

Welcome

Prof. Eric A. Suess

January 20, 2021

Welcome to Stat. 652 Statistical Learning

This course will be about Data Science and the use of Statistical Learning/Machine Learning/Artificial Intelligence to analyze data.

Our focus will be on using the setup of Machine Learning with *training data*, *test data*, and *validation data*.

We will discuss *accuracy* using *confusion tables*.

Welcome to Stat. 652 Statistical Learning

We will also be using modern R packages to fit the models we will be learning about.

We will be using h2O.ai and h2O Driverless AI.

We will be using Tensorflow for R and keras for R.

We may discuss Spark for R.

Statistical Topics

- ▶ Sampling from a population
- ▶ Sample statistics
- ▶ Sampling distributions
- ▶ Bootstrapping
- ▶ Outliers
- ▶ Linear Regression - Prediction
- ▶ Confounding Variables
- ▶ Logistic Regression - Classification
- ▶ Problems with p -values

Statistical Learning, Machine Learning, Artificial Intelligence, and Predictive Analytics

- ▶ Supervised Learning
- ▶ Unsupervised Learning

Supervised Learning

Classifiers and Regression/Prediction

- ▶ Decision Trees
- ▶ Random Forests
- ▶ Boosting
- ▶ Bagging
- ▶ k-Nearest Neighbors (kNN)
- ▶ Naive Bayes
- ▶ Artificial Neural Networks (ANN)
- ▶ Deep Learning
- ▶ Ensemble Methods
- ▶ Forecasting

Evaluation Models

- ▶ Training, Validation, Testing Data
- ▶ Cross-validation
- ▶ Confusion Matrix
- ▶ ROC curves
- ▶ Bias-variance Trade-off
- ▶ Regularization

Unsupervised Learning

Clustering

- ▶ Hierarchical Clustering
- ▶ k-Means
- ▶ DBSCAN

Dimension Reduction

- ▶ Singular Value Decomposition (SVD)
- ▶ Principal Component Analysis (PCA)
- ▶ Factor Analysis
- ▶ Multidimensional Scaling (MDS)

Simulation

- ▶ Simulating variability

References

There are many many excellent references that will be useful for this class.

There are some references provided on the syllabus.

There will be many links given on the website.

My current favorite podcast about data science is DataFramed.