Course Title: Statistics for Artificial Intelligence, Machine Learning, and Data Science: An Introduction

Course Code: STAT 05

Instructor: Greg Ryslik

Course Overview:

This course will provide a basic high-level introduction to the mathematics and statistics that underpin many of the modern machine learning and AI algorithms. This course will cover two broad areas of statistics: inference and prediction. The inference portion will introduce common statistical concepts that allow us to understand a population and test hypotheses (such as perform A/B tests, calculate and interpret p-values). The prediction section will begin with the simplest of algorithms (linear regression) and gradually touch upon more advanced topics such as random forests and cross validation. Real world examples will be used from the fields of healthcare, genetics, marketing and manufacturing.

Please note: This course has no specific prerequisites and can be taken on a variety of levels. Beginners are encouraged to listen to the lectures and learn basic concepts. By the end of the course, beginners should have a sense of what these algorithms do. On a higher level, intermediate students can work some of the introductory problems that will be provided. On an advanced level (for students with a substantial math background who are interested in becoming data scientists), difficult math and stats problems will be covered.

Students should be aware that this course will not make them an AI expert (this is not possible in nine weeks). On a basic level, the course will give students a taste of what statistics for AI is all about. On the highest level, the course will give students of a strong sense of what they need (and should be excited about) in order to pursue a career in this area.

The instructor recommends reading chapters 1-6 of "CK-12 Basic Probability and Statistics - A Short Course" before the course begins. The material is very basic and covers some things like the mean, median, standard deviation and what are independent events etc. Students can access this material here: https://www.ck12.org/book/CK-12-Basic-Probability-and-Statistics-A-Short-Course/

For students who are looking to work through some more in-depth problems throughout the course (on a medium/high level), they should take a look at the first few subsections of modules 4, 5, and 6 as a quick refresher:

https://courses.lumenlearning.com/introstats1/

Please contact the Stanford Continuing Studies office with any questions
365 Lasuen St., Stanford, CA 94305
continuingstudies@stanford.edu
650-725-2650
Grade Options and Requirements:

- No Grade Requested (NGR)
  - This is the default option. No work will be required; no credit shall be received; no proof of attendance can be provided.

- Credit/No Credit (CR/NC)
  - Students must attend at least 80% of class sessions.

*Please Note: If you require proof that you completed a Continuing Studies course for any reason (for example, employer reimbursement), you must choose the Credit/No Credit option. Courses taken for NGR will not appear on official transcripts or grade reports.*

Tentative Weekly Outline:

**Class 1**
An overview and introduction of statistics and why it’s especially important in the age of AI. The goal of this first class is to get people thinking about probability, statistics, and common misconceptions. We will work through some key topics such as sample space (via the Monty Hall problem) and some fundamental building blocks such as the Central Limit Theorem (how it applies). We will also cover the two main branches of statistics – prediction vs. inference, how they inter-relate and how they differ. This will be shown via real world examples.

**Class 2**
*Start of Inference.* We will cover what is the idea of inference, why we would want to perform inference on a population of interest and the basics behind a hypothesis test. This will be illustrated through examples based in genetics, politics and others. We’ll talk about p-values and how they should be interpreted and what they mean by “p-value hacking”.

**Class 3**
We’ll introduce the case of what happens when the data is categorical. Here we will cover some basic topics that appear everywhere in the health sciences (and gambling!) such as “odds” and “Odds-ratio”. We’ll hint at logistic regression that we’ll cover in class 6. We’ll show how to test for independence.
Class 4:
Beginning of the Prediction part of the class. Now that the students understand p-values, I’d like to introduce the topic of regression. While I’ll avoid a proof of how the estimates are created, I’ll go through a linear regression example and discuss how it is interpreted and what is meant by the Beta Values, the Std Errors, the p-values, etc. I’ll also talk about prediction and the dangers of making a prediction outside the range of the initial inputs.

Class 5:
I’ll begin to introduce more advanced prediction topics here such as cross-validation, R^2 and will briefly touch on logistic regression and why it’s useful. This will serve as a good gateway to classification and as a gateway to some of the more advanced AI algorithms.

Class 6:
This class will be used to wrap up some final topics in classical statistics and causality before we begin some of the more advanced machine learning topics. I’d like to introduce Bayes Theorem in the context of rare diseases as well as explore the topic of correlation vs causation. Finally, I’ll briefly mention what is a sample-size calculation and what does it mean to “power a study”.

Class 7:
I will introduce one of the easier to understand AI algorithms – Random Forests. I will explain how it is used in the context of classification and how it relates to topics such as prediction. I will link it to topics such as cross-validation and begin to close the loop between statistics and AI. We’ll briefly cover what do we do when we have “small data” – aka the opposite of big data.

Class 8:
This pre-ultimate course will be used as a demo of how statistics directly impacts machine learning. We’ll cover the basics of Artificial Neural Networks from first principles. We’ll also provide an introduction to convolutional neural nets and how they are used in computer vision. Finally, we’ll delve into how neural nets are fitted and how activation functions are used.

Class 9:
The final course will introduce Principle Components Analysis (PCA) and how it is used in dimension reduction. We’ll then provide a global summary of what the course has covered. We’ll refresh the difference between inference and prediction. We’ll then do a broad overview of some common AI algorithms. Finally, we will use the remainder of the class to answer any programming, statistics and general AI & Data Science questions.
Tentative Weekly Zoom Schedule:

Mondays, 5:30 PM – 7:00 PM (Pacific Standard Time)

Please contact the Stanford Continuing Studies office with any questions
365 Lasuen St., Stanford, CA 94305
continuingstudies@stanford.edu
650-725-2650