Welcome

Prof. Eric A. Suess

January 22, 2020

Today

- I will introduce the class
- Discuss the book(s) and homework
- Introduce the class website
- Begin the class with the material in Chapter 1 of the book

Introduction

In this class we will be learning about

- **Statistical Learning**
- **Statistical Machine Learning**
- **Machine Learning**
- **Predictive Analytics**
- **Big Data**
- **Artifical Intelligence**
- **Deep Learning**
- **Data Science**

"Modern Applied Statistics" using computers and lots of data.

Introduction

What is Statistical Learning? The answer depends on who you talk to.

The author of our book says things, such as,

"... machine learning provides a set of tools that use computers to transform data into actionable knowledge."

or

"... making sense of complex data."

or

Introduction

"... computer scientist Tom M. Mitchell states that a machine learns whenever it is able to utilize its experience such that its performance improves on similar experiences in the future."

Homework this week

Do an extensive Google search on the terms

- Statistical Learning
- Statistical Machine Learning
- Machine Learning
- etc.

Report on what these terms mean.

Do they mean the same thing?

Or are there differences?

Introduction to Learning with Data

Data has long been **expensive**. *Experimental* data needed to basically be collected by hand.

Recently (not really, but ok) data has become **very cheap**. *Observational* data is being collected automatically and stored in large amounts.

There are still opportunities in the traditional areas of data analysis, but there are new frontiers for people who have the skills to access, interact with, model and analyze, the large amounts of data that are already stored electronically.

Got to know something about **databases**!

Introduction to Machine Learning

What is the difference between **machine learning** and **data mining**?

The main difference is that machine learning is used to group similar observations based on important variables or to develop models for prediction.

While data mining is used to search for "hidden nuggets" in data.

However, much of the same "tools" are discussed in both areas.

Introduction to Machine Learning

The basic learning process:

- **Data input:** It utilizes observations, memory storage, and recall to provide a factual basis for further reasoning.
- **Abstraction:** Involves the translation of data into broader representations. Modeling.

$y = f(x_1, x_2, ...x_p)$

- **Generalization:** It uses abstracted data to form a basis for future action.
- **Evaluation:** Try to make improvements.

Introduction to Machine Learning

Assessing the success of learning:

Models are not perfect, but some are useful. (Is this the exact quote? Who said this?)

This is what is often discussed as checking the assumptions and/or validating the model.

Over fitting is a problem that needs to be avoided.

Introduction to Machine Learning

Steps to apply machine learning to your data:

1. *Collect* data
2. *Exploring* and *preparing* the data
3. *Training* a model on the data
4. *Evaluating* the model performance
5. *Improving* model performance

Introduction to Machine Learning

Selecting a machine learning algorithm:

Understand the data.

What are the **observations** or **unit of observation** or **examples**?

What are the **variables** or **features**?

Are the variables **numeric** or **categorical**?

Knowing your data can lead to a possible model and learning algorithm.

Introduction to Machine Learning

Types of machine learning algorithms:

1. **Predictive Models**, **Classification**, **Supervised Learning**
2. **Clustering**, **Segmentation Analysis**, **Unsupervised Learning**
3. **Descriptive Models**, **Patterns**

In Statistics we consider **Inference vs. Prediction**

Introduction to Machine Learning

See page 21 of our book for a table that presents the different types of algorithms and what the task is that they are used for.

The table includes the chapters of the book that covers each topic.

R

R is a very good platform for doing machine learning. There are many packages that have been written for implementing the algorithms. This is the platform that we are going to focus on for the course.

- Much of the **data science pipeline**, in many companies, is in Python, but R can be called from python and vice versa.

For Big Data see SparkR

Python

Python is another good platform for doing machine learning. The scikit-learn package has grown to be very useful for machine learning.

- Much of the **data science pipeline**, in many companies, is in Python.

Strengths and weaknesses

Both R and Python have there strengths and weaknesses.

There **was** a bit of a competition between the two.

It would be good to become familiar with both and good at one.

Summary

The author summarizes Chapter 1 with, “Machine Learning originated at the intersection of statistics, database science, and computer science. It is a powerful tool, capable of finding actionable insights in large quantities of data.”

Get better at using IT tools!!!