<u>Machine Learning Concepts and Basics Notes</u>

Linear Regression:

- $y \approx B_0*x_0+ B_1*x_1 = y_p$
    - B_i : unknown constants
    - X_i : independent variable
    - y: dependent variable

Not all x_i can be mapped to y_i perfectly there will almost always be an amount of deviation, hence the use of "≈"(almost equal to) symbol.

- The best fit trend line is an educated guess regarding what the linear relationship between x and y are.
    - R^2 value/correlation measures how related x and y are, closer to 1 the better.
- Sum of the squared errors (SSE) is the sum of the differences between y and predicted y_p
- T-test is used to find the correlation between 2 dependent variables.

Logistical Regression:

- Classification model, primarily binary
    - Find relation between predictor variables and categorical variables
- Subtypes:
    - Binary log: Yes/No decisions
    - Nominal log: 3+ categories with not natural ordering to levels i.e. search engines, colors, letters
    - Ordinal log: 3+ categories with a natural ordering to levels i.e. effectiveness of course, severity of medical condition, restaurant rating
- Metrics:
    - Wald Test: tests significance of individual coefficients (similar to t-test)
    - Goodness-of-fit-tests: used to find if observed values match expected value under model
        - Most common used: Chi-squared method

Decision Trees:

- Decision trees are non-parametric supervised learning method used for classification and regression
    - Non-parametric: a statistical method in which the data are not assumed to come from prescribed models that are determined by a small number of parameters
- Random Forest Classifiers: each tree votes for a value and the result that has the most votes is chosen as the predicted value.
    - Primary concept of random forest and other forest type algorithms is:
        - A large number of uncorrelated models(trees) operating as a committee will outperform any individual constituent models
        - Low correlation between trees is important since each tree's error is covered by its neighbors
- Random Forest Regression: operates almost the same way as classifiers except all the results given by each tree are averaged together to generate a single value.

**For more workshops and information visit go.tamiu.edu/arc*