CIS*4020 Assignment 1

Regression and Clustering

Handed out: Oct 5th

Due date: Oct 26th

The objectives of this assignment include:

i. Gaining familiarity with programming in Python
ii. Becoming comfortable with regression
iii. Learning to work with data, choose appropriate variables
iv. Gain an understanding of clustering algorithms
v. Explore challenges associated with unlabeled data

Requirements:

You are expected to hand in a completed version of the Jupyter notebook provided:

1. Details of what steps are required, and associated marks are included in the Jupyter notebook itself.
2. Notes on getting setup are included in this document.

Note that your implementation for any given part of the problem may be simple (or complex) depending on design decisions made. I am most interested in your analysis of the problem, and design decisions that were made, and also presenting the opportunity to think critically about the problems presented. There are placeholders for you to include discussion, and the discussion you include is important to show that you have understood conceptual aspects of the problem.

Please rename the Jupyter notebook provided to lastname_firstname_A1.ipynb for your submission.

Getting started:

The recommended way of getting started follows. If you already have a great deal of experience with Python, you may already have a specific setup in place and should feel free to continue with this.

If this is entirely new to you, you might consider starting with the Anaconda Individual Distribution:

https://www.anaconda.com/products/individual

Here it is recommended that you install an up to date Anaconda distribution with Anaconda's latest stable Python 3 version. There are instructions for Linux, iOS and Windows users.

One simple method of working on the notebook itself is to do so within your web browser. Launching an Anaconda prompt (if windows), and the jupyter notebook command will typically take you to the the browser directly.

From here, you can create new notebooks, or in this case, navigate to the location of the A1.ipynb file and click on it.

This is a relatively simple way of getting off the ground. For any missing packages, these can be installed at the prompt (or in shell) using pip.

There are IDEs to work with Python, and typically your bits of Python could be broken down into individual .py files, however the notebook will allow you to work interactively, and produce your assignment report in the same place as your code.

If you are looking for a richer experience, you could consider running jupyter lab (there is also the option of suppressing the web interface entirely if you look deeper). There are also other many other options for a UI to edit jupyter notebooks that might appeal to you.

Summary of Assignment 1 elements (Detailed requirements are in the notebook markdown itself)

Part I

a. Simple implementation of incremental least squares (from scratch)
b. Reading in data related to movies, split into train, validation, test
c. Developing features, understanding properties of the data, visualization
d. Train a regression model (using sklearn or other)
e. Train a non-linear regression model with and without regularization

Part II

a. Simple implementation of k-means clustering (from scratch)
b. Determining number of clusters, justifying choice

Bonus – Guessing what the mystery data is, and showing why (don't spend much time on this, it is nearly impossible to guess 😊)

Part III

a. Implementing a simple recommender system for 1 person – based on the data provided, choose 1 individual in the data and generate recommendations based on data from other users