Report: A1 Linear Regression

Brandon Maness

[Bmaness3@uncc.edu](mailto:Bmaness3@uncc.edu)

800828979

Problem Definition

The goal of this assignment is to be able to accurately predict the price a listed house will sell at based on it’s features, location, and rating using a linear regression model.

Method

I am using linear regression for this task. This involves analyzing and learning the data, finding correlations between features by analyzing scatter plots, normalizing the data through a standard normalization method, performing a least squares calculation to find weights for regression, and using gradient and mini-batch descent for our actual regression. I then tested the model using a testing set and predicted values to ensure accuracy.

Dataset Description

The dataset is large, sporting 21 feature columns each with 21,613 datapoints for a total of 453,837 datapoints in the set. The data was split at a 90/10 percent ratio between training and testing, respectively. The features of the dataset are as follows:

id: identifier for the house

date: date the house was sold

price: price the house was sold at

bedrooms: number of bedrooms

bathrooms: number of bathrooms

sqft\_living: The square-footage of the inside of the house

sqft\_lot: The square-footage of the lot the house is on

floors: amount of floors in the house

waterfront: 0 if not waterfront, 1 if waterfront

view: rates the view of the home from 0 to 4

condition: rates condition of home from 1 to 5

grade: grade given to describe quality of the house from 1 to 13

sqft\_above: square-footage of all floors minus the basement

sqft\_basement: square-footage of the basement

yr\_built: year the house was built

yr\_renovated: year renovated, 0 if never renovated

zipcode: zipcode of the house

lat: lattitude of the house

long: longitude of the house

sqft\_living15: Some modified version of sqft\_living

sqft\_lot15: Some modified version of sqft\_lot

Experimental Details

The data was gotten from Kaggle, and collected from multiple home sales over a number of years. In processing the data, I removed the id and date fields as they held no relevance, I also found and removed any null or NaN values. Normalization was necessary as a lot of the points (bathrooms, bedrooms, floors) could not be compared well because they were integer values as can be seen in the scatter plots.

I did not ger far enough into the assignment to tune hyperparameters or fully develop a learning algorithm due to sickness. I’m hoping for an extension so I can further work on this assignment.