Code – Along Saturday

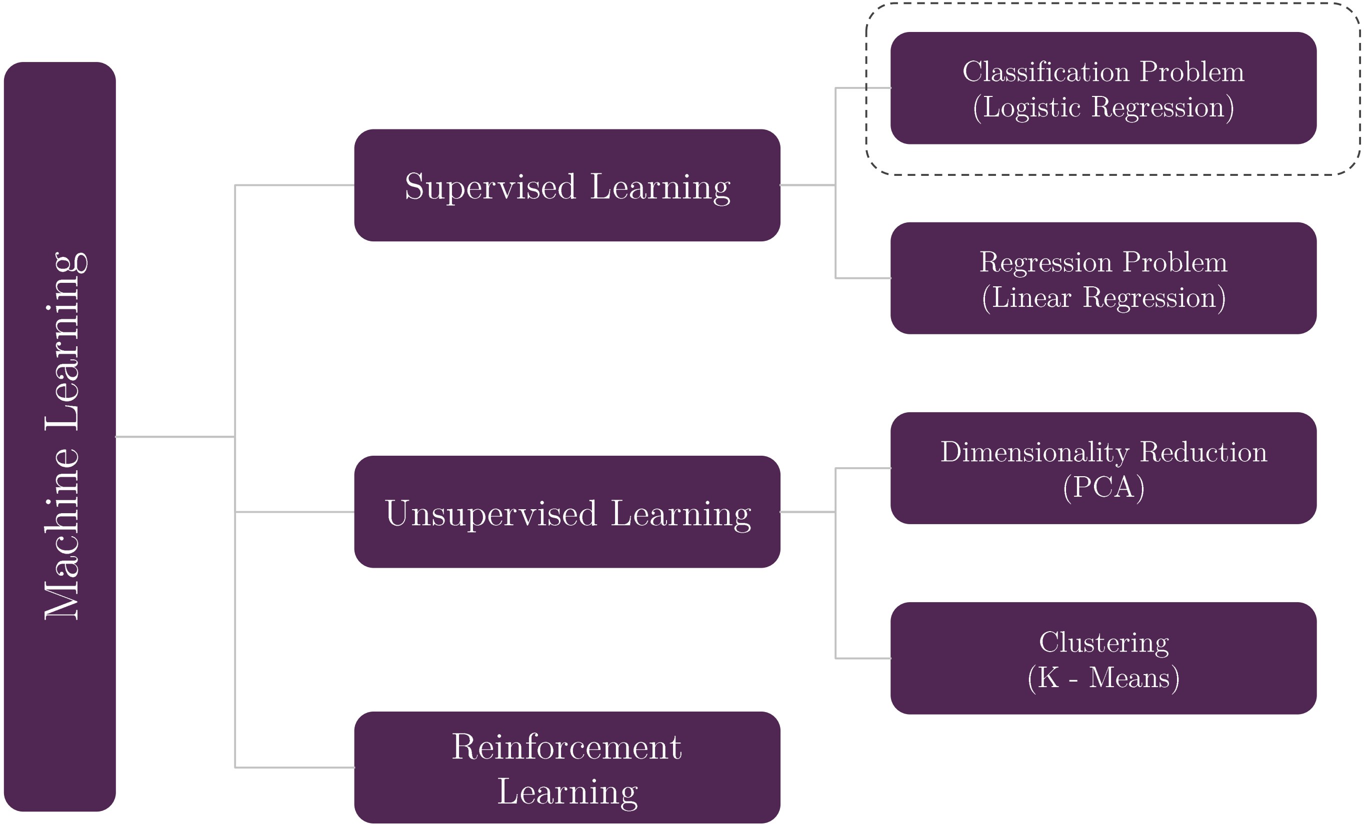
Exploratory data Analysis & Feature Engineering

Prerequisite

Must have things

- Anaconda should be installed with python *.* version
- Basic understanding of Matrix Algebra
- Basic Understanding of Python

Machine Learning Classification



Steps in Building Machine Learning Models

Step 1

Get the Data

- Creating Isolated Environment
- Importing data
- Quick look on Data Structure
- Creating a test set

Step 2

Data Exploration and Visualization

- Data Exploration
- Data Visualization
- Looking for Correlations

Step 3

Data Preparation for ML Algorithms

- Data Cleaning
- Handling Text Attributes
- Creating Pipelines

Step 4

Train and Fine - tuning the model

- Training and Evaluation on training set
- Better Evaluation using Cross Validation
- Finalizing the model
- Prediction on test set

Let's get started...

Get the data

- Create Workspace / Virtual Environment
- Importing Data
- Quick look at the data structure
- Creating a test set

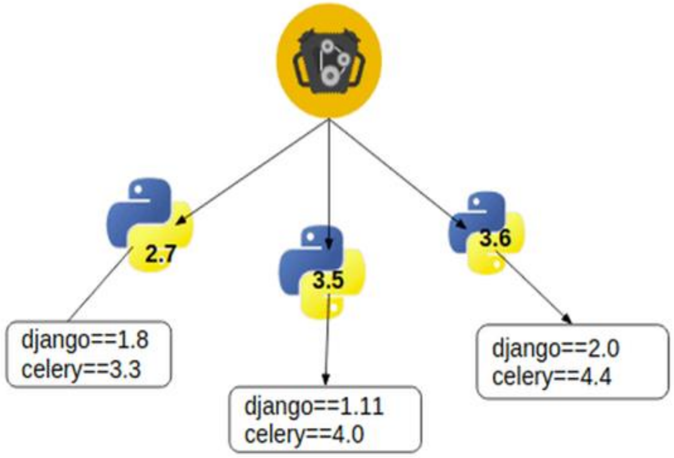
Create Workspace / Virtual Environment

- Also known as Isolated environment

- Allows us to work with different version of python and its packages for different projects

- Two ways to install packages:
  - conda
  - pip

- Local packages installation



Demo

Get the data

- Create Workspace / Virtual Environment
- Importing Data
- Quick look on the data structure
- Creating a test set

Get the data

- Create Workspace / Virtual Environment
- Importing Data
- Quick look on the data structure
- Creating a test set

Quick look on the data structure

- The data is related with direct marketing campaigns of a Portuguese Banking Institution.
- These marketing campaigns are based on phone calls
- Classification goal is to predict if the customer will subscribe a term deposit (variable y)
- Find the data on below link:
  - https://archive.ics.uci.edu/ml/datasets/bank+marketing

Information about data features – 1

- Bank – Client Data

  1. age (numeric)
  2. job (categorical: 'admin.', 'blue-collar', 'entrepreneur', 'housemaid', 'management', 'retired', 'self-employed', 'services', 'student', 'technician', 'unemployed', 'unknown')
  3. marital : marital status (categorical: 'divorced', 'married', 'single', 'unknown'; note: 'divorced' means divorced or widowed)
  4. education (categorical: 'basic.4y', 'basic.6y', 'basic.9y', 'high.school', 'illiterate', 'professional.course', 'university.degree', 'unknown')
  5. default: has credit in default? (categorical: 'no', 'yes', 'unknown')
  6. housing: has housing loan? (categorical: 'no', 'yes','unknown')
  7. loan: has personal loan? (categorical: 'no', 'yes', 'unknown')

Information about data features – 2

- Contact of the current campaign:
    - contact: contact communication type (categorical: 'cellular’, 'telephone')
    - month: last contact month of year (categorical: 'jan', 'feb', 'mar', ..., 'nov', 'dec’)
    - day of week: last contact day of the week (categorical: 'mon’, 'tue’, 'wed’, 'thu’, 'fri')
    - duration: last contact duration, in seconds (numeric).

Information about data features – 3

- other attributes:
  - campaign: number of contacts performed during this campaign and for this client (numeric, includes last contact)
  - pdays: number of days that passed by after the client was last contacted from a previous campaign (numeric; 999 means client was not previously contacted)
  - previous: number of contacts performed before this campaign and for this client (numeric)
  - poutcome: outcome of the previous marketing campaign (categorical: 'failure','nonexistent','success’)

Information about data features – 4

- social and economic context attributes
  - emp.var.rate: employment variation rate - quarterly indicator (numeric)
  - cons.price.idx: consumer price index - monthly indicator (numeric)
  - cons.conf.idx: consumer confidence index - monthly indicator (numeric)
  - euribor3m: euribor 3 month rate - daily indicator (numeric)
  - nr.employed: number of employees - quarterly indicator (numeric)

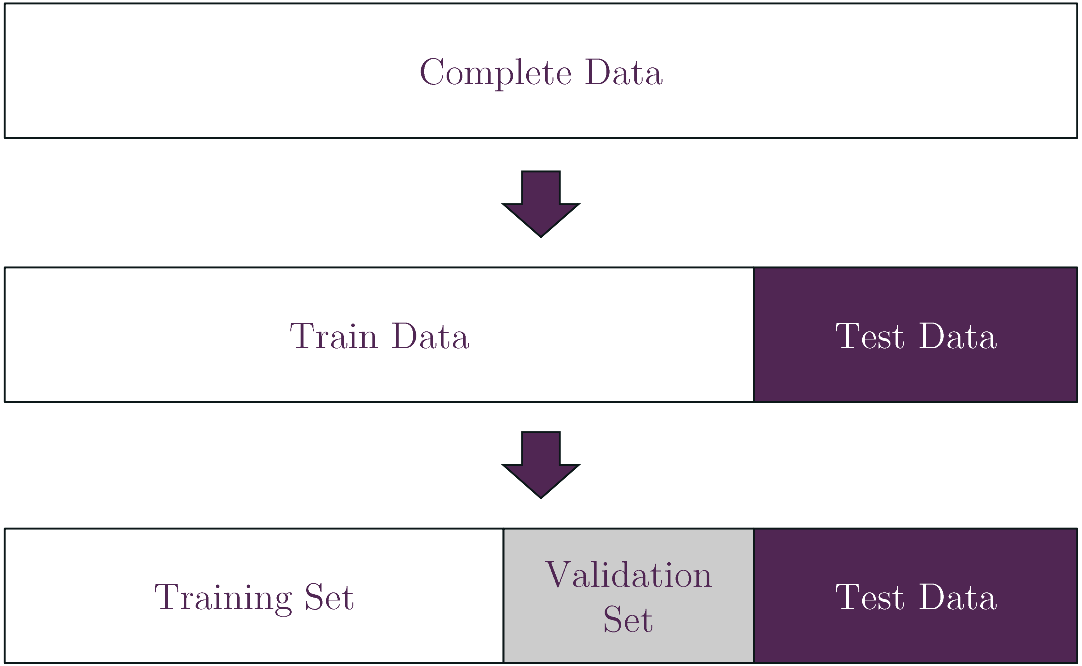
- Output variable (desired target):
  - y - has the client subscribed a term deposit? (binary: 'yes', 'no')

Get the data

- Create Workspace / Virtual Environment
- Importing Data
- Quick look on the data structure
- Creating a test set

Creating a test set

- Ratio is dependent on the size of the data
- The following percentages (for Train:Validation:Test) are considered:
  - 70:20:10
  - 90:5:5
  - 97:2:1



Data Exploration and Visualization

- Data Exploration
- Data Visualization
- Looking for Correlations

Data Exploration and Visualization

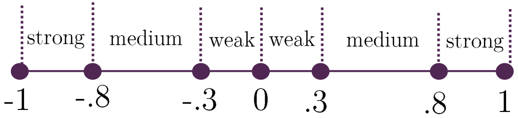
- Data Exploration
- Data Visualization
- Looking for Correlations

Data Exploration and Visualization

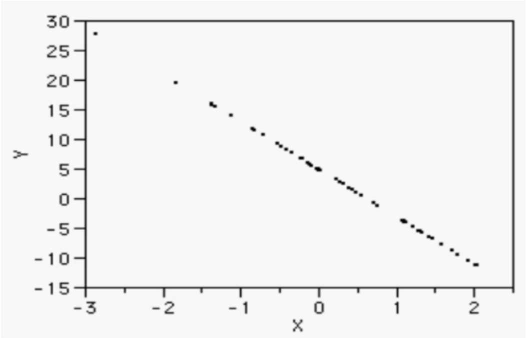
- Data Exploration
- Data Visualization
- Looking for Correlations

Linear Correlation

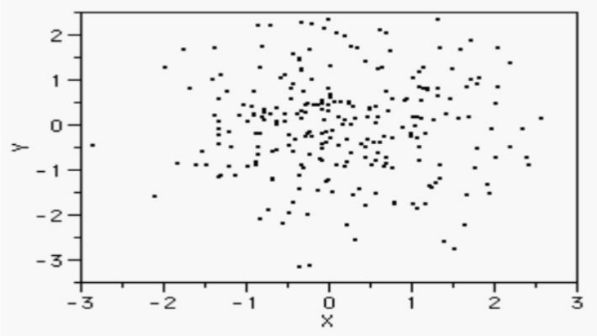
- It reflects the degree of linear relationship between two variables
- It is symmetric
- Correlation between x and y is same as correlation between y and x
- It ranges from -1 to +1



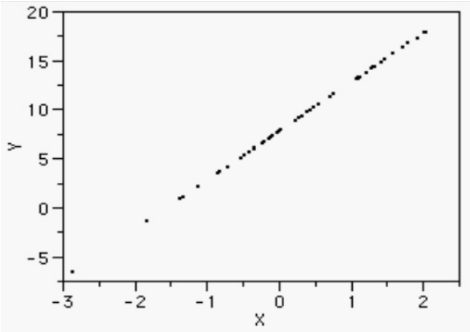
$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$



r = -1



r = 0



r = 1

Preparing data for ML algorithms

- Data Cleaning / Missing Value Treatment
- Handling Text and Categorical Attributes
- Feature Scaling
- Transformation pipelines

How to deal with missing values

- Method 1: By removing the rows with missing value
- Method 2: By Imputation
  - For numerical variables, missing values will replace by median or mean value of the variable
  - For categorical variables, missing values will replace by mode
- Method 3: By considering as a separate class

Preparing data for ML algorithms

- Data Cleaning / Missing Value Treatment
- Handling Text and Categorical Attributes
- Feature Scaling
- Transformation pipelines

Label Encoding & One – Hot Encoding

- Label Encoding
  - Give every class of a categorical variable a unique numerical ID
  - It doesn't increase the dimension of the data

Platform
Desktop
Mobile
tablet

Platform
1
2
3

- One – Hot Encoding
  - Transform a categorical variable of m classes into m binary features
  - Also known as converting data into sparse format

platform
desktop
mobile
tablet

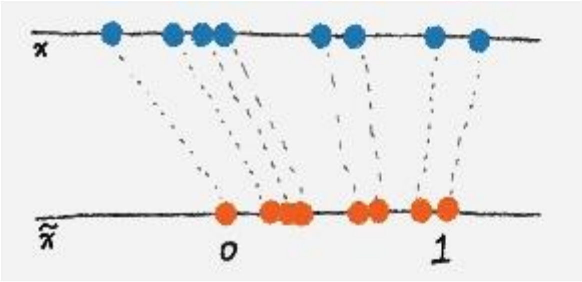
platform=desktop	platform=mobile	platform=tablet
1	0	0
0	1	0
0	0	1

Preparing data for ML algorithms

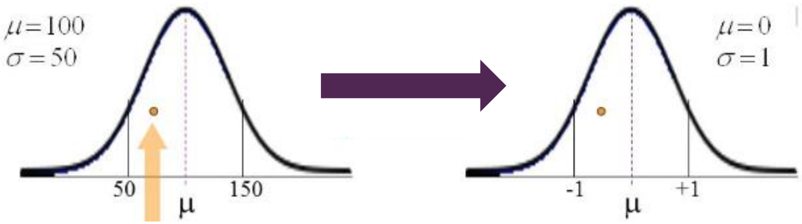
- Data Cleaning / Missing Value Treatment
- Handling Text and Categorical Attributes
- Feature Scaling
- Transformation pipelines

Different types of Scalers

- MinMaxScaler
  - It squeezes all the values within the range of 0 and 1
  - $\tilde{x} = \dfrac{x - \min(x)}{\max(x) - \min(x)}$



- StandardScaler
  - It shifts the distribution of each attribute to have a mean of 0 and a standard deviation of 1 (unit variance)
  - $\tilde{x} = \dfrac{x - mean(x)}{\mathrm{sd}(x)}$



Preparing data for ML algorithms

- Data Cleaning / Missing Value Treatment
- Handling Text and Categorical Attributes
- Feature Scaling
- Transformation pipelines

Scikit-Learn Design Principles

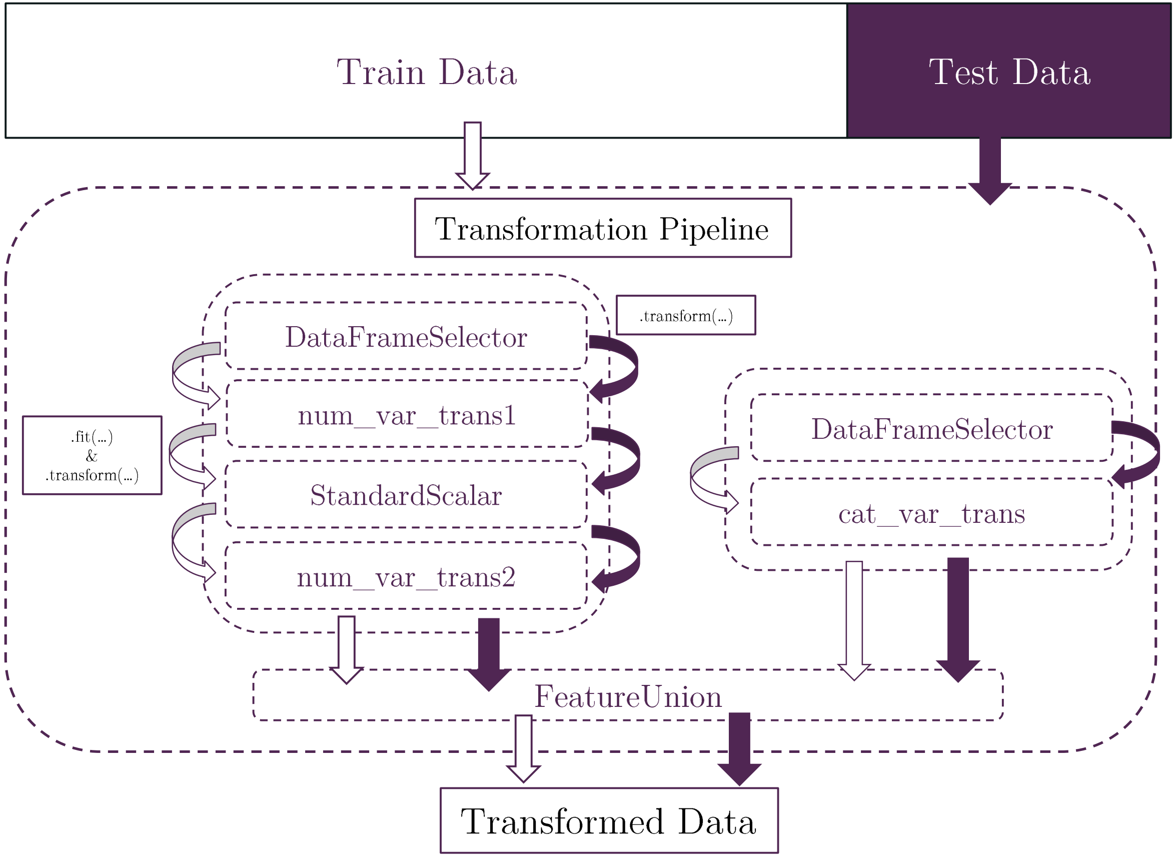
- Consistency
  - *Estimators* (any object that can estimate some parameters based on the dataset)
    - Estimation is performed by the $fit()$ method and takes dataset as a parameter
  - *Transformers* (Transformation of the dataset)
    - It is performed by $transform()$ method with dataset as a parameter and returns the transformed dataset
  - *Predictors* (Making predictions for given dataset)
    - It is performed by $predict()$ method with dataset as a parameter

$fit\_transform()$

Faster

Transformation Pipeline

- Pipeline is a chain of several steps
- Easy to reproduce and productise the data



Train and Fine-Tuning the model

- Training and evaluating on the training set
- Better evaluation using Cross-Validation
- Finalizing the model
- Predicting the test set

Confusion Matrix

		Predicted Class	
		0	1
Observed Class	0	TN	FP
	1	FN	TP

TP	True Positive
TN	True Negative
FN	False Negative
FP	False Positive

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

$$Precision = \frac{TP}{TP + FP}$$

$$Recall\ (Sensitivity\ or\ TPR) = \frac{TP}{TP + FN}$$

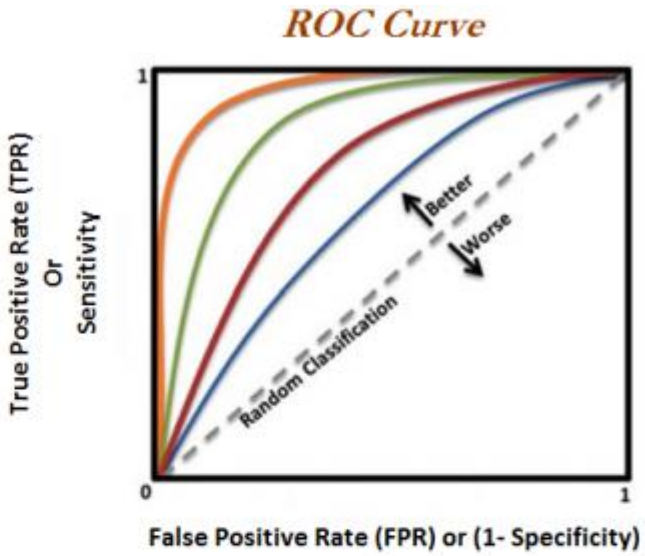
$$f1\ score = \frac{2 * Precision * Recall}{Precision + Recall}$$

$$FPR\ (1 - Specificity) = \frac{FP}{FP + TN}$$

$$Specificity\ (TNR) = \frac{TN}{FP + TN}$$

ROC Curve & AUC

- ROC Curve
  - Also known as Receiver Operating Curve
  - Plot of test Sensitivity on the Y – axis versus its False Positive Rate (FPR) on the X – axis
  - Each discrete point on the graph called the Operating Point
- AUC (Area Under the Curve)
  - AUC provides the overall measure of test accuracy
  - Higher the AUC the better the overall performance of the test



Train and Fine-Tuning the model

- Training and evaluating on the training set
- Better evaluation using Cross-Validation
- Finalizing the model
- Predicting the test set

K - Fold Cross Validation

Train Data				
Training Set	Training Set	Training Set	Validation Set	E_1
Training Set	Training Set	Validation Set	Training Set	E_2
Training Set	Validation Set	Training Set	Training Set	E_3
Validation Set	Training Set	Training Set	Training Set	E_4

$$E_{4-fold} = \frac{E_1 + E_2 + E_3 + E_4}{4}$$

- Helps in identifying the overfitting and underfitting of data
- Helps in getting the best hyperparameter of the model

Train and Fine-Tuning the model

- Training and evaluating on the training set
- Better evaluation using Cross-Validation
- Finalizing the model
- Predicting the test set

Train and Fine-Tuning the model

- Training and evaluating on the training set
- Better evaluation using Cross-Validation
- Finalizing the model
- Predicting the test set

Questions?