CS501 - Introdution to Data Science

Class Overview.

This course is an introdution to Data Science. Data Science is a continuously evolving discipline, that touches many areas of Computer Science and Mathematics. Qualified Data Scientists are in great demand and enjoy promising career paths.

This semester the class is offered as an online class. Students should consult the class web page regularly (at least daily). In addition I will schedule Zoom meetings as needed. If you are unable to attend the meetings, video/audio transcripts will be available.

In this introductory course we will take a somewhat simplified approach to the subject. We will view it a process with the following steps outlined below.

1. Problem specification.
2. Data acquisition.
3. Data cleaning and preparation.
4. Data exploration.
5. Data modeling.
6. Model testing.

It should be emphasized that each of the steps is a loop. At each step we may discover that previous steps need to be revisited.

Course Outline

The course will be organized around four major projects. Each project will require a new set of skills. So we will begin each project by developing the requisite skills. For each project, we will follow the outline above.

The first project will emphasize data cleaning and formatting and introduce a few basic classification methods. There will be a brief review of elementary statistical tests. The second project will focus on data presentation and Computer Graphics. There will be a review of Computer Graphics in Python. The third project will be a character recognition project that will require more advanced tools, including neural networks. The fourth project will use deep learning to develop a more sophisticated solution to a problem chosen by the student (subject to approval).

- Week 0 - Class Protocols, Course Objectives.
- Weeks 1-3, Project One: Author Identification
  - The Problem: Given a text document, identify certain parameters of the document. Examples may include:

    * Author identification
    * Classification by topic.
    * Sentiment analysis (e.g., deciding if a movie review is positive or negative)
  - Objectives:
    * Learn to use basic Linux command line tools to clean and organized text data.
    * Review elementary statistical ideas.
    * Learn a few basic classification methods, including
      · k-Nearest Neighbor
      · Naive Bayesian Classification
  - Week 1 - Using command line tools for Data Manipulation, Cleaning and Formatting
    * Regular expressions.
    * grep, sed, awk, etc.
  - Week 2 - Python review.
    * Regular expressions in Python.
    * Parsing a Project Gutenberg book.
    * Numpy.
  - Week 3 - Solution methods,
    * K-nearest Neighbor.
    * Naive Bayesian Classification.
    * Related problems: spam filtering, credit card fraud.

- Weeks 3-5, Project Two: Data Visualization
  - Objective: using Computer Graphics to understand the structure of datasets.
  - Project: Understanding Covid-19 data.
  - Tools:
    * Scatter plots
    * Line plots.
    * Bar, pie and doughnut charts.
    * Box plots.
    * Histograms.
    * Treemaps.
    * Heat maps.
    * 3-D plotting.
  - Techniques: classification versus regression.
    * Linear regression
    * Logistic regression.
    * Modeling gradient descent.

- Weeks 5-10: Project Three, Machine Learning

  - Project: Character recognition
  - Objectives:
    * Train a neural network to identify characters from images (jpegs, pngs, etc.).
  - Techniques
    * Learn to manipulate data frames in environments like *Numpy* and *R*.
    * Understand feed-forward networks.
    * Understand backpropagation and stochastic gradient descent.

- Weeks 11-15: Project Four: Deep learning
  - The Problem: chosen by the student, subject to constraints.
  - Techniques
    * Convolutional neural networks.
    * Support vector machines
    * Decision trees
    * Random forests
    * et cetera ...

Requirements.

- Students are expected to have completed CS202 and CS303.
- Students are expected to check the class web page http://cs.indstate.edu/cs501 regularly.
- Students are expected to complete any readings that appear on the `Links` section of the class web page.
- Students are to complete assignments that appear in the `Assignments` section of the class web page.
- Student are expected to complete assignments on time.

Goals.

Time: MWF 1:00

Room: A-017 Root Hall

Professor: G. Exoo

Email: `cs501@cs.indstate.edu`

Office: Online

Office Hours: TBA

Grading:

Class Participation	20%
Weekly assignments	70%
Final Exam	10%

Grading Scale:

92	-	100	A
90	-	91	A-
88	-	89	B+
82	-	87	B
80	-	81	B-
78	-	79	C+
72	-	77	C
70	-	71	C-
68	-	69	D+
60	-	67	D
0	-	59	F

Important Web Links

Class Web Page

CS501 on cs.indstate.edu

Academic Integrity Policy

Student Guide

It's On BLUE.

Indiana State University fosters a campus free of sexual misconduct including sexual harassment, sexual violence, intimate partner violence, and stalking and/or any form of sex or gender discrimination. If you disclose a potential violation of the sexual misconduct policy I will need to notify the Title IX Coordinator. Students who have experienced sexual misconduct are encouraged to contact confidential resources listed below. To make a report to the Title IX Coordinator, visit the Equal Opportunity and Title IX website.

Confidential Resources:

The ISU Student Counseling Center
HMSU 7th Floor
812-237-3939.

Campus Ministries

For more information on your rights and available resources.

All things related to Covid-19

The ISU Victim Advocate: Leah Reynolds

HMSU Room 813

812-237-3829 (office)

812-243-7272

`leah.reynolds@indstate.edu`