COMP 4432 Machine Learning

Lesson 1: Introduction

Agenda

- Syllabus
- Office Hours
- Biography
- Class Introductions
- Introduction to ML

In General

- In machine learning, a model is used to describe the process that results in the data observed (training data).

Syllabus

- Course Overview

  *This course explores machine learning techniques and theory. The course covers how to use popular machine learning libraries to develop, train, evaluate, and deploy predictive models on **prepared** data. Both design principles (machine learning types and tasks) and technical tools/languages will be covered.*

  This course will provide students with a foundation of machine learning grounded in the mathematical models behind the techniques, and will cover the theory and computational methods inherent in the application of machine learning models.

Syllabus

- Course Overview
  This course explores machine learning techniques and theory. The course covers how to use popular machine learning libraries to develop, train, evaluate, and deploy predictive models on prepared data. Both design principles (machine learning types and tasks) and technical tools/languages will be covered.

  This course will provide students with a foundation of machine learning grounded in the ***mathematical models*** behind the techniques, and will cover the ***theory*** and computational methods inherent in the ***application*** of machine learning models.

Syllabus

- Late assignments policy
  - One week grace then 10% off per week
  - Cannot accept anything after the quarter ends
- Honor code

Course Expectations

- Submit assignments on time

Course Expectations

- Submit assignments on time
- Comment your code
  - Not step-by-step, but decision points
  - Assume grader will not ask questions

Course Expectations

- Submit assignments on time
- Comment your code
  - Not step-by-step, but decision points
  - Assume grader will not ask questions
- Read, watch, and research
- Bring questions to the live sessions

Office Hours

- Automatic disqualifiers?
  - Weekday business hours?

Assignment 1 (due next Tuesday)

- **Part 1**: Data Loading and Preparation. Load the diabetes dataset into two numpy arrays: one for the feature set and one for the target. **Pick a single feature to try to predict the target (disease progression).** Document the reason you chose the feature you did. Break your single feature and target sets into training and test sets with the last 20 rows being in the test set.
- **Part 2**: Model Training. Instantiate a linear regression model, and train it with your single feature and target sets.
- **Part 3**: Prediction and Measurement. List the first 10 predictions on your single feature ~~training~~ **test** set. Print out the feature coefficient and the root mean squared error of your model.
- **Part 4**: Visualization. Print out a scatter plot with the feature you chose on the x-axis, and progression on the y-axis. Plot the regression line on this same graph with appropriate labels on each axis.

Instructor Bio

- Lucas (Luke) Sawle
- PhD from University of Denver
  - Theoretical Biophysics
    - Protein Folding Problem and Protein Stability
    - Large Scale Simulation and Analytics
- BSs in Physics and Applied Mathematics
- Data Scientist since 2016
  - Telecom (Corporate Finance and Strategy)
  - Banking (Anti-Money Laundering)
  - Healthcare

Class Introductions

- Name
- Location (Home)
- Academic experience
- Professional experience
- Coding experience
- Fields of Machine Learning that interest you?
- Expectations from this course

Asynchronous Material

- Thoughts on the definition and/or history
- Supervised versus Unsupervised
- Classification versus Regression
- Outlier detection
  - One-Class SVM
  - Isolation Forest

Asynchronous Material

- *What can go wrong?*
  - Data issues
  - Real world data is not good, let alone perfect
  - Large part of day-to-day
  - Feature engineering (Subject matter expertise)

Asynchronous Material

- Under- versus Over-fitting
  - Regularization
  - Hyperparameter tuning
    - Tree based models
    - XGBoost
    - Early stopping

Asynchronous Material

- Building Blocks
  - A lot of math…

Asynchronous Material

- Building Blocks
  - A lot of math…
  - Calculus
    - Derivatives and gradients

Asynchronous Material

- Building Blocks
  - A lot of math…
  - Calculus
    - Derivatives and gradients
  - Linear Algebra
    - Multiplication
      - Matrix **A** is (i x j). Matrix **B** is (m x n)
        - (i x j)(m x n). Only possible if $j = m$
        - Product ***AB*** will have dimension (i x n)
      - **AB** =/= **BA**

Linear Regression

$$J(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left(h_\theta(x^{(i)}) - y^{(i)} \right)^2$$

$$h_\theta(x^{(i)}) = \theta_0 x_0^{(i)} + \theta_1 x_1^{(i)} + \ldots + \theta_n x_n^{(i)} = \theta^T x^{(i)}$$

$$\theta = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \vdots \\ \theta_f \end{pmatrix} \quad x^{(i)} = \begin{pmatrix} x_0^{(i)} \\ x_1^{(i)} \\ \vdots \\ x_f^{(i)} \end{pmatrix}$$

$$J(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left(\theta^T x^{(i)} - y^{(i)} \right)^2$$

Linear Regression (Closed Form)

$$J(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left(\theta^T x^{(i)} - y^{(i)} \right)^2$$

$$J(\theta) = \frac{1}{2n} (X\theta - y)^T (X\theta - y)$$

$$\frac{\partial J}{\partial \theta} = 2X^T X\theta - 2X^T y = 0$$

$$\theta = (X^T X)^{-1} X^T y$$

Linear Regression (Gradient Descent)

$$J(\theta) = \frac{1}{2n} \sum_{i=1}^{n} \left(\theta^T x^{(i)} - y^{(i)} \right)^2$$

$$\theta_j \leftarrow \theta_j - \alpha \frac{\partial}{\partial \theta_j} J(\theta)$$

$$\frac{\partial J(\theta)}{\partial \theta_j} = \frac{1}{n} \sum_{i=1}^{n} \left(h_\theta(x^{(i)}) - y^{(i)} \right) x_j^{(i)}$$