DATA SCIENCE WITH PYTHON

A THREE DAYS WORKSHOP

Aatif Imtiaz Butt, Ph.D

Department of Physics and Applied Mathematics
Pakistan Institute of Engineering and Applied Sciences

January 11-13, 2022

1. Why Pursue Career in Data Science?
2. Introduction to Python and its Libraries
3. NumPy for Scientific Comuting
4. Structured Databases with Pandas
5. Data Visualization Tools
6. Data Science Methodology
7. Machine Learning with Python

Motivation and Pathway to Data Science

$ $ $

Country	Mean	Std	Size
United States	$107,517.14	$57,177.29	4509
New Zealand	$94,279.41	$81,386.29	58
Australia	$92,893.31	$45,040.37	185
Germany	$83,695.92	$80,376.15	97
Canada	$74,856.31	$25,320.06	244
United Kingdom	$65,566.59	$40,704.08	667
Sweden	$62,227.62	$14,577.25	87
Netherlands	$60,869.7	$39,286.6	82
South Africa	$54,704.39	$24,272.97	57
India	$33,226.41	$78,978.6	112

Table displaying mean salary for a data professional by Country

www.towardsdatascience.com
www.salaryexpert.com

Instagram Accounts to Follow

1. **@python.hub** with 855K Followers
2. **@learn.machinelearning** with 232K Followers
3. **@pythoncoder2.0** with 83.5K Followers

PATHWAY TO A SOLID DATA SCIENTIST

Certifications:

1. **IBM Data Science Professional Certification**
2. **MIT's MicroMasters Program on Statistics and Data Science**
3. **Cloud Computing with Microsoft's Azure or Amazon's AWS**

YouTube Channels for Data Science:

1. **Derek Banas**
2. **freeCodeCamp.org**

Books to Keep:

1. **Python Data Science Handbook by Jake VanderPlas**
2. **Introduction to Programming in Python by Robert Sedgewick, Kevin Wayne and Robert Dondero**

Introduction to Python

Python Libraries for High Performance

- NumPy – Math Library for Efficient and Effective Computation
- SciPy – Collection of Numerical Algorithms and Domain Specific Toolboxes
- Pandas – Exploratory Tool for Structured Databases
- Matplotlib – Popular Plotting Package
- Scikit-Learn – Collection of Algorithms and Tools for Machine Learning

NumPy for Scientific Comuting

Any data that need be analyzed need be transformed into arrays of numbers

Sound Clips are 1D arrays of Intensity vs Time

Digital Images are 2D arrays of numbers representing pixel brightness across the panel

Manipulating Matrices and Vectors is at the heart of Game Development

NumPy enables us to effectively load, store and manipulate in-memory dense data in Python

We will use JupyterLab in the Lab to learn about various features of NumPy

Pandas for Databases and Spreadsheets

Built on top of NumPy, Pandas can be considered as enhanced version of NumPy structured arrays in which rows and columns are identified with Labels

Pandas provide numerous tools to explore Tabular Data in many Databases and Spreadsheets

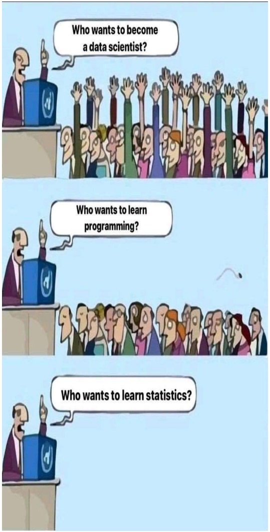
Pandas is widely used in Data Preparing, Data Cleaning and Data Analysis

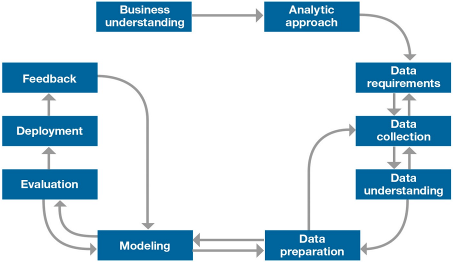
We will use JupyterLab in the Lab to learn about various features of Pandas

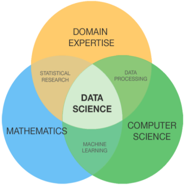
Visualization

Plotly is a more sophisticated data visualization tool that is better suited for creating elaborate plots more efficiently

Data Science Methodology







Soft Introduction to Machine Learning

Artificial Intelligence vs **Machine Learning** vs **Deep Learning**

A Subfield of Computer Science that gives computers the ability to learn without being explicitly programmed

Objective: A Good Decision Tree based on Historical Data

- Real Estates: price of an asset?
- Healthcare: if cell growth is benign or malignant?
- Finance: if loan application be approved?
- Entertainment: personalized recommendations by Netflix, Amazon, YouTube

Major Machine Learning Techniques

- **Regression/Estimation** – predicting continuous values
- **Classification** – predicting category of a case
- **Clustering** – finding structure of data
- **Association** – identifying frequently co-occuring events
- **Anomaly Detection** – discovering unusual cases
- **Sequence Mining** – predicting next event
- **Dimension Reduction** – reducing dataset size
- **Recommendation Systems** – recommending items

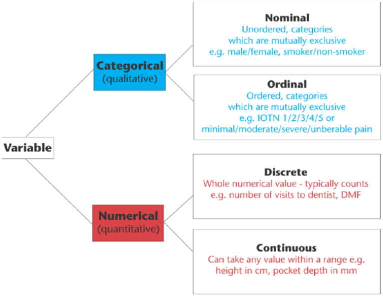
Algorithms: Supervised vs Unsupervised

Supervised Learning

- Labeled Data
- Regression Models: Predict trend using previously labeled data
- Classification Models: Classify labeled data
- Numerous algorithms; controlled environment

Unsupervised Learning

- Unlabeled Data
- Clustering Models: Find patterns and groupings in unlabeled data
- Fewer algorithms; uncontrolled environment



Regression Models

A Database with Categorical and Continuous Headers

MODELYEAR	MAKE	MODEL	VEHICLECLASS	ENGINESIZE	CYLINDERS	TRANSMISSION	FUELTYPE	FUELCONSUMPTION_CITY	FUELCONSUMPTION_HWY	FUELCONSUMPTION_COMB	FUELCONSUMPTION_COMB_MPG	CO2EMISSIONS
2014	ACURA	ILX	COMPACT	2	4	AS5	Z	9.9	6.7	8.5	33	196
2014	ACURA	ILX	COMPACT	2.4	4	M6	Z	11.2	7.7	9.6	29	221
2014	ACURA	ILX HYBRID	COMPACT	1.5	4	AV7	Z	6	5.8	5.9	48	136
2014	ACURA	MDX 4WD	SUV - SMALL	3.5	6	AS6	Z	12.7	9.1	11.1	25	255
2014	ACURA	RDX AWD	SUV - SMALL	3.5	6	AS6	Z	12.1	8.7	10.6	27	244
2014	ACURA	RLX	MID-SIZE	3.5	6	AS6	Z	11.9	7.7	10	28	230
2014	ACURA	TL	MID-SIZE	3.5	6	AS6	Z	11.8	8.1	10.1	28	232
2014	ACURA	TL AWD	MID-SIZE	3.7	6	AS6	Z	12.8	9	11.1	25	255
2014	ACURA	TL AWD	MID-SIZE	3.7	6	M6	Z	13.4	9.5	11.6	24	267
2014	ACURA	TSX	COMPACT	2.4	4	AS5	Z	10.6	7.5	9.2	31	212

Simple Regression: One independent variable is used to estimate the dependent variable. Simple regression can be linear or non-linear.

Multiple Regression: Multiple independent variables are considered to estimate the dependent variable. Multiple regression can be linear or non-linear. Beware of OverFit Modeling.

Algorithms: Ordinal Regression; Poisson Regression; Fast Forest Quantile Regression; Linear, Polynomial, Lasso, Stepwise, Ridge Regression; Bayesian Linear Regression; Neural Network Regression; Decision Forest Regression; Boosted Decision Tree Regression; KNN (K-Nearest Neighbour);

Each Regression Algorithm has its own specific conditions to when it is best suited

Simple Linear Regression

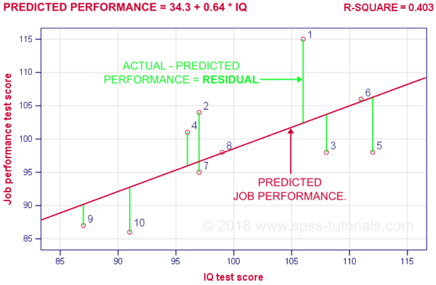
Least Square Fitting:

Dataset: (x_i, y_i) where $i : 0 \rightarrow n$
$[x_i]$ is an array of independent variable
$[y_i]$ is an array of dependent variable

$\hat{y} = \theta_0 + \theta_1 x$ is a fit line, where:
θ_0 is the intercept
θ_1 is the coefficient
$|\hat{y}_i - y_i|$ is the residual error at x_i

How to distinguish between a Good Fit and a Bad Fit?

Minimizing the sum of the squares of residual errors yields:
$\theta_0 = \frac{\overline{x^2}\,\overline{y} - \overline{x}\,\overline{xy}}{\overline{x^2} - \overline{x}^2}$ and $\theta_1 = \frac{\overline{xy} - \overline{x}\,\overline{y}}{\overline{x^2} - \overline{x}^2}$
and
$\delta_y = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$



Evaluation Metrics in Regression Modeling

Evaluation Metrics quantitatively measure the performance of the Model on Prediction

Mean Absolute Error: Easiest of the metrics to understand

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

Mean Squared Error: Focus is geared towards large errors

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

Root Mean Squared Error: Most popular because of same units

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

Evaluation Metrics – *continued*

Relative Absolute Error:

$$\mathrm{RAE} = \frac{\sum_{i=1}^{n} |y_i - \hat{y}_i|}{\sum_{i=1}^{n} |y_i - \bar{y}_i|}$$

Relative Squared Error: Widely adopted by Data Science Community

$$\mathrm{RSE} = \frac{\sum_{i=1}^{n} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{n} (y_i - \bar{y}_i)^2}$$

R-Squared: Popular metric for the accuracy of the model

$$\mathrm{R}^2 = 1 - \mathrm{RSE}$$

LAB: Simple Linear Regression

We wish to design a new car having least value of CO_2 emission. We suspect that engine capacity, number of cylinders and fuel consumption may play significant role in varying levels of CO_2 emission in automobile industry.

A dataset of 1067 cars manufactured in 2014 is available for analysis. It contains a total of 13 fields, some categorical and continuous.

Let's make use of NumPy, Pandas, Visualization Tools and Scikit-Learn libraries to systematically explore the data and conclude a solid prediction.

Multiple Linear Regression

Two Applications:

1. Effectiveness of a given independent variable on prediction
2. Predict the impact due to changes in independent variable

$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \cdots + \theta_n x_n$$

$$\hat{y} = \Theta^T X$$

$$\Theta^T = [\theta_0 \ \theta_1 \ \theta_2 \cdots \theta_n], X = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} \tag{1}$$

Θ is called **Parameters** or **Weight Vector** of Regression Equation

Multiple Linear Regression – *continued*

The whole idea is to find the best fit hyper-plane in higher dimensions

Estimating the Weight Vector Θ:

- Ordinary Least Squares
  - Uses linear algebra operations
  - Takes a long time for dataset with rows greater than 10k
- An Optimization Algorithm
  - Gradient Descent
  - Good choice for dataset with rows greater than 10k

How many independent variables should be used to estimate the dependent variable?

Adding too many independent variables without theoretical justification may lead to OverFit Model.

Can we use Categorical fields as independent variables?

LAB: Multiple Linear Regression

Using the same dataset as in the case of Simple Linear Regression, we try to make prediction using multiple independent variables.

We will also try to use as many independent variables as possible to observe an OverFit Modeling.

Explained Variance Score $= 1 - \frac{\mathrm{Var}(\hat{y} - y_i)}{\mathrm{Var}(y_i)}$

Polynomial Regression

Is this an example of Simple Non-Linear Regression?

$$\hat{y} = \theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \cdots + \theta_n x^n$$

No, it is NOT!!!

$$x \to x_1, x^2 \to x_2, x^3 \to x_3, \ldots, x^n \to x_n$$

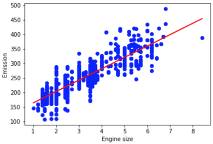
yields

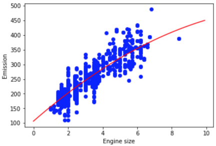
$$\hat{y} = \theta_0 + \theta_1 x_1 + \theta_2 x_2 + \theta_3 x_3 + \cdots + \theta_n x_n$$

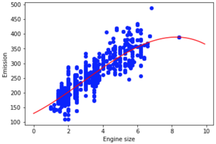
We just converted a polynomial Regression of degree n in one independent variable to a problem of Multiple Linear Regression in n independent variables.

We will explore more in the Lab

Is straight line a Good Fit?
Polynomial of degree 2?
Polynomial of degree 3?







Non-Linear Regression

The dependent variable should be a non-linear function of the Weight Vector Θ

Examples are:

$$\hat{y} = \theta_0 + \theta_0\theta_1 x$$

$$\hat{y} = \theta_0 + \theta_0\theta_1^x$$

$$\hat{y} = \log\left[\theta_0 + \theta_1 x + \theta_2 x^2 + \theta_3 x^3 + \cdots + \theta_n x^n\right]$$

$$\hat{y} = \frac{\theta_0}{1 + \theta_1^{x+\theta_2}}$$

Let's get our hands on the Lab to learn first-hand about Non-Linear Regression Modeling

Classification

It is a supervised learning approach

We categorize some unknown items into discrete set of categories or classes

Target Attribute is a categorical variable with discrete values

- Binary Classification: loan default predictor (identify bad risk customers) and churn detection
- Multi-Class Classification: risk evaluation of Covid-19 vaccinations in elderlies

Algorithms:

- K-Nearest Neighbour
- Decision Trees (ID3, C4.5, C5.0)
- Logistic Regression
- Support Vector Machines (SVM)
- Naive Bayes
- Linear Discriminant Analysis
- Neural Networks

K Nearest Neighbours

A method for classifying cases based on their similarity to other cases in the neighbourhood

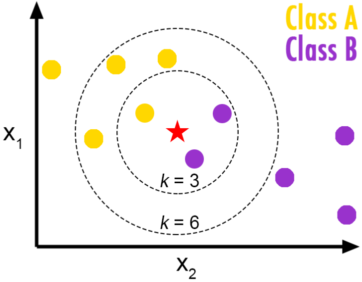
Based on an assumption that similar cases with same class labels are near each other

Algorithm Flow:

- Pick a value of K
- Calculate the distance of unknown case from all cases
- Search for K observations in the training data that are nearest to the unknown data point
- Predict the response of the unknown data point using the most popular response value from the K nearest neighbours

What Bothers Us?

- How to select the correct K?
- How to calculate the distance of unknown case from all cases?

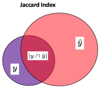


Way Out:

- Train and Test the Model on a range of K values and go with the best K value
- Assume Minkoski space and calculate distances in the multi-dimensions

Evaluation Metrics in Classification

Jaccard Index:
Classifier close to 1.0 has better accuracy



$$J(y,\hat{y}) = \frac{|y \cap \hat{y}|}{|y \cup \hat{y}|} = \frac{|y \cap \hat{y}|}{|y| + |\hat{y}| - |y \cap \hat{y}|}$$

$$y : [0,0,0,0,0,1,1,1,1,1]$$

$$\hat{y} : [1,1,0,0,0,1,1,1,1,1]$$

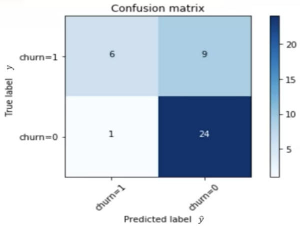
$$J(y,\hat{y}) = \frac{8}{10+10-8} = 0.66$$

CONFUSUSION MATRIX	ACTUAL	
PREDICTED	True Positive (TP)	False Positive (FP)
	False Negative (FN)	True Negative (TN)

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

$$\text{F1-Score} = \frac{2*Precision*Recall}{Precision+Recall} \qquad \text{Recall} = \frac{TP}{TP+FN}$$

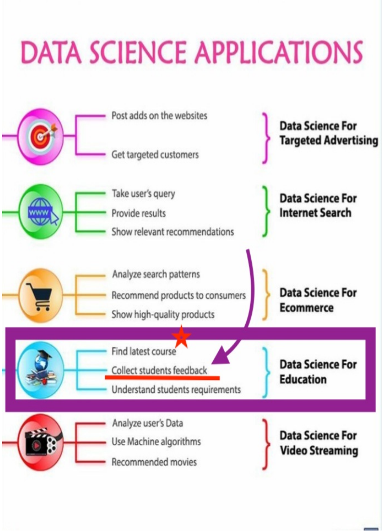
Confusion Matrix:
F1-Score is the harmonic average of precision and recall.



Log Loss:
Classifier with lower Log Loss has better accuracy

$$\text{Log Loss} = -\frac{1}{n}\sum y \log \hat{y} + (1-y)\log(1-\hat{y})$$

Your Feedback Matters!!!



THANK YOU ALL FOR ACTIVELY ENGAGING IN THE WORKSHOP

FOR ASKING QUESTIONS AND GIVING VALUABLE SUGGESTIONS

TOGETHER WE EVOLVE