. IQR method

3. Z-score method

4. knn

5. LOF

6. Clustering

Understanding Boxplots

```
IQR = Q3-Q1

Lower Limit = Q1 - 1.5*IQR

Upper Limit = Q3 + 1.5*IQR
```

Outliers are the data points **below and above** the **lower and upper limit**.

- **Minimum** – It is the minimum value in the dataset excluding the outliers
- **First Quartile (Q1)** – 25% of the data lies below the First (lower) Quartile.
- **Median (Q2)** – It is the mid-point of the dataset. Half of the values lie below it and half above.
- **Third Quartile (Q3)** – 75% of the data lies below the Third (Upper) Quartile.
- **Maximum** – It is the maximum value in the dataset excluding the outliers.

Quartile Computation Example

- 100,110,110,110,120,120,130,140,140,150,170,220
- **Median (Q2)** = (120+130)/2 = **125** ;
- First Quartile - take the first six values and find their median.
- 100,110,110,110,120,120,130,140,140,150,170,220
- **Q1** = (110+110)/2 = **110**
- Third Quartile, take the next six and find their median.
- 100,110,110,110,120,120,130,140,140,150,170,220
- **Q3** = (140+150)/2 = **145**
- If the total number of values is odd then we exclude the Median while calculating Q1 and Q3.
- **IQR** = Q3-Q1 = 145-110 = **35**
- **Lower Limit** = Q1-1.5*IQR = 110-1.5*35 = **57.5**
- **Upper Limit** = Q3+1.5*IQR = 145+1.5*35 = **197.5**
- **Minimum = 100**
- **Maximum = 170**
- The outliers which are outside this range are –
- **Outliers = 220**

1. Box plots

- Box plots visualize data distribution.
- plots the q1 (25th percentile), q2 (50th percentile or median) and q3 (75th percentile)
- *(q1–1.5*(q3-q1))* and *(q3+1.5*(q3-q1))*.
- Outliers, if any, are plotted as points above and below the plot.

IQR
Q1
Q3
Q1 - 1.5*IQR
Q3 + 1.5*IQR
Median
−4σ
−3σ
−2σ
−1σ
0σ
1σ
2σ
3σ
4σ
-2.698σ
-0.6745σ
0.6745σ
2.698σ
Probability Density
0.40
0.35
0.30
0.25
0.20
0.15
0.10
0.05
0.00
0.35%
24.65%
50%
24.65%
0.35%

a) If the Median is at the **center** of the Box and the **whiskers** are almost the **same on both the ends** then the data is **Normally Distributed.**

b) If the Median lies **closer to the First Quartile** and if the **whisker at the lower end is shorter** (as in the above example) then it has a **Positive Skew (Right Skew).**

c) If the Median lies **closer to the Third Quartile** and if the **whisker at the upper end is shorter** then it has a **Negative Skew (Left Skew).**

how a score compares to other scores in the same set

Quartile Uses

- Suppose Jimmy works as a software developer He earns a monthly gross wage of $6,200. He would like to find out how his wages compare with those working in the same occupation and industry.
- He obtained data on the 25th percentile, median and 75th percentile wage levels
- Jimmy's wage is above the 75th percentile is, he is among the top 25% of earners among software developers in the industry.
- The 25th percentile is the value at which 25% of data lie below that value, and 75% of data lie above that value.
- **50th Percentile** - Also known as the Median. The median cuts the data set in half. Half of the data lie below the median and half lie above the median.
- **75th Percentile** - third, or upper, quartile. - value at which 25% of the data lie above that value and 75% of the answers lie below that value.

https://www.analyticsvidhya.com/blog/2021/05/feature-engineering-how-to-detect-and-remove-outliers-with-python-code/

2. IQR method

- *Lower Bound = q1–1.5*IQR*
- *Upper Bound = q3+1.5*IQR*
- Any value below the lower bound and above the upper bound are considered to be outliers.

3. Z-score method

- Z-score is the number of standard deviations a value of a variable is away from the variable' mean.
- *Z-Score = (X-mean) / Standard deviation*

Example:
weights of newborns are normally distributed with a mean of about 7.5 pounds and a standard deviation of 0.5 pounds. If a certain newborn weights 7.7 pounds, What is the z-score?

- $z = (x - \mu) / \sigma$
- $z = (7.7 - 7.5) / 0.5$
- $z = 0.4$

- If x = 10 pounds?

Distance-based Approaches

- Data is represented as a vector of features
- Three major approaches
  - Nearest-neighbor based
  - Density based
  - Clustering based

Nearest-Neighbor Based Approach

- Approach:
  - Compute the distance between every pair of data points
  - There are various ways to define outliers:
    - Data points for which there are fewer than *p* neighboring points within a distance *D*
    - The top n data points whose distance to the kth nearest neighbor is greatest
    - The top n data points whose average distance to the k nearest neighbors is greatest

https://towardsdatascience.com/local-outlier-factor-lof-algorithm-for-outlier-identification-8efb887d9843

REACHABILITY DENSITY (RD)

- maximum of K-distance of Xj and the distance between Xi and Xj.

$$\mathrm{RD}(X_i, X_j) = \max\left(K-\text{ distance }(X_j)\,,\ \text{distance }(X_i, X_j)\right)$$

Illustration of reachability distance with K=2

if a point Xi lies within the K-neighbors of Xj, the reachability distance will be K-distance of Xj (blue line), else reachability distance will be the distance between Xi and Xj (orange line).

Density-based: LOF approach

- For each point, compute the density of its local neighborhood
- Compute local outlier factor (LOF) of a sample *p* as the average of the ratios of the density of sample *p* and the density of its nearest neighbors
- Outliers are points with largest LOF value

In the NN approach, p_2 is not considered as outlier, while LOF approach find both p_1 and p_2 as outliers

K-DISTANCE AND K-NEIGHBORS

- **K-distance** is the distance between the point, and it's K^{th} nearest neighbor.
- **K-neighbors** denoted by $N_k(A)$ includes a set of points that lie in or on the circle of radius K-distance.
- If K=2, K-neighbors of A will be C, B, Here, the value of K=2 but the
- $||N_2(A)|| = 2$.

Manhattan_Distance(A,B) =	1
Manhattan_Distance(A,C) =	2
Manhattan_Distance(A,D) =	3
Manhattan_Distance(B,C) =	1
Manhattan_Distance(B,D) =	4
Manhattan_Distance(C,D) =	3

K-neighborhood (A) = {B,C} , ||N2(A)|| =2
K-neighborhood (B) = {A,C}, ||N2(B)|| =2
K-neighborhood (C)= {B,A}, ||N2(C)|| =2
K-neighborhood (D) = {A,C}, ||N2(D)|| =2

REACHABILITY DENSITY (RD)

$$\mathrm{RD}(X_i, X_j) = \max\left(K - \text{ distance } (X_j)\,,\ \text{distance } (X_i, X_j)\right)$$

Manhattan_Distance(A,B) =	1
Manhattan_Distance(A,C) =	2
Manhattan_Distance(A,D) =	3
Manhattan_Distance(B,C) =	1
Manhattan_Distance(B,D) =	4
Manhattan_Distance(C,D) =	3

RD(A,B) = 1
RD(A,C) = 2
RD(B,A) = 2
RD(B,C) = 2
….

LOCAL REACHABILITY DENSITY (LRD)

$$LRD_k(A) = \frac{1}{\sum_{Xj \in N_k(A)} \frac{\mathrm{RD}(A,Xj)}{||N_k(A)||}}$$

$$LRD_2(A) = \frac{1}{\frac{RD(A,B)+RD(A,C)}{||N_2(A)||}} = \frac{\frac{1}{1+2}}{2} = 0.667$$

$$LRD_2(B) = \frac{1}{\frac{RD(B,A)+RD(B,C)}{||N_2(B)||}} = \frac{\frac{1}{2+2}}{2} = 0.50$$

$$LRD_2(C) = \frac{1}{\frac{RD(C,B)+RD(C,A)}{||N_2(C)||}} = \frac{\frac{1}{1+2}}{2} = 0.667$$

$$LRD_2(D) = \frac{1}{\frac{RD(D,A)+RD(D,C)}{||N_2(D)||}} = \frac{\frac{1}{3+3}}{2} = 0.337$$

LRD is inverse of the average reachability distance of A from its neighbors.

more the average reachability distance (i.e., neighbors are far from the point), less density of points are present around a particular point.

Low values of LRD implies that the closest cluster is far from the point.

LOCAL OUTLIER FACTOR (LOF)

$$LOF_k(A) = \frac{\sum_{Xj\ \in N_k(A)} LRD_k(Xj)}{||N_k(A)||} \times \frac{1}{LRD_k(A)}$$

LOF is the ratio of the average LRD of the K neighbors of A to the LRD of A

$$LOF_2(A) = \frac{LRD_2(B) + LRD_2(C)}{\|N_2(A)\|} \times \frac{1}{LRD_2(A)} = \frac{0.5 + 0.667}{2} \times \frac{1}{0.667} = 0.87$$

$$\mathrm{LOF}_2(B) = \frac{LRD_2(A) + LRD_2(C)}{\|N_2(B)\|} \times \frac{1}{LRD_2(B)} = \frac{0.667 + 0.667}{2} \times \frac{1}{0.5} = 1.334$$

$$\mathrm{LOF}_2(C) = \frac{LRD_2(B) + LRD_2(A)}{\|N_2(C)\|} \times \frac{1}{LRD_2(C)} = \frac{0.5 + 0.667}{2} \times \frac{1}{0.667} = 0.87$$

$$\mathrm{LOF}_2(D) = \frac{LRD_2(A) + LRD_2(C)}{\|N_2(D)\|} \times \frac{1}{LRD_2(D)} = \frac{0.667 + 0.667}{2} \times \frac{1}{0.337} = 2$$

Generate data with outliers

- import numpy as np
-
- np.random.seed(42)
-
- X_inliers = 0.3 * np.random.randn(100, 2)
- X_inliers = np.r_[X_inliers + 2, X_inliers - 2]
- X_outliers = np.random.uniform(low=-4, high=4, size=(20, 2))
- X = np.r_[X_inliers, X_outliers]
-
- n_outliers = len(X_outliers)
- ground_truth = np.ones(len(X), dtype=int)
- ground_truth[-n_outliers:] = -1

https://scikit-learn.org/stable/auto_examples/neighbors/plot_lof_outlier_detection.html

Plot results

```
import matplotlib.pyplot as plt
from matplotlib.legend_handler import HandlerPathCollection


def update_legend_marker_size(handle, orig):
    "Customize size of the legend marker"
    handle.update_from(orig)
    handle.set_sizes([20])


plt.scatter(X[:, 0], X[:, 1], color="k", s=3.0, label="Data points")
# plot circles with radius proportional to the outlier scores
radius = (X_scores.max() - X_scores) / (X_scores.max() - X_scores.min())
scatter = plt.scatter(
    X[:, 0],
    X[:, 1],
    s=1000 * radius,
    edgecolors="r",
    facecolors="none",
    label="Outlier scores",
)
plt.axis("tight")
plt.xlim((-5, 5))
plt.ylim((-5, 5))
plt.xlabel("prediction errors: %d" % (n_errors))
```

Fit the model for outlier detection (default)

- from sklearn.neighbors import LocalOutlierFactor
-
- clf = LocalOutlierFactor(n_neighbors=20, contamination=0.1)
- y_pred = clf.fit_predict(X)
- n_errors = (y_pred != ground_truth).sum()
- X_scores = clf.negative_outlier_factor_

Local Outlier Factor (LOF)
Data points
Outlier scores
4
2
0
−2
−4
−4
−2
0
2
4
prediction errors: 8

Clustering-Based

- Basic idea:
  - Cluster the data into groups of different density
  - Choose points in small cluster as candidate outliers
  - Compute the distance between candidate points and non-candidate clusters.
    - If candidate points are far from all other non-candidate points, they are outliers

Applications

- Credit card fraud detection
- telecommunication fraud detection
- network intrusion detection
- fault detection
- many more

Fraud detection

- Detection of criminal activities occurring in commercial organizations.
- Malicious users might be:
  - Employees
  - Actual customers
  - Someone posing as a customer (identity theft)
- Types of fraud
  - Credit card fraud
  - Insurance claim fraud
  - Mobile / cell phone fraud
  - Insider trading
- Challenges
  - Fast and accurate real-time detection
  - Misclassification cost is very high

Intrusion detection

- Intrusion detection
  - Monitor events occurring in a computer system or network and analyze them for intrusions
  - Intrusions defined as attempts to bypass the security mechanisms of a computer or network
- Challenges
  - Traditional intrusion detection systems are based on signatures of known attacks and cannot detect emerging cyber threats
  - Substantial latency in deployment of newly created signatures across the computer system
- Anomaly detection can alleviate these limitations

Healthcare informatics

- Detect anomalous patient records
  - Indicate disease outbreaks, instrumentation errors, etc.
- Key challenges
  - Only normal labels available
  - Misclassification cost is very high
  - Data can be complex: spatio-temporal

Industrial damage detection

- Detect faults and failures in complex industrial systems, structural damages, intrusions in electronic security systems, suspicious events in video surveillance, abnormal energy consumption, etc.
  - Example: aircraft safety
    - anomalous aircraft (engine) / fleet usage
    - anomalies in engine combustion data
    - total aircraft health and usage management

- Key challenges
  - Data is extremely large, noisy, and unlabelled
  - Most of applications exhibit temporal behavior
  - Detected anomalous events typically require immediate intervention

Image processing

- Detecting outliers in a image monitored over time
- Detecting anomalous regions within an image
- Used in
  - mammography image analysis
  - video surveillance
  - satellite image analysis
- Key Challenges
  - Detecting collective anomalies
  - Data sets are very large

Variants of anomaly detection problem

- Given a dataset D, find all the data points $\mathbf{x} \in$ D with anomaly scores greater than some threshold t.

- Given a dataset D, find all the data points $\mathbf{x} \in$ D having the top-n largest anomaly scores.

- Given a dataset D, containing mostly normal data points, and a test point **x**, compute the anomaly score of **x** with respect to D.

Summary – Data Preprocessing

- Missing Data
- Noisy Data
- Outliers