INT104 ARTIFICIAL INTELLIGENCE

LECTURE 2- DATA PRE-PROCESSING

Sichen Liu
Sichen.Liu@xjtlu.edu.cn
Xi'an Jiaotong-Liverpool University
西交利物浦大学

CONTENT

- Data Collection
- Discover and Visualize the Data
- Data Preprocessing
- Data Cleaning
- Data Transformation
- Data Reduction

Data Type

Structured Example: tables

- Highly organized
- Usually with a label

Cust.Id	sex	employed	income	marital	vehicles	age	State of residence
2068	F	NA	11300	Married	2	49	Michigan
2073	F	False	0	Married	3	40	Florida
2848	M	TRUE	4500	Never Married	3	22	Georgia
5641	M	TRUE	20000	Never Married	0	22	New Mexico

Unstructured Example: free text

"It was found that a female with a height between 65 inches and 67 inches had an IQ of 125–130"

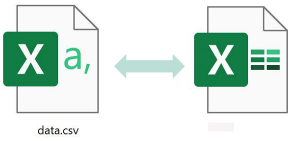
Data Collection

- Lots of places that host/share data online, or you can collect them yourself.
- Open data collections
- Social media data
- Multimodal data

Data Storage and Presentation

- CSV (Comma Separated Values)

```
treat,before,after,diff
No Treatment,13,16,3
No Treatment,10,18,8
No Treatment,16,16,0
Placebo,16,13,-3
```


data.csv

- TSV (Tab Separated Values)

```
Name<TAB>Age<TAB>Address
Ryan<TAB>33<TAB>1115 W Franklin
Paul<TAB>25<TAB>Big Farm Way
Jim<TAB>45<TAB>W Main St
Samantha<TAB>32<TAB>28 George St
```

Data Storage and Presentation

- XML (Extensible Markup Language)

```
<?xml version="1.0" encoding="UTF-8"?>
<bookstore>
    <book category="information science" cover="hardcover">
        <title lang="en">Social Information Seeking</title>
        <author>Chirag Shah</author>
        <year>2017</year>
        <price>62.58</price>
    </book>
    <book category="data science" cover="paperback">
        <title lang="en">Hands-On Introduction to Data
          Science</title>
        <author>Chirag Shah</author>
        <year>2019</year>
        <price>50.00</price>
    </book>
</bookstore>
```

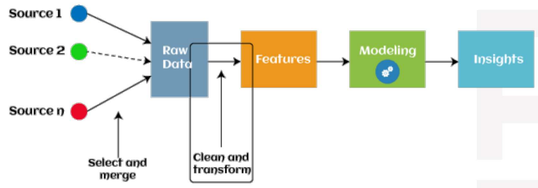
- JSON (JavaScript Object Notation)

```
{
  "squadName" : "Super Hero Squad",
  "homeTown" : "Metro City",
  "formed" : 2016,
  "secretBase" : "Super tower",
  "active" : true,
  "members" : [
    {
      "name" : "Molecule Man",
      "age" : 29,
      "secretIdentity" : "Dan Jukes",
      "powers" : [
        "Radiation resistance",
        "Turning tiny",
        "Radiation blast"
      ]
    },
```

Data Visualization

- Data Visualization in Python
  - Matplotlib
  - Seaborn
  - Pandas.plot
  - ………
- Common Format
  - Line Charts
  - Bar Graphs
  - Histograms
  - Scatter Plots
  - Heat Maps

Data Pre-processing



Goal: to improve the quality of data, reduce errors and inconsistencies, and prepare the data for further analysis or modeling.

Data Pre-processing

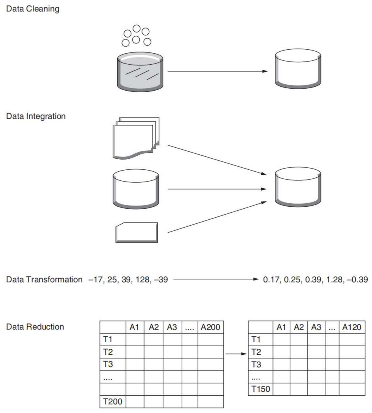
- Feature: an individual measurable property or characteristic of a phenomenon.
- Instance: a sample or data point, refers to a single observation or example in the dataset
- Target variable
- Dataset: A dataset is a collection of instances, features, and target variables that are used to train and test machine learning models.

Feature

Cust.Id	income	vehicles	age
2068	11300	2	49
2073	0	3	40
2848	4500	3	22
5641	20000	0	22

Instance

Data Pre-processing



Data Cleaning

- Data Munging

Example: “Add two diced tomatoes, three cloves of garlic, and a pinch of salt in the mix.”

Table 2.2 Wrangled data for a recipe.

Ingredient	Quantity	Unit/size
Tomato	2	Diced
Garlic	3	Cloves
Salt	1	Pinch

Data Cleaning

- Handling Missing Data
  - Get rid of the corresponding instance.
  - Get rid of the whole column.
  - Set the values to some value (zero, the mean, the median, etc.).
- Smooth Noisy Data
  - Identify or remove the outliners
  - Try to resolve the inconsistent

  (there is no one way to remove noise, or smooth out the noisiness in the data)

Practice: Data Cleaning

#	Country	Alcohol (L/person)	Deaths (Per 100k)	Heart (Per 100k)	Liver (Per 100k)	Free healthcare
1	Australia	2.5	785	211	15.30000019	Y
2	Austria	3.000000095	863	167	45.59999847	Y
3	Belg/Lux	2.900000095	883	131	20.70000076	N
4	Canada	2.400000095	793	NA	16.39999962	Y
5	Denmark	2.900000095	971	220	23.89999962	Y
6	Finland	0.800000012	970	297	19	N
7	France	9.100000381	751	11	37.90000153	N
8	Iceland	-0.800000012	743	211	11.19999981	Y
9	Ireland	0.699999988	1000	300	6.5	Y
10	Israel	0.600000024	-834	183	13.69999981	Y
11	Italy	27.900000095	775	107	42.20000076	Y
12	Japan	1.5	680	36	23.20000076	N
13	Netherlands	1.799999952	773	167	9.199999809	N
14	New Zealand	1.899999976	916	266	7.699999809	Y
15	Norway	0.0800000012	806	227	12.19999981	N
16	Spain	6.5	724	NA	NA	Y
17	Sweden	1.600000024	743	207	11.19999981	N
18	Switzerland	5.800000191	693	115	20.29999924	N
19	UK	1.299999952	941	285	10.30000019	Y
20	US	1.200000048	926	199	22.10000038	N
21	West Germany	2.700000048	861	172	36.70000076	Y

Data Integration

How to integrate multiple databases or files:

Combine

- Combine data from multiple sources into a coherent storage place (e.g., a single file or a database).

Resolve conflicts

- Different representations or different scales; for example, metric vs. British units.

Remove redundant

- The same attribute may have different names in different databases.
- One attribute may be a “derived” attribute in another table; for example, annual revenue.
- Correlation analysis may detect instances of redundant data

Data Transformation

Data must be transformed so it is consistent and readable (by a system)

- Handling Text and Categorical Attributes

  i.e, [“cat1”], [“cat2”], [“cat3”], [“cat4”]

  - Ordinal encoder: `from sklearn.preprocessing import OrdinalEncoder`

    [0], [1], [2], [3]

  - One-hot encoder: `from sklearn.preprocessing import OneHotEncoder`

    [1,0,0,0], [0,1,0,0], [0,0,1,0], [0,0,0,1]

Data Transformation

- Normalization
  - Min–max normalization.

$$x_{scaled} = \frac{x - x_{min}}{x_{max} - x_{min}}$$

  - Z-score normalization.

    Normalizing every value in a dataset such that the mean of all of the values is 0 and the standard deviation is 1

$$x_{scaled} = \frac{x - mean}{sd}$$

  - Normalization by decimal scaling.

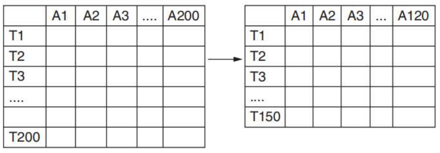
$$x_{scaled} = \frac{x}{10^j}$$

Practice: Data Transformation

#	Country	Alcohol (L/person)	Deaths (Per 100k)	Heart (Per 100k)	Liver (Per 100k)	Free healthcare
1	Australia	2.5	785	211	15.30000019	Y
2	Austria	3.000000095	863	167	45.59999847	Y
3	Belg/Lux	2.900000095	883	131	20.70000076	N
4	Canada	2.400000095	793	NA	16.39999962	Y
5	Denmark	2.900000095	971	220	23.89999962	Y
6	Finland	0.800000012	970	297	19	N
7	France	9.100000381	751	11	37.90000153	N
8	Iceland	-0.800000012	743	211	11.19999981	Y
9	Ireland	0.699999988	1000	300	6.5	Y
10	Israel	0.600000024	-834	183	13.69999981	Y
11	Italy	27.900000095	775	107	42.20000076	Y
12	Japan	1.5	680	36	23.20000076	N
13	Netherlands	1.799999952	773	167	9.199999809	N
14	New Zealand	1.899999976	916	266	7.699999809	Y
15	Norway	0.0800000012	806	227	12.19999981	N
16	Spain	6.5	724	NA	NA	Y
17	Sweden	1.600000024	743	207	11.19999981	N
18	Switzerland	5.800000191	693	115	20.29999924	N
19	UK	1.299999952	941	285	10.30000019	Y
20	US	1.200000048	926	199	22.10000038	N
21	West Germany	2.700000048	861	172	36.70000076	Y

Data Reduction

Data reduction is a key process in which a reduced representation of a dataset that produces the same or similar analytical results is obtained.



	A1	A2	A3	....	A200
T1					
T2					
T3					
....					
T200					

	A1	A2	A3	...	A120
T1					
T2					
T3					
....					
T150					

Feature Selection

- Filter methods – features are selected and ranked according to their relationships with the target;
- Wrapper methods – it's a search for well-performing combinations of features
- Embedded methods – perform feature selection as part of the model training process.

longitude	latitude	housing_median_age	total_rooms	total_bedrooms	population	households	median_income	median_house_value	ocean_proximity
-122.23	37.88	41	880	129	322	126	8.3252	452600	NEAR BY
-122.22	37.86	21	7099	1106	2401	1138	8.3014	358500	NEAR BY
-122.24	37.85	52	1467	190	496	177	7.2574	352100	NEAR BY
-122.25	37.85	52	1274	235	558	219	5.6431	341300	NEAR BY
-122.25	37.85	52	1627	280	565	259	3.8462	342200	NEAR BY
-122.25	37.85	52	919	213	413	193	4.0368	269700	NEAR BY
-122.25	37.84	52	2535	489	1094	514	3.6591	299200	NEAR BY

Looking for Correlations

Correlation is a statistical analysis that is used to measure and describe the *strength* and *direction* of the relationship between two variables.

Pearson's r correlation:

$$r = \frac{N\sum xy - \sum x \sum y}{\sqrt{[N\sum x^2 - (\sum x)^2][N\sum y^2 - (\sum y)^2]}}$$

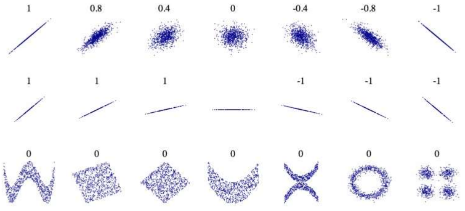


Figure 2-14. Standard correlation coefficient of various datasets (source: Wikipedia; public domain image)

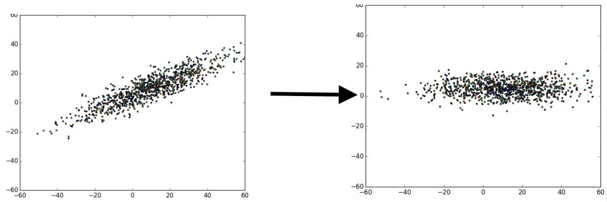
Feature Extraction

Technique in which new features are extracted from the existing ones.

- Identifying and selecting the most relevant and informative features from dataset
- Transforming them into a lower-dimensional space while preserving the most important information.

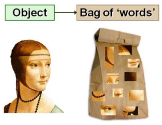
Examples

PCA

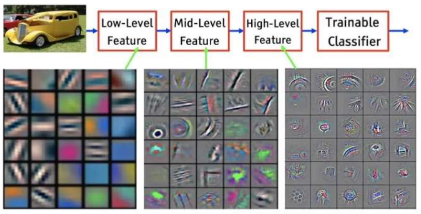


Examples

Bag of words:



CNN:


Low-Level Feature
Mid-Level Feature
High-Level Feature
Trainable Classifier