**Python ??**

• **Inventor- Van Rossum** • **General purpose high level language** • **Evolved since past 25 years** • **Clear readable syntax** • **Good practice required – clean readable code**

• **Dynamic language – everything at runtime**

• **Interpreted language – quick iteration and test**

• **Lot less code** • **Portable**

• **Code security** • **Slow**

• **Second most popular language**

• **Instagram,** • **Netflix** • **Google** • **Spotify** • **Uber**

• **OpenShift** • **Dropbox**

1

**Applications**

▪ **Fraud detection.**

▪ **Web search results.**

▪ **Real-time ads on web pages** ▪ **Credit scoring and offers.** ▪ **Prediction of equipment failures.**

▪ **New pricing models.**

▪ **Network intrusion detection.**

▪ **Recommendation Engines** ▪ **Customer Segmentation** ▪ **Text Sentiment Analysis** ▪ **Predicting Customer Churn** ▪ **Pattern and image**

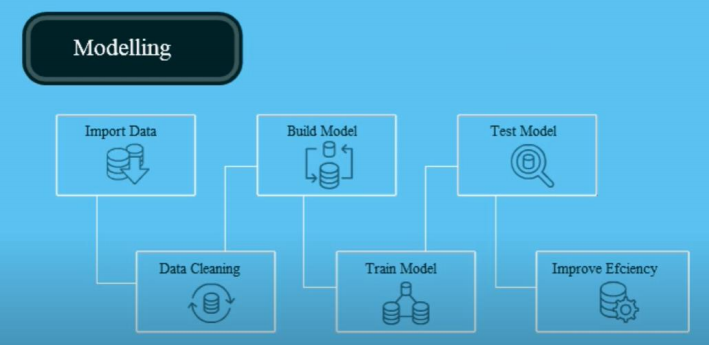
**recognition.**

▪ **Email spam filtering.** ▪ **Financial Modeling**

2

DS vs AI vs ML vs DL

3

5

Machine Learning

• Modelling uses machine learning algorithms, in which the machine learns from the data just like humans learn from their experiences.

• Speech Recognition, Image recognition, AI Camera on Phone, Healthcare, Insurance, Customer Churn, Customer Default • Derived from data - derivatives

6

3 broad categories of machine learning

• **Regression**: The output variable to be predicted is a **continuous variable**, e.g. scores of a student

• **Classification**: The output variable to be predicted is a **categorical variable**, e.g. classifying incoming emails as spam or ham

• **Clustering**: **No pre-defined notion of label** allocated to groups/clusters formed, e.g. customer segmentation

•

7

Supervised / Unsupervised

• **Supervised learning methods**

– Past data with labels is used for building the model – **Regression** and **classification** algorithms fall under this category • **Unsupervised learning methods**

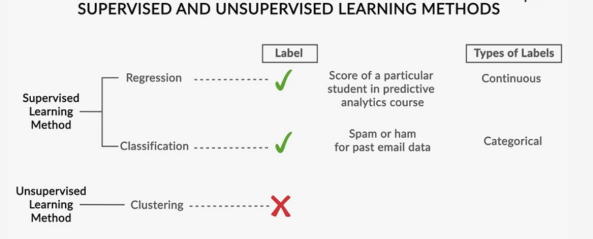
– No pre-defined labels are assigned to past data

– **Clustering** algorithms fall under this category

– PCA

8

Illustration

9

**Evaluating Supervised Learning**

▪ **Split data in two parts: a training set and a test set.** ▪ **Train the model using only the training set and then measure (using r-squared, Classification accuracy, Logarithmic loss, confusion matrix, AUC, F1 score, MSE, MAE) the model’s accuracy by asking it to predict values for the test set, and compare that to the known, true values.**

10

**Train / Test in practice**

▪ **Need to ensure both sets are large enough to contain representatives of all the variations and outliers in the data you care about**

▪ **The data sets must be selected randomly** ▪ **Train/test is a great way to guard against overfitting**

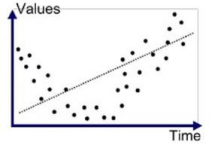
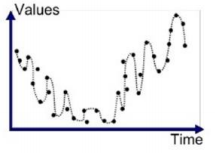
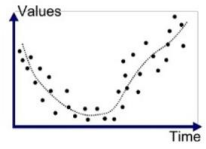
11

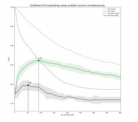
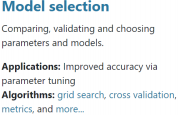
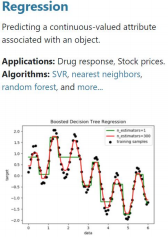
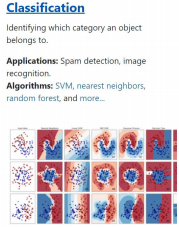
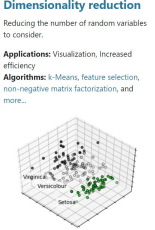
**Train/Test is not Infallible**

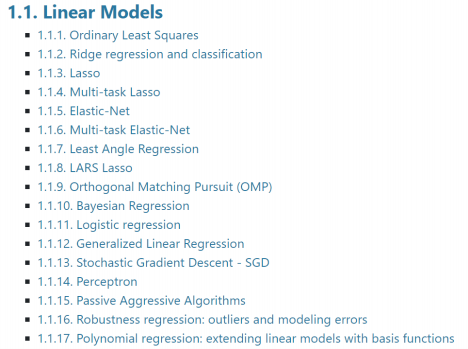
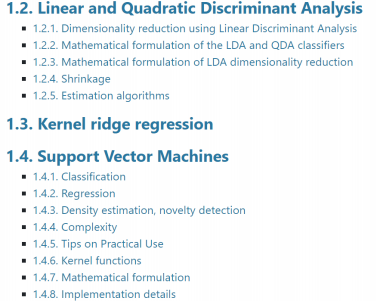
▪ **Sample sizes are too small**

▪ **Or due to random chance that train and test sets look remarkably similar**

▪ **Overfitting / Underfitting can still happen**

****12

13

14

15

16

Linear Regression

• R2 / RSS / TSS



17

Linear Regression

18

Linear Regression

19

Equation



20

Linear Regression



21

Model Steps







22

Models

• Logistic Regression

• Naïve Bayes

• K Means Clustering

• Hierarchical Clustering / K Mode Clustering / DB Scan Clustering • SVM

• Tree Model

• Model Selection

• Boosting

23

NLP / Deep Learning

• Lexical Processing

• Syntatic Processing

• Semantic Processing

• Chatbots

• Neural Network

• CNN

• RNN

24

Neural Network



25

Neural Network

26

Character Recognition

27

Correlation

• Measures the relative strength of the *linear* relationship between two variables

• Unit-less

• Ranges between –1 and 1

• The closer to –1, the stronger the negative linear relationship • The closer to 1, the stronger the positive linear relationship • The closer to 0, the weaker any positive linear relationship

Scatter Plots of Data with Various Correlation Coefficients

**Y**

**X**

**Y**

**X**

**Y**

**X**

r = -1 r = -.6 r = 0

**Y**

**X**

**Y**

**X**

**Y**

**X**

r = +1 r = +.3

r = 0

Linear Correlation

Linear relationships Curvilinear relationships Y

Y

Y

X

Y

X

X X

Linear Correlation

Strong relationships Weak relationships Y

Y

X

X

Y

Y

X

X

Linear Correlation

No relationship

Y

X

Y

X

33

**Lighter Topic**

34

**Skill Mappings**

36

**Data Science Insurance** 

**Perspective**

37

**Unsupervised Learning**

▪ **Unsupervised learning is used against data that has no historical labels.**

▪ **The goal is to explore the data and find some structure within.**

38

**Reinforcement Learning**

▪ **Reinforcement learning is often used for robotics, gaming and navigation.**

▪ **With reinforcement learning, the algorithm discovers through trial and error which actions yield the greatest rewards.**

39

**Reinforcement Learning**

▪ **You have some sort of agent that “explores” some space** ▪ **Learns the value of different state changes in different conditions** ▪ **Those values inform subsequent behavior of the agent** ▪ **Yields fast on-line performance once the space has been explored**

****40

**Regression**

▪ **Relationship between a scalar dependent variable y and one or more** 

**explanatory variables (or independent variables)** ▪ **Commonly used for predictive analysis**

41

**Regression**

****42

**How this works** ▪ **Usually : “least squares”** ▪ **Minimizes the squared** 

**error between each point and the line**

▪ **The slope is the correlation between the two variables** ▪ **This is the same as**

**maximizing the likelihood of the observed data**

43

http://localhost:8890/notebooks/Desktop/COE\_29th%20March%202017/Data%20science/Python/ML\_De mos/LinearRegression.ipynb

44

**Multivariate regression (Multiple Regression)**

▪ **What if more than one variable** 

**influences**

▪ **Example: predicting a price for a**

**car based on its many attributes**

**(body style, brand, mileage, etc.)**

45

46

**K-Means Clustering**

▪ **Attempts to split data into K groups that are closest to K centroids**

▪ **Unsupervised learning – uses only the positions of each data point**

▪ **Can uncover interesting groupings of people/ things / behavior**

**Example: Where do millionaires live?**

▪ **What genres of music / movies / Cars ?**

▪ **Create your own stereotypes from demographic data**

47

http://localhost:8890/notebooks/Desktop/COE\_29th%20March%202017/Data%20science/Python/ML\_De mos/KMeans.ipynb

48

**Bayesian Methods**

****49

**What about all the other words?**

****50

**Classification problem**

****

51

http://localhost:8890/notebooks/Desktop/COE\_29th%20March%202017/Data%20science/Python/ML\_De mos/NaiveBayes.ipynb

52

**Recommender Systems**

****53

**User-Based Collaborative Filtering**

▪ **Build a matrix of things each user bought/viewed/rated** ▪ **Compute similarity scores between users**

▪ **Find users similar to you**

▪ **Recommend stuff they bought/viewed/rated that you haven’t yet.**

54

**User-Based Collaborative Filtering**55

**User-Based Collaborative Filtering**

56

**Problems with User-Based CF**

▪ **People are fickle; tastes change** ▪ **There are usually many more people than things**

▪ **People do bad things odd times**

57

**Item-Based Collaborative Filtering**

▪ **A movie will always be the same movie – Does it change?**

▪ **There are usually fewer things than people** ▪ **Harder to game the system**

58

**Item-Based Collaborative Filtering**

▪ **Find every pair of movies that were watched by the same person**

▪ **Measure the similarity of their ratings across all users who watched both**

▪ **Sort by movie, then by similarity strength**

59

**Item-Based Collaborative Filtering**

****

60

**Item-Based Collaborative Filtering**

61

**Item-Based Collaborative Filtering**

62

**Item-Based Collaborative Filtering**

63

**K-Nearest Neighbor (KNN)**

▪ **Used to classify new data points based on “distance” to known data**

▪ **Find the K nearest neighbors, based on your distance metric**

64

http://localhost:8890/notebooks/Desktop/COE\_29th%20March%202017/Data%20science/Python/ML\_De mos/SimilarMovies.ipynb

65

Back up slides

66

67