**Machine Learning**

**Assignment # 2**

**Regression (CLO 3)**

**Deadline: To be announced in the class**

**Part 1.** **Simple Linear Regression**

Predicting House Prices (One feature)

In this part we will use data on house sales in King County, where Seattle is located, to predict house prices using simple (one feature) linear regression. You will:

* Use the kc\_house\_data.csv for this assignment. The dataset is already split into test and train data. Use the kc\_house\_train\_data.csv to build your model and kc\_house\_test\_data.csv to measure its performance.
* Write a function to compute the Simple Linear Regression weights using the gradient descent method discussed in the class.
* Write a function to make predictions of the output given the input feature

**Things to do**

**1.** Firstly, load the dataset. This can be done using Pandas and NumPy libraries. The official documentation for Pandas is available at <http://pandas.pydata.org/pandas-docs/stable/getting_started/basics.html>

A good tutorial for understanding basic data manipulation in Python is available at

<http://cs231n.github.io/python-numpy-tutorial/>

**2.** Explore the dataset. Plot different features of the dataset against the sale price to visualize the correlations in the data. You may use scatter plots for visualizations.

**3.** Implement a python function for Simple regression model with single feature. Write a generic function that accepts a column of data (e.g, an Array) ‘input\_feature’ and another column ‘output’ and returns the Simple Linear Regression parameters ‘intercept’ and ‘slope’. Use the Gradient descent method discussed in the class.

**4.** Use “sqft\_living” to train the Simple Regression model for predicting house sale price.

**5.** Use test data to measure the performance of your model. To report the performance use RSS (Residual sum of squares). RSS or the sum of squared errors of prediction (SSE), is the sum of the squares of residuals (deviations predicted from actual empirical values of data). It is a measure of the discrepancy between the data and an estimation model. A small RSS indicates a tight fit of the model to the data.

**6.** Instead of using ‘sqft\_living’ to estimate prices we could use ‘bedrooms’ (a count of the number of bedrooms in the house) to estimate prices. Using your function from (3) calculate the Simple Linear Regression slope and intercept for estimating price based on bedrooms.

7. Now that we have 2 different models compute the RSS from BOTH models on TEST data.

**8. Extend the simple univariate model to multiple variable regression model for house prediction.**