Homework 1

EE425X - Machine Learning: A signal processing persepective

In this homework you will be learning a Linear Regression model on two types of data sets. The problem here is as follows. You are given a set of m independent training data points, $y, \boldsymbol{x}$ that satisfy

$$y = \theta^T \boldsymbol{x} + e$$

and your goal is to estimate θ using training data that is either simulated or real. Denote the estimate as $\hat{\theta}$. Here y is a scalar and $\boldsymbol{x}$ is an n length vector.

1 Simulated (Synthetic) Data

Generate your own data to simulate the linear regression model $y = \boldsymbol{\theta}^T \boldsymbol{x} + e$. Generate m such independent training data vectors.

Use all three types of approaches for training, sample code for which is given in the Python-intro handout: Pseudo-Inverse, Solution of Normal Equations, and Gradient-Descent. Report the normalized error $\|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}\|_2 / \|\boldsymbol{\theta}\|_2$ and the time taken in each case.

Also comment on how the data is generated and how the error relates to the error e. Repeat this experiment for the following settings:

(a) Pick $m = 30$, $n = 5$, and set $e = 0$. Let $\boldsymbol{\theta} = [1, 4, 2, 10, 23]$.
(b) Pick $m = 30$, $n = 5$, and set $\mathbb{E}[e^2] = 10^{-6}$. Let $\boldsymbol{\theta} = [1, 4, 2, 10, 23]$.
(c) Pick $m = 100$, $n = 5$, and set $\mathbb{E}[e^2] = 10^{-6}$. Let $\boldsymbol{\theta} = [1, 4, 2, 10, 23]$.
(d) Pick $m = 1000$, $n = 5$, and set $\mathbb{E}[e^2] = 10^{-6}$. Let $\boldsymbol{\theta} = [1, 4, 2, 10, 23]$.
(e) Pick $m = 100$, $n = 5$, and set $\mathbb{E}[e^2] = 10^{-4}$. Let $\boldsymbol{\theta} = [1, 4, 2, 10, 23]$.

2 Real Data

In this problem you will be applying linear regression to real world data. Download the data from `https://archive.ics.uci.edu/ml/datasets/Airfoil+Self-Noise#`.

This contains $m = 1503$ training data points and there are $n = 5$ features. The last column of this data-set (6-th) column represents the output, y.

Here, you do not have access to the true θ so instead report the following error metric: $\sum_{i=1}^{m} (y^{(i)} - \hat{\boldsymbol{\theta}}^T \boldsymbol{x}^{(i)})^2 / m$.

This is a large data set, so only implement Gradient Descent. Comment on the choice of `max-iter` and the learning-rate μ.

Extra Credit: Now, standardize the features (1-5 column) to ensure they have zero mean and unit variance, and repeat the experiment. Report the same results.

Extra Credit: Implement (batch) Stochastic Gradient Descent and comment on how the batch-size affects convergence.

3 What to turn in?

Submit a short report that discusses all of the above questions. Also submit your codes with clear documentation. Grading will be based on the quality of report and accuracy of implemented codes.