Chapter 1 Notes - STA 629

Anthony Bernardi

August 26, 2024

1 Introduction and Preamble

This is somewhat of an introduction to the course. There will be 5 homeworks and a final project, with the lowest homework grade dropped. There is a possibility that the final project will resemble a Master's Thesis, so this is something worth looking into. More specifically, it might be worth talking with ZW about this. This Final Project can also be analogous to something like a Kaggle Competition submission. Assignments are to be submitted via PDF, along with source and code files.

2 Chapter 1 - Introduction

Machine Learning can be thought of as building models and algorithms and using computers to perform specific tasks on given data. Think of this as an intersection between CS, Applied Math, and Stats. This is also an intersection with predictive modeling, namely, generating the most accurate prediction. We can think of the Applied Math piece as; Optimization and Numerical Analysis (e.g. Simpson's Rule for Integration) Mostly focused on predictive modeling rather than Gen AI. It ought to be noted that most (90 The common ground between CS, Math, and Stat; prediction, probability bounds, clustering, and graphical models.

A Quick Example: Spam Emails

Consider the following. X_i = email

$Y_i = 0$ if email (valid) and $Y_i = 1$ if spam.

In this problem, we'll have to use feature extraction, to try and get a sense of which phrases and terms are important in spam filtering in emails.

Ultimately, this is a classification problem (labeling an email 0 or 1.) Our metric is a simple misclassification rate. Numerically, these phrases and words get 'tokenized' and we see the distribution of each of these phrases and words, it gets coded into the model in this way.

Example 2: Regression

We will want to consult the book for this, but this is a standard regression problem in prediction.

Example 3: Handwriting Digit Recognition In this example from the textbook, we have a grid of grayscale maps used to predict a handwritten digit. What we are wondering is, how can the computer predict what is a 0, a 1, etc.?This information is stored in a manner analogous to RGB pixels. In the feature engineering for this problem, this information gets *vectorized* more specifically, marginal histograms are used.

Example 4: Microarray Gene Expression Data We'll now consider a slightly more involved example. $Y_i = TumorTypes\ X_i = GeneExpression"levels"$

This is thought of as a classic "large p, small n" problem in Machine Learning. The Y_i presence makes this a supervised learning problem, and Y_i can contain labels such as benign, malignant, etc. *Semi-Supervised Learning* is where we have a label but possibly some novel sub-types in prediction. A heatmap would be used in this problem case, as is provided in the book.

We will now consider the "Life Cycle of ML" which includes the following;

1. Model Deployment

2. Monitoring/Collecting Data

We will also want to consider that data can be renewed, and the model can subsequently perform worse, which requires our attention.

Supervised vs. Unsupervised Learning