Installation & Data Preprocessing

Python 3 | Anaconda | Datasets

miTu skillologies

Contents

- Installing Python 3
- Installing Anaconda
- Dataset acquisition
- Importing the Libraries
- Importing the Dataset
- Missing Data
- Categorical Data
- Splitting: Training & Testing
- Feature Scaling

Installing Python3

- Before installing any of the packages, it is mandatory to update your OS with recent patches
- The supported files, libraries, security patches need to updated at least once in a week
- BEGIN in Terminal:
  - **sudo apt-get update**
  - **sudo apt-get install python3.5**
- The first command as it says, updates your system repositories
- The second one installs python 3.5, to check type
  - **python3**

Installing Anaconda

- Traverse into directory where Anaconda's Shell extension file is stored using terminal
  - **cd /Package/**
- Install it using bash
  - **bash Anaconda3-4.3.1-Linux-x86_64.sh**
- Installing is not sufficient, we need to mention it the Python version that they must use
  - **conda install python=3.5**
- To ensure it is running in same env., follow this
  - **python3.5** #check the header for anaconda

> USING Python IDE

- PyCharm IDE installation
- Shortcuts:

Ctrl + Shift + F10	**RUN current Program**
Alt + Shift + F10	**RUN only selected file**
Alt + Shift + X	**RUN recently exected only**
Ctrl + Alt + E	**RUN in Py Console**
Ctrl + `	**Open Utility Menu**

- Create Project from Start up menu
- Create New Python file

Dataset Acquisition

- Unzip the given Zipped file; named ***Machine Learning A-Z ds.zip***
- In the folder with name Data Pre-processing, you will see your dataset named ***Data.csv***
- Drag and Drop the ***Data.csv*** on Project directory in PyCharm

Importing Dataset

- **import**
  - using this statement we can import packages / libraries inside python
  - to import dataset we need special library to perform Dataset import
  - **Pandas**, is the required dataset for same
- **import pandas**
  - Using package name entirely increases keystrokes, to save it we give Alias/Name name
- **import pandas as pd**
  - Now we can use **pd**, everytime we need to call it

Missing Data

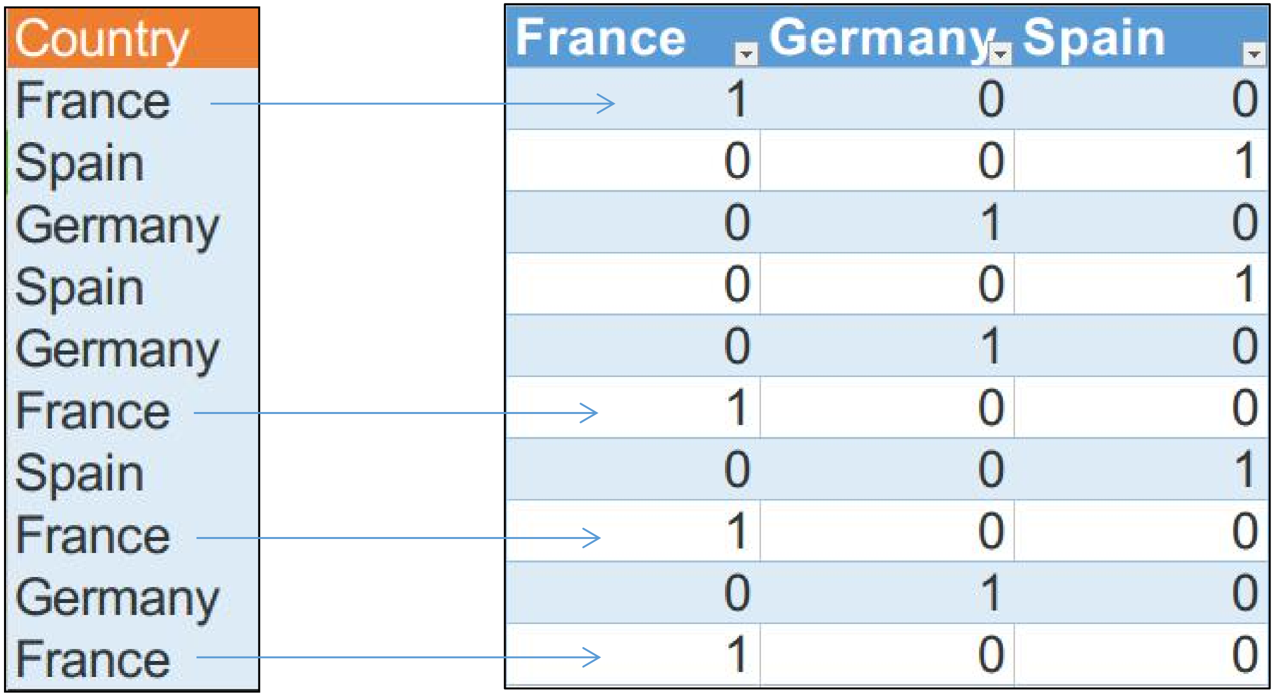
- Predicting the missing values using Averaging / Mean
- **Preprocessing** from **SKlearn** can handle such tasks using **Imputer**

Categorical Data

- Categorizing the Repetative Strings into values
- Values so that they can be given into equations
- Let's convert them to numbers

Categorical Data

DUMMY ENCODING



Country
France
Spain
Germany
Spain
Germany
France
Spain
France
Germany
France

France	Germany	Spain
1	0	0
0	0	1
0	1	0
0	0	1
0	1	0
1	0	0
0	0	1
1	0	0
0	1	0
1	0	0

Splitting - Training & Testing

- Machine Learning performance improves with new Co-relations
  - Eg.

VS

MEMORIZATION

CORELATION

Feature Scaling

- Varying nature of Data
  - AGE: 27 to 47
  - SALARY: 40K to 80K
- Lose of Scaling
- ML are based on Euclidean distances

Euclidean

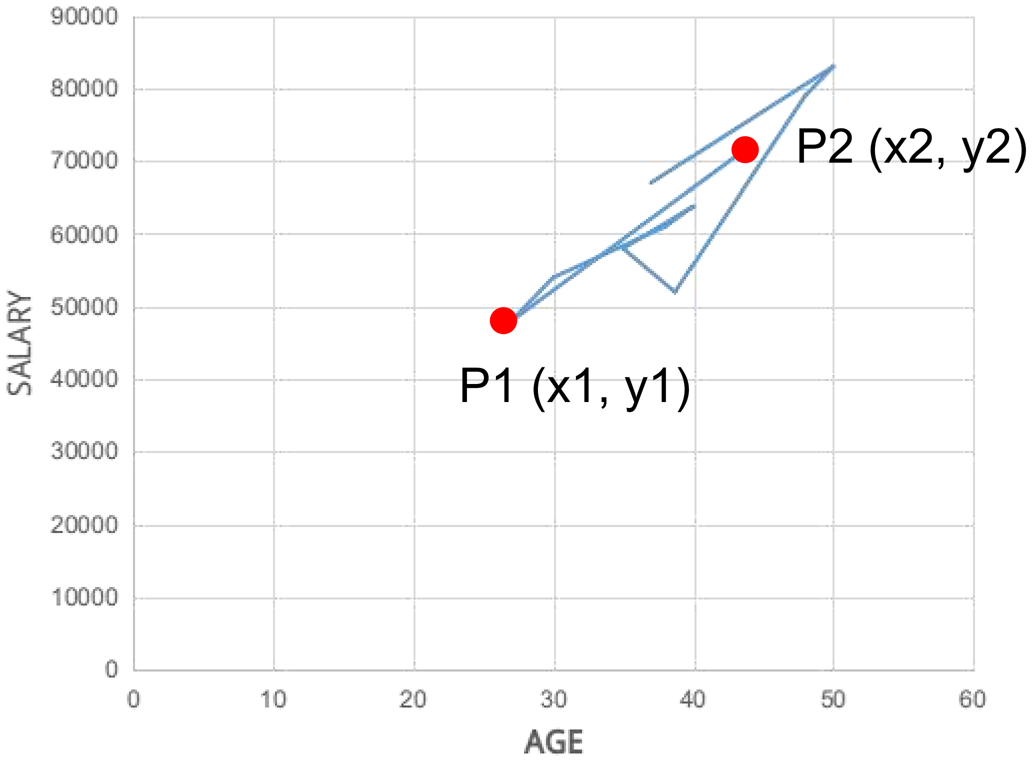
juːˈklɪdɪən | adjective

is Two data points is the Sq root of Sum of the squared co-ordinates

Feature Scaling

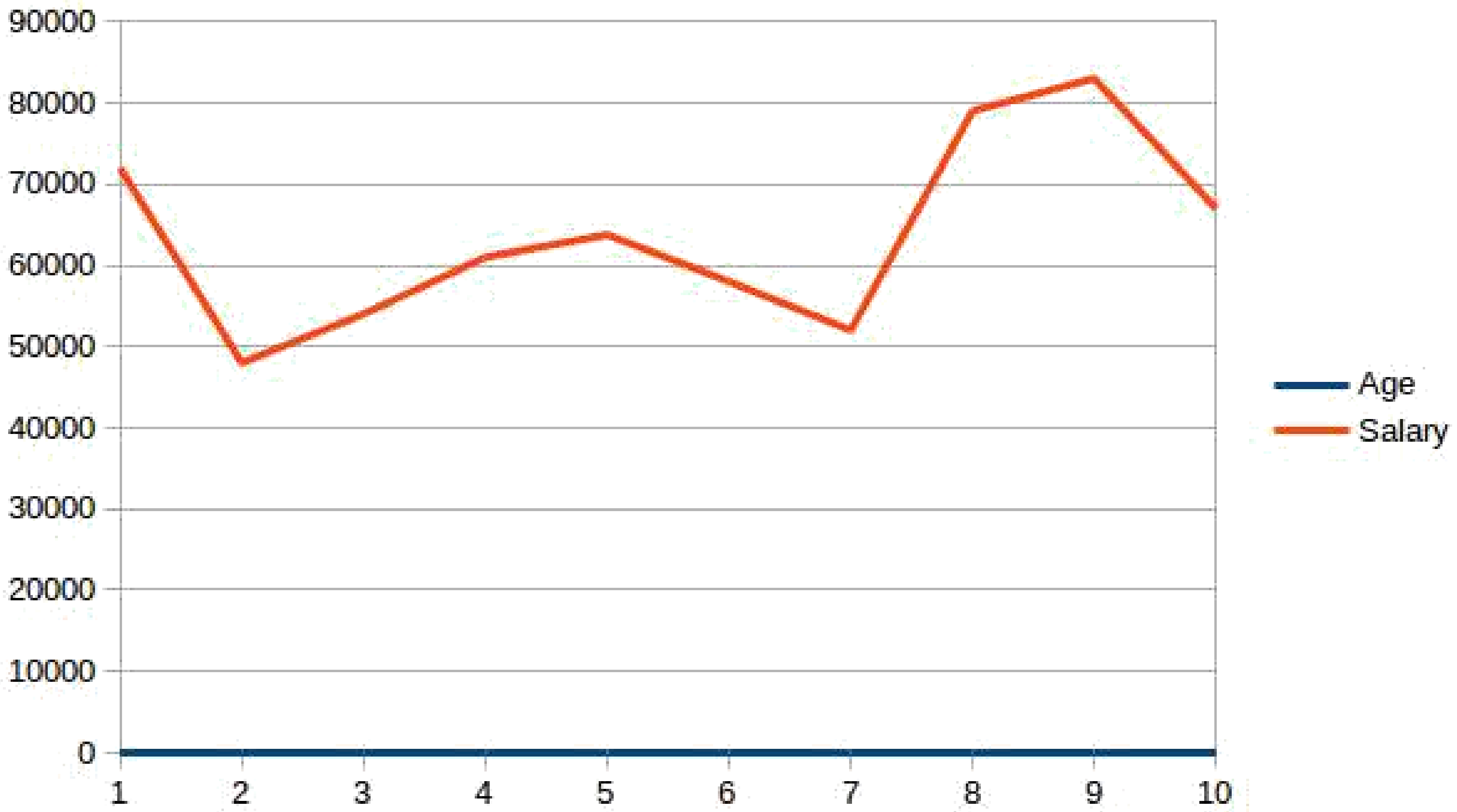
- Euclidean Distance

Distance b/w P1 & P2 = $\sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$



Feature Scaling

- Actual Plotting of Values



Feature Scaling

EUREKA!

- Scale values from -1 to +1 to get both the AXES in same range
- Eliminate Domination

Congratulations!
DAY 1 Accomplished!

@mitu_skillologies

/mITuSkillologies

@mitu_group