Week 1: Introduction to Regression

A. What to Expect:

In the weeks to come we will be learning the concepts of machine learning problems that can be broadly divided into:

1. **Supervised learning**, in which one uses ground truth data (aka targets or labels) to train models and then uses those trained models to make predictions;
2. **Unsupervised learning**, in which one attempts to find patterns in the data without ground truth labels;
3. **Semi-supervised learning**, in which only a fraction of the data is labeled, which one then uses to infer labels for similar, unlabeled data within the dataset.

Supervised learning, in turn can be further divided into:

1. **Regression**, in which one's targets/labels can have a continuous range of values;
2. **Classification**, in which one's targets/labels can have only a discrete set of values.

For Week 1, we focus on regression.

Before we dive into different types of regression models, we'll explore **gradient descent**, the core of supervised machine learning algorithms in general. Every supervised ML model has a **loss function**, which measures the difference between **ground truth target/label values** and **predicted target/label values**, and which the user attempts to minimize with respect to the model's **parameters**. To minimize the loss function, one periodically updates the model parameters using numerical approximations of the loss function's gradients with respect to the parameters.

Perhaps the most accessible example of regression involves attempting to find a curve of best fit through a set of points in an x-y plane. In this case, x represents a single feature (e.g. foot length) from which one seeks to predict the target y (e.g. height). Attempting to draw a straight line which best fits the (x, y) = (foot length, height) points is called **univariate linear regression**. If you attempt to predict y from multiple features (e.g. arm length in addition to foot length), and thus attempt to find a (hyper-)plane of best fit, the analogous term is **multivariate linear regression**. **Polynomial regression** differs from **linear regression** (whether uni- or multivariate) only in that one must add features corresponding to (combinations of) higher powers of the original features.

We then delve into **regularization**. This is a technique which usually makes a model perform worse on **training data** (data used to update the model's parameters) but hopefully perform better on **testing data** (which the model hasn't seen before and from which it makes new predictions). To return to our (foot length, height) example, assuming no two people in the training dataset have the same foot length but different heights, it's possible to find a polynomial curve which passes through every person's data point exactly. However, this model will likely be highly complicated (i.e., it will likely involve a polynomial of high degree) and likely not generalize well; in other words, it's highly unlikely that such a complicated model will accurately predict a new person's height given their foot length. Moreover, adding columns to the dataset to account for high-degree features can dramatically increase one's storage and computational costs. Thus, regularization is introduced. Regularization adds a term (possibly more) to

the model's loss function, penalizing it if the model's parameters become too large in magnitude and/or too numerous.

Finally, we learn about more advanced regression techniques, such as **tree-based models** and **generalized linear models**. Tree-based models approximate the continuous range of possible target values as a set of closely spaced but still discrete categories. At each node, tree-based models find critical values for one feature in the dataset which allow the model to split data points into different classes. We will continue to learn about tree based models for classification in Week 2.

B. Week 1 Learning Objectives:

- To implement Linear and Non-linear regression models on real-life sample problems.
- To understand the output metrics and parameters to assess regression models.
- To visualize and report model outcomes after benchmarking across several methods.
- To analyze underlying limiting conditions (under predictions/over predictions)[correlation and causation].