Data Science

TERMINAL LEARNING OBJECTIVE

Students will understand the fundamentals of data science

Given classroom environment, instructor assistance, workstation, physical and virtual ranges, and access to the training materials, students will be able to do what is described in the terminal learning objective.

Agenda

Friday 23JUL21

- LSA1: Understand data science.
- LSA2: Understand the components of Data Science statistics.
- LSA3: Understand Data Processing in Data Science.
- LSA5: Understand Data Science Metrics and Techniques.
- LSA6: Understand the concepts associated with Data Analysis.
- LSA7: Understand the concepts associated with Data Visualization.

LSA1: Understanding Data Science

Outcome: At the end of this learning step, students will be able to define data science.

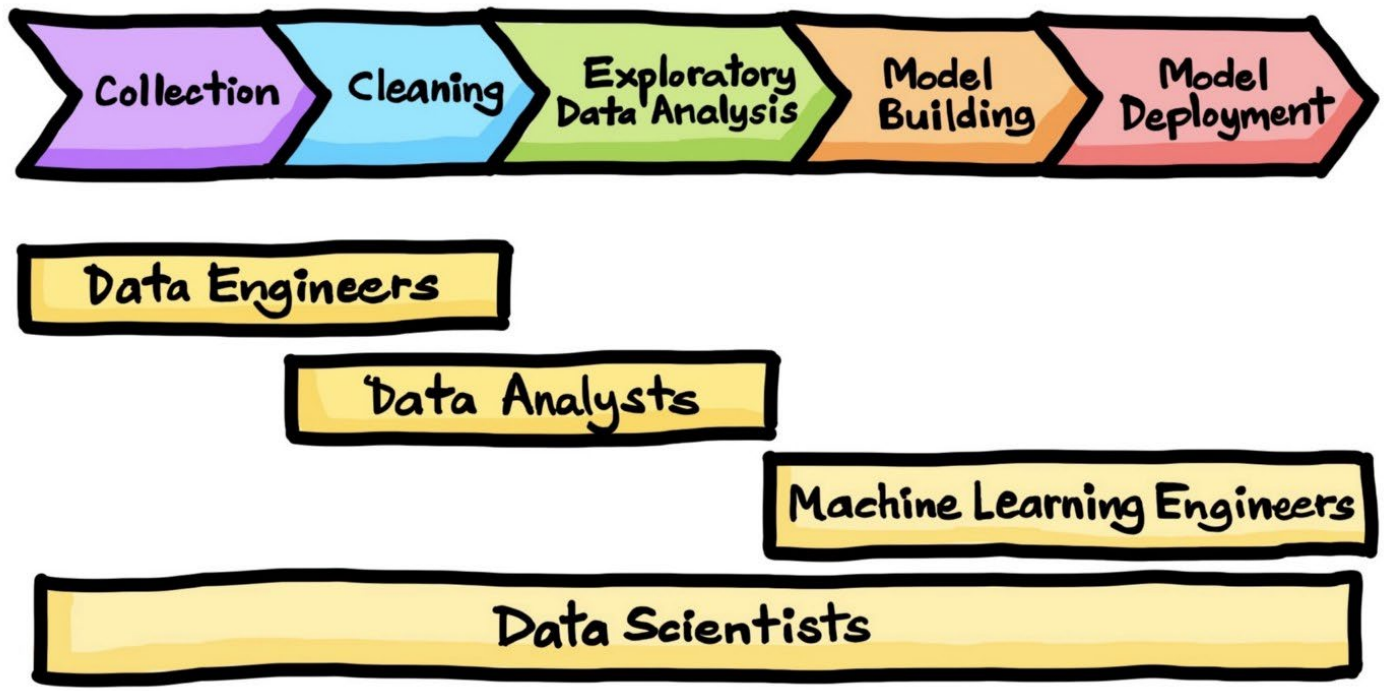
What is data science?

Data Science is the study of the generalizable extraction of knowledge from data.[1]

1. Dhar, V. (2013). Data science and prediction. Communications of the ACM, 56(12), 64-73.


DATA SCIENCE
LANDSCAPE
Software Engineering
Parallel Computing
Optimize Code
Software Development Best Practices
Tidy Code
Data Structure
Model Deployment
Web Development
Data Pre-processing
Handling Missing Data
Data Cleaning
Feature Engineering
Obtaining Data
Feature Selection
Mathematics
Discrete Mathematics
Matrices
Linear Algebra
Optimization
Probability Theory
Geometry
Calculus
Real analysis
Statistics
Inferential Statistics
Hypothesis Testing
Experimental Design
Descriptive Statistics
Programming
Scala
Julia
C/C++
SQL
Spark
Hadoop
R
Python
Bash
Java
Data Visualization
Exploratory Data Analysis
Types
Comparison
Composition
Relationship
Distribution
Machine Learning
Back-Propagation
GAN
Deep Learning
CNN
Neural Network
Support Vector Machine
Random Forest
Trees
XG Boost
Decision Trees
Algorithm
Classification
Regression
Supervised Learning
Principal Component Analysis
K-means
Clustering
Unsupervised Learning
Soft Skills
Lifetime Learning
Writing
Curiosity
Storytelling
Problem Solving
Presentation
Domain Knowledge
Grit
Creativity
Communication
Critical Thinking
BY: CHANIN NANTASENAMAT
DATA PROFESSOR
http://youtube.com/dataprofessor
FEBRUARY 14, 2020

THE DATA SCIENCE PROCESS



(Drawn by Chanin Nantasenamat in collaboration with Ken Jee)

Data Science Applications

- Voice Recognition
  - Siri, Google Voice, Alexa
- Fraud detection
- Medical Image Analysis
- Sports
  - Find undervalued players (Think Money Ball) Liverpool did this an won the 2019 Premier League championship
  - Determine the best positions on the field or court
- E-Commerce
  - Automated Ad placement
  - Product Recommendation
- Social Media
  - LinkedIn recommend connections
  - Tinder recommends Matches
  - Facebook presents potential connections

Voice Recognition

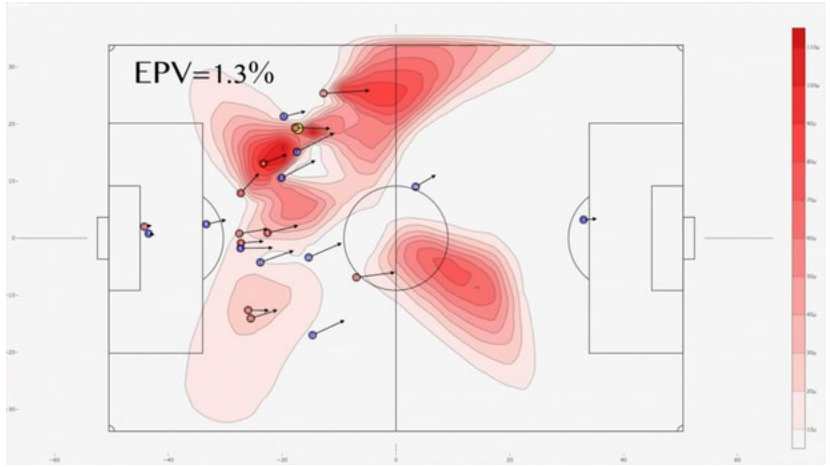
Video recommendation systems

E-Commerce Applications

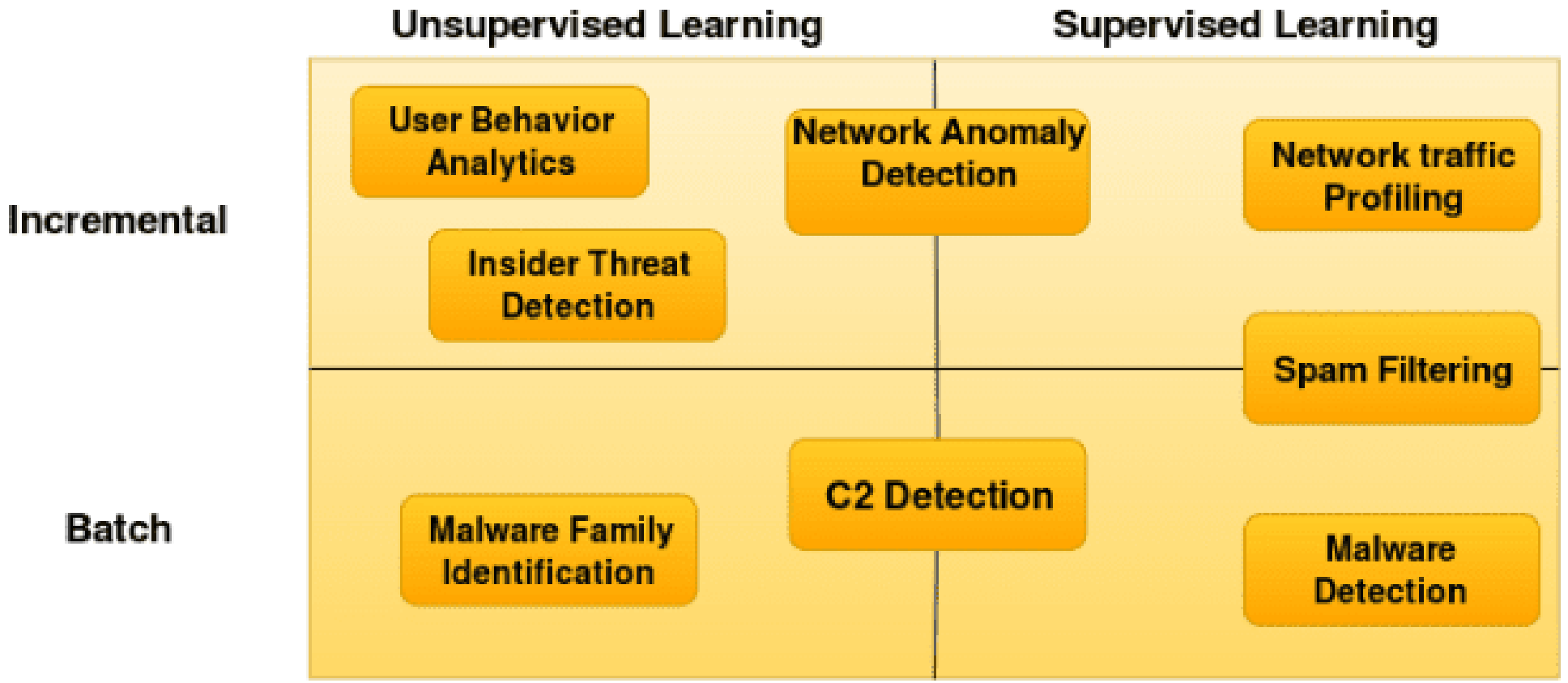
- Product Recommendations
- Personalized Advertisement

Sports – Liverpool 2019

- Find undervalued players (2019 Liverpool)
- Determine best strategy on the field that increases goal probability
- Won 2019-2020 Premier League title



Cyber Security Applications



https://www.blueteamsacademy.com/ml-ids/

Conclusions

- Data Science spans multiple disciplines and seeks to find actionable insights from data
- Applications exist in every domain

Questions

LSA2: Understand the components of Data Science statistics.

Outcome: At the end of this learning step, students will be able to identify and define the components of data science statistics.

Central Tendency

- Used to get a simple understanding of a feature or variable
- Informs how biased a dataset is.
- Utilize Mean and/or Median to determine central tendency
- Median is more robust to outliers

Mean – average value

$$Mean = \frac{Sum\ of\ all\ data\ values}{Number\ of\ data\ values}$$

$$\bar{x} = \frac{\sum x}{n}$$

$$\bar{x} = \frac{1 + 2 + 3 + 4 + 5}{5} = \frac{15}{5} = 3$$

```
import statistics as stats
array = [1, 2, 3, 4, 5]
stats.mean(example_array)
```

Median – middle value

Odd list length:

- Median is $\frac{n+1}{2}$ element in ordered list.

Even list length:

- Median is the average of the $\frac{n}{2}$ and $\frac{n+1}{2}$ elements in ordered list.

```
import statistics as stats

array = [1, 2, 3, 3, 4, 5]

stats.median(example_array)

median = 3
```

MODE

The most frequent element.

```
import statistics as stats

array = [1, 2, 3, 3, 4, 5]

stats.mode(example_array)

mode = 3
```

Quartiles

- Three values that split the provided data into four equal parts (25%, 50%, 75%).
- Gives us insight to if a value is an outlier or not (Typically in the top/bottom 5%).

```
import statistics as stats

array = [1, 2, 3, 4, 9, 5]

stats.quantiles(array, n=4)=

[1.75, 3.5, 6.0]
```

Variance

- Measurement to determine how dispersed individual elements are from the mean.
- Captures variability of elements from the mean

$$\sigma^2 = \frac{\sum_{i=1}^{n} X_i - \mu}{n}$$

- σ^2 = Variance
- X_i = Element *i*
- μ = population mean

```
import statistics as stats

array = [1, 2, 3, 4, 9, 5]

stats.variance(array)=
8
```

Standard Deviation

- Measurement to determine how dispersed individual elements are relative to the mean.
- Square root of the Variance

$$\sigma = \sqrt{\frac{\sum_{i=1}^{n} X_i - \mu}{n}}$$

- σ = Standard Deviation
- X_i = Element *i*
- μ = population mean
- n = number of elements

```
import statistics as stats

array = [1, 2, 3, 4, 9, 5]

stats.stdev(array)= 2.82843
```

Covariance

- Measures the relationship between two variables.
- Does not measure dependency between two variables.
- Positive Covariance
  - Two variables go in same direction
- Negative Covariance
  - Two variables go in opposite directions

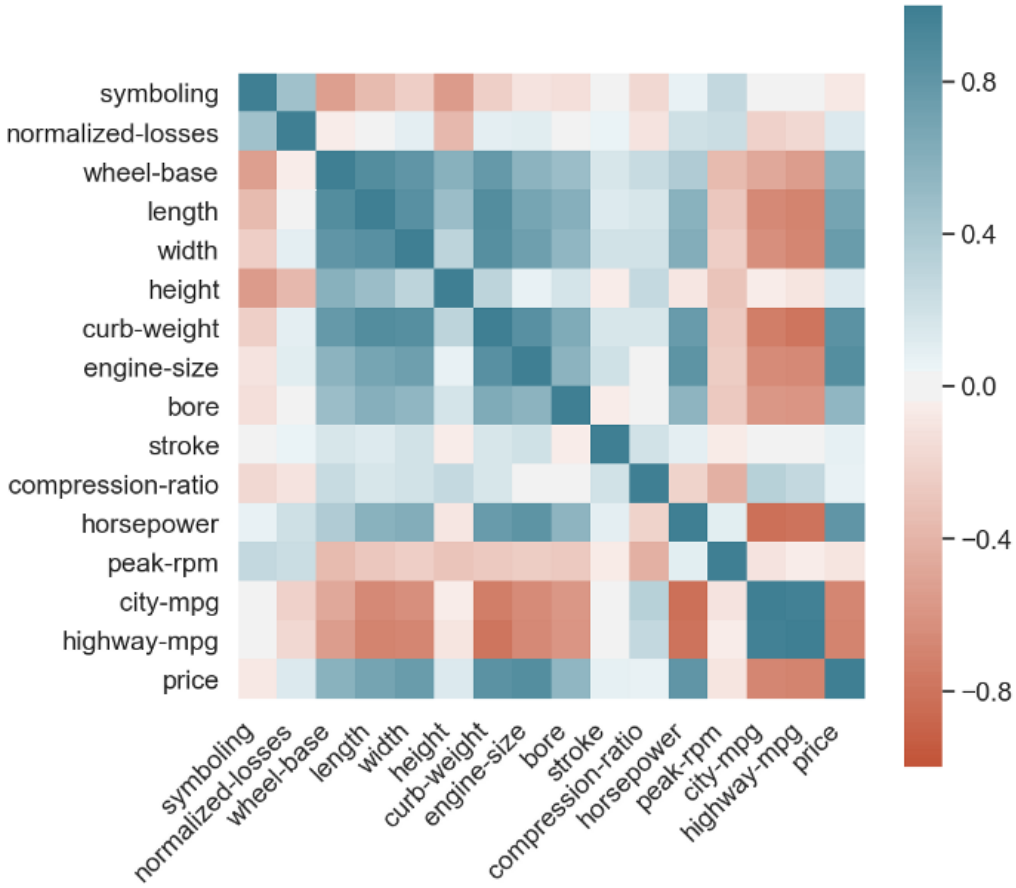
$$COV(X,Y) = \frac{\sum(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

```
from numpy import cov
array1 = [1, 2, 3, 4, 9, 5]
array2 = [1, 2, 3, 4, 9, 5]
cov(array1, array2)= 1
```

- $COV(X,Y) =$ covariance of *X* and *Y*
- X_i = the values of *X*
- $\bar{X}$= the sample mean of *X*
- Y_i = the values of *Y*
- $\bar{Y}$= the sample mean of *Y*
- n = number of elements

Correlation

- Normalized covariance (-1.0, 1.0)
- If the correlation between two variable is 1, then the change in variable one results in a proportional move in the same direction for variable 2.
- Useful for Principal Component Analysis (PCA) or other dimensionality reduction techniques.

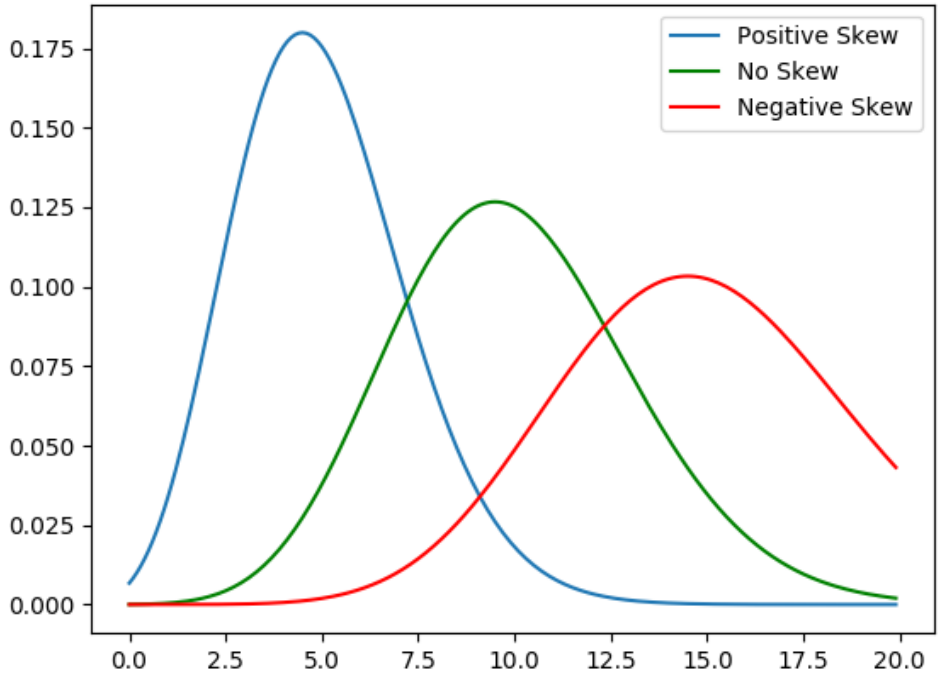


Skewness

- Measures asymmetry
- Positive value means skewed to the left.
- Negative value means skewed to the right

$$skewness = \frac{3x(Mean\ -Median)}{Standard\ Deviation}$$

- Provides insight to if the distribution is Gaussian
- scipy.stats.skew(array)



LSA3: Understand Data Processing in Data Science.

Outcome: At the end of this learning step, students will be able to identify and define data processing techniques.

What is Data Preprocessing/processing?

A data science procedure in which data transforms into a clean, machine-interpretable format.

Encoding schemes

- One hot encoding: Convert label encoding to a non-ordering relationship variable
- Binarization: Convert variable into 0s and 1s based on a fixed threshold.
- Discretization: Split a continuous variable into categories or groups
  - Equal Width, Equal Frequency, K means

Normalization / Rescale

Normalization gives equal weight/importance to each variable

- (Rescaling) Min-Max normalization
- Mean normalization
- Z-score normalization

Processing values

Missing Values

- Remove entries with missing values
- Impute values for those missing

Duplicate Values

- Remove Duplicates

Inconsistent Values

- Create consistent format (Date, Address, Phone number, etc.)

- Outliers
  - Remove

Why do we need to Process Data?

- Missing, Duplicate, Inconsistent values
- Outliers
- Feature Enrichment
- Feature Encoding
- Dimensionality Reduction
- Imbalanced data

One Hot Encoding

Car Maker	Categorical Value	Price
Passat	1	12000
Civic	2	10000
Accord	3	14000
Accord	3	15000

Passat	Civic	Accord	Price
1	0	0	1200
0	1	0	10000
0	0	1	14000
0	0	1	15000

Label Encoding

One-hot Encoding

Binarization

Passat	Civic	Accord	Price
1	0	0	1200
0	1	0	1000
0	0	1	14000
0	0	1	15000

Passat	Civic	Accord	Price
1	0	0	0
0	1	0	0
0	0	1	1
0	0	1	1

Discretization

- Transforming continuous data into discrete data (categories or bins)
- Height is a continuous variable that can be transformed into short, average, tall bins
  - Short (less than 60 in)
  - Average (between 60-71 inches)
  - Tall (greater than 71 inches)

Why Transformations?

- Utilized to make your data work better for you
- Most raw data is skewed, but some models (linear and logistic regression) follow normal distributions.
- Map a skewed distribution to a normal distribution through a transform

Log Transform

- Attempt to get distribution close to a normal distribution
- Not applied to features with negative values
- Applied for skewed-right data.
- Scale changes to multiplicative scale (linearly distributed data)

Square Root Transform

- Attempt to get distribution close to a normal distribution
- Compresses high values, so lower values are more spread out.
- Could apply this to the dependent variable, but results become less interpretable.
- Useful for right skewed distributions.

Square Transform

- Attempt to get distribution close to a normal distribution
- Useful for left skewed distributions.

Feature enrichment

The addition and/or transformation of features to increase the variance within the dataset to produce more robust models.

- Supplement data with data from other sources
- Feature Selection
- Feature Transformation
- Data Imputation

Questions

LSA5: Understand Data Science Metrics and Techniques.

Outcome: At the end of this learning step, students will be able to identify, define, and apply data science metrics and techniques.

Agenda

- Cross Validation
- Grid Search / parameter tuning
- Metrics
  - Loss Functions
  - Area under the curve (AUC)
  - Type 1 error and Type 2 error
  - Confusion matrix
  - Validation Accuracy

Loss Functions

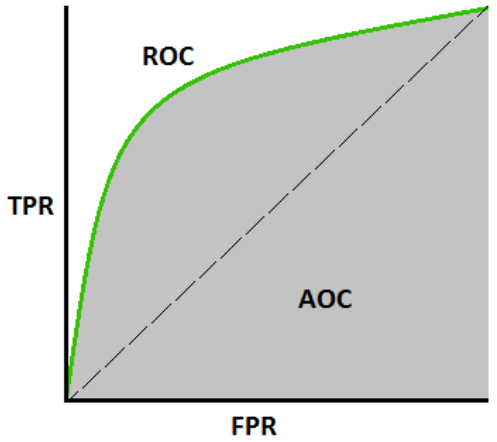
- Machine learning is algorithmic and relies on loss functions to achieve expected performance
- Regression Losses
  - Mean Squared Error (L2 loss)
  - Mean Absolute Error (L1 loss)
  - Huber
- Categorical Losses
  - (Cross-Entropy) Negative log likelihood
  - Hinge

AREA UNDER THE CURVE (AUC)

- Higher AUC = Better prediction performance
- Utilized for Binary Classification
- TPR = True positive rate
- FPR = False positive rate

$$\text{TPR /Recall / Sensitivity} = \frac{TP}{TP + FN}$$

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



Type 1 vs Type 2 Error

- Type 1 (false positive):
  - Test says you have COVID, but you don’t
- Type 2 (false negative):
  - Test says you don’t have COVID, but you do

	Actual -- True/False	
Predicted -- Positive/Negative	True Positive	False Positive (Type I)
	False Negative (Type II)	True Negative

Confusion Matrix

- Accuracy (all **correct** / all) = TP + TN / TP + TN + FP + FN
- Misclassification (all **incorrect** / all) = FP + FN / TP + TN + FP + FN
- Precision (**true** positives / **predicted** positives) = TP / TP + FP
- Sensitivity aka Recall (**true** positives / all **actual** positives) = TP / TP + FN
- Specificity (**true** negatives / all **actual** negatives) =TN / TN + FP

Confusion Matrix

		Actual	
		Pregnant	**Not**
Predicted	**Pregnant**	45 **TP**	55 **FP**
	Not	5 **FN**	395 **TN**

Accuracy	88%
Misclassification	12%
Precision	45%
Sensitivity	90%
Specificity	88%

Parameter Tuning

- There are several parameters that can hold different values
- We must decide which values are best for our model.
- One method is grid search:
  - Take all combinations of desired parameters and test each one. Select the parameter combination that yields the highest performance value for the selected performance metric.

Cross-Validation

- Split data into k-splits.
- Hold out one split as the test data and use the remaining k-1 splits as test data.
- Record the performance metric for model
- Repeat k times
- Less variance in performance across k models results in higher validation of model

Questions

LSA6: Understand the concepts associated with Data Analysis.

Outcome: At the end of this learning step, students will understand data analysis concepts.

Exploratory Data Analysis (EDA)

- Investigate data sets and summarize their main characteristics
- Utilize statistics and data visualization methods
- Guides how to manipulate the data to best answer the problem you are addressing
- Detect outliers or anomalous events before modeling

Why do EDA?

- Inform assumptions
- Identify obvious patterns, outliers, and variable relationships
- Confirms you have the right data you need for the problem
- Identify models not correct for the problem

Four types of EDA

- Univariate non-graphical
- Univariate graphical
- Multivariate nongraphical
- Multivariate graphical

Univariate non-graphical EDA

- Examine one variable
- Does not look at correlation or relationships
- Describes the data using mean, median, min, max, quartiles, standard deviation.

Univariate Graphical EDA

- Examine one variable
- Does not look at correlation or relationships
- Histogram
- Box plot

Multivariate non-graphical EDA

- Examine multiple variables
- Looks at correlation or relationships
- Typically, correlation and covariance.

Multivariate Graphical EDA

- Examine multiple variables
- Utilizes graphics to display relationships between multiple variables or data sets
- Scatter Plot
- Multivariate bar plot or histogram
- Heat map

LSA7: Understand the concepts associated with Data Visualization.

Outcome: At the end of this learning step, students will identify data visualization concepts and create data visualizations.

Agenda

- Scatter Plot
- Line Plot
- Histogram
- Bar Plot
- Contour Plot
- Facets

What is a graphic?

Common statistical plots:

- Scatter plot
- Line plot
- Box plot
- Histogram
- Bar plot

What are the common elements of a graphic?

- Some type of data
- Maps data to aspects of the plot
- Geometric in nature
- Statistical transformation
- Coordinate systems

How are these plots Related?

- Scatter Plot
  - Maps variables to x-axis and y-axis
  - Uses points to represent each observation
- Line Plot
  - Maps variables to x-axis and y-axis
  - Uses lines to connect each observation
- Histogram
  - Maps bins to x-axis and frequencies to y-axis
  - Uses bars to represent bin frequency
- Bar Plot
  - Maps categorical variable to x-axis and counts to y-axis
  - Uses bars to represent category frequency
- Box Plot
  - Maps 5-number summary (min, lower-hinge, median, upper-hinge, max) to y-axis
  - Uses shapes to represent values (box and whiskers)

How are these plots Related?

- Contour Plot
  - Maps three-dimensional surface to a two-dimensional surface
  - Maps variables to x-axis and y-axis
  - Uses contour lines to connect points of same response value
  - Color bands between contours to represent range of values
- Facets
  - Matrix of two-dimensional graphs partitioned by a third variable

Why use visuals?

- Easier to interpret
- Layers allow for flexibility in the amount of information displayed
- Wow factor for customers

Python packages for Visualization

- Matplotlib
- Seaborn
- Plotly

Application for Data Visualization

- Exploratory Data Analysis
- Present results
- Interactive user application

LSA4: Demonstrate the ability to solve statistical and data processing problems.

Outcome: At the end of this learning step, students will solve statistical and data processing problems.

LSA8: Demonstrate the ability use metrics, data analysis and visualization techniques to understand data

Outcome: At the end of this learning step, students will use metrics and visualization to analyze data.