Weekly Homework 4

Ava Chong
CS 1675: Intro to Machine Learning

February 14, 2019

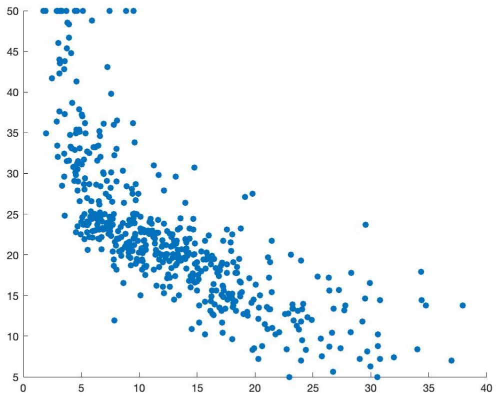
Problem 1. Exploratory Data Analysis

(a) There is 1 binary attribute: Charles River dummy variable
(b) Correlations between each attribute:

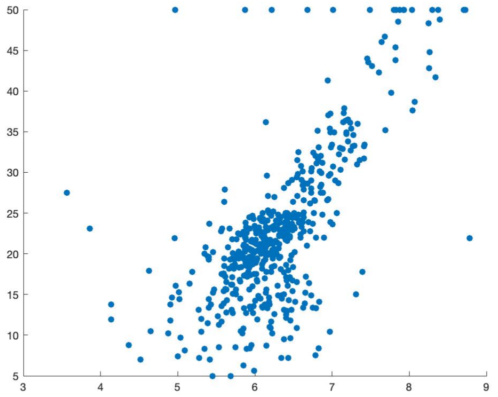
CRIM = -0.388
ZN = 0.360
INDUS = -0.483
CHAS = 0.175
NOX = -0.427
RM = 0.695
AGE = -0.376
DIS = 0.249
RAD = -0.381
TAX = -.468
PTRATIO = -0.507
B = 0.333
LSTAT = -0.737

(c) The scatter plots that looks most linear were those for LSTAT and RM

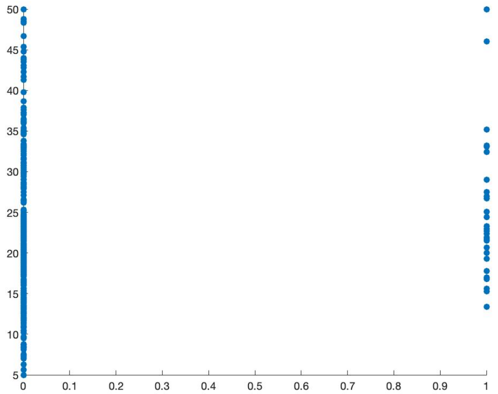
LSTAT Scatter-plot



RM Scatter-plot



The scatter plot that looks most nonlinear would be the binary attribute CHAS
CHAS Scatter-plot



(*d*) RAD and TAX have the greatest mutual correlation at .91

Problem 2. Linear Regression

(*a*) See code

(*b*) See code
(*c*) See code
(*d*) Resulting weights:

CRIM = -0.0979
ZN = 0.0489
INDUS = -0.0253
CHAS = 3.4508
NOX = -0.355
RM = 5.816
AGE = -0.00331
DIS = -1.0205
RAD = 0.226
TAX = -0.0122
PTRATIO = -0.3880
B = 0.01702
LSTAT = -0.485

Mean Squared Error for the training set: 24.4759
Mean Squared Error for the testing set: 24.2922

The testing set had a lesser error, making it better.

Problem 3. Online Gradient Descent
(*a*) See Code
(*b*) Mean Squared Error for the training set: 608.446
Mean Squared Error for the testing set: 487.656
The errors are much worse indicating that solving it this way is worse.
(*c*) Using un-normalized data caused the weights to be too large to use.
(*d*) I found that when playing with the parameters, the mean squared error got smaller as the number of iterations increased.