Assignment 1

Tanuj Kumar ta.kumar@mail.utoronto.ca 1002197133

October 2017

1 Question 1

1.1 Summary of Data

- **Number of data points:** 506
- **Number of features:** 13
- **Dimensions of design matrix:** 506×13
- **Target:** House values in Boston
- **Data:** 13 distinct features that potentially affect house values.
  - **CRIM:** Per capita crime rate by town
  - **ZN:** Proportion of residential land zoned for lots over 25000 square feet
  - **INDUS:** Proportion of non-retail business acres per town
  - **CHAS:** Charles River binary dummy variable (1 if the tract is next to the river, 0 otherwise)
  - **NOX:** Nitric oxides concentration
  - **RM:** Average number of rooms per dwelling
  - **AGE:** Proportion of owner-occupied units built prior to 1940
  - **DIS:** Weighted distances to five Boston employment centres
  - **RAD:** Index of accessibility to radial highways
  - **TAX:** Full-value property tax rate
  - **PTRATIO:** Pupil-teacher ratio by town
  - **B:** Proportion of black residents based on a specific area-growth formula (see: historic US policies on *redlining* and racist housing policy affecting land value)
  - **LSTAT:** Percent lower "status" of population

1.2 Feature Plots

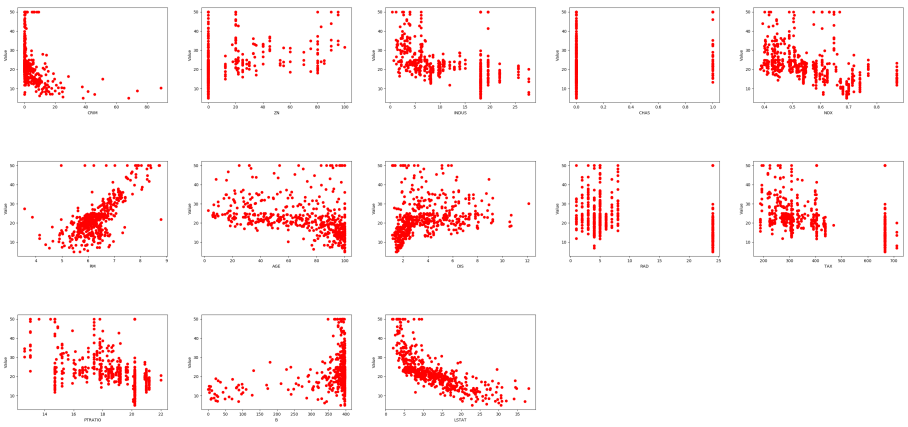

Figure 1: Plot of the features against the respective target values. From top left to right and bottom: CRIM, ZN, INDUS, CHAS, NOX, RM, AGE, DIS, RAD, TAX, PTRATIO, B, LSTAT

1.3 Feature Weight Tabulation

feature	weight
bias	31.8
crim	-11.7
zn	0.0445
indus	0.0279
chas	2.46
nox	-17.6
rm	4.31
age	-0.00511
dis	-1.45
rad	0.297
tax	-0.0111
ptratio	-0.920
b	0.0106
lstat	-0.495

1.4 Three Error Metrics

Three specific error metrics are used (with provided formulas):

- **Mean square error:**
  - $MSE = \frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i - y_i)^2$
  - **Average MSE error:** 19.05

- **Justification.** The mean square error is the natural choice based on the fact that linear regression used the above as its primarily loss function in the first place, when calculating $\mathbf{w}^*$.

- **Mean absolute error:**
  - $MAE = \frac{1}{n}\sum_{i=1}^{n} ||\hat{y}_i - y_i||$
  - **Average MAE error:** 3.83
  - **Justification.** The mean absolute error tends to be more robust to outliers, which shows with the lower loss values, and thus can better capture potential noise.

- **Mean square log error:**
  - $MSLE = \frac{1}{n}\sum_{i=1}^{n}(\log(y_i + 1) - \log(\hat{y}_i + 1))^2$
  - **Average MSLE:** 0.03
  - **Justification.** As we see with logistic regression, the logarithmic error can heavily limit the influence of outliers on the data and minimize the loss immensely with regards to the presence of noise. A loss this low may potentially run the risk of overfitting.

1.5 Significant Features

Based on weight from multiple runs of the regression, these are the top five features:

1. **CRIM**. High negative correlation. Intuitiely reasonable, because housing values plummet the more crime there is in a particular neighbourhood or town.

2. **NOX**. High negative correlation. Again, intuitively reasonable, as higher concentrations of nitric oxide can result in effects including headaches, dizziness, nausea, and loss of balance, due to effects on blood vessels. These negative physiologial effects likely contribute to lower housing values because of the discomfort in living in places with higher concentrations. Nitric oxides are some of the common pollutants produced by heavy industry.

3. **RM**. Medium positive correlation. The more rooms upon initial construction, the higher the house value. This describes the difference in price between a small two-room shack and a large mansion.

4. **CHAS**. Medium positive correlation. Riverfront properties have historically and presently resulted in higher values, worldwide (with the exception of potentially polluted rivers).

5. **DIS**. Medium negative correlation. This is perhaps best explained by the function of suburbs: the separation of home and work spaces enabling the suburban commute in exchange for parcels of private property and sprawl. While the "Boston employment centre" comment is vague, it is likely that higher-valued homes are farther away because those who can afford them are

often not employees but are likelier to be employers or higher management, paid higher in wages and so have more financial flexibility to own both larger homes, potentially work from home or possess more flexible hours, ands imply have the ability to commute more conveniently.

2 Question 2

2.1 Derivation of Locally Weighted Least Squares

Recall the basic LRLS function:

$$L(\mathbf{w}) = \frac{1}{2}\sum_{i=1}^{N} a^{(i)}(y - \mathbf{w}^T\mathbf{x})^2 + \frac{\lambda}{2}||\mathbf{w}||^2$$

We can write these in terms of matrices. Let $\mathbf{X}$ be the design matrix, $\mathbf{w}$ be the vector of weights (bias inclusive for both), $\mathbf{y}$ be the target vector, and $\mathbf{A}$ be the matrix such that $\mathbf{A}_{ii} = a^{(i)}$. Then:

$$L(\mathbf{w}) = \frac{1}{2}\mathbf{A}(\mathbf{y} - \mathbf{X}\mathbf{w})^2 + \frac{\lambda}{2}||\mathbf{w}||^2$$

$$L(\mathbf{w}) = \frac{1}{2}\mathbf{A}(\mathbf{y} - \mathbf{X}\mathbf{w})^T(\mathbf{y} - \mathbf{X}\mathbf{w}) + \lambda\mathbf{w}^T\mathbf{w}$$

$$L(\mathbf{w}) = \frac{1}{2}\mathbf{A}(\mathbf{y}^T\mathbf{y} + \mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - 2\mathbf{w}^T\mathbf{X}^T\mathbf{y}) + \frac{1}{2}\lambda\mathbf{w}^T\mathbf{w}$$

$$L(\mathbf{w}) = \frac{1}{2}\mathbf{A}\mathbf{y}^T\mathbf{y} + \frac{1}{2}\mathbf{A}\mathbf{w}^T\mathbf{X}^T\mathbf{X}\mathbf{w} - \mathbf{A}\mathbf{w}^T\mathbf{X}^T\mathbf{y} + \frac{1}{2}\lambda\mathbf{w}^T\mathbf{w}$$

And we will take the gradient of this with respect to $\mathbf{w}$:

$$\nabla_{\mathbf{w}} L(\mathbf{w}) = \mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{w} - \mathbf{X}^T\mathbf{A}\mathbf{y} + \lambda\mathbf{w}$$

To find the optimized parameters, we must minimize the gradient:

$$\nabla_{\mathbf{w}} L(\mathbf{w}) = 0$$

$$\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{w} - \mathbf{X}^T\mathbf{A}\mathbf{y} + \lambda\mathbf{w} = 0$$

$$\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{w} + \lambda\mathbf{w} = \mathbf{X}^T\mathbf{A}\mathbf{y}$$

And we solve for the minimizing weight vector, $\mathbf{w}^*$

$$\mathbf{X}^T\mathbf{A}\mathbf{X}\mathbf{w}^* + \lambda\mathbf{w}^* = \mathbf{X}^T\mathbf{A}\mathbf{y}$$

$$\mathbf{w}^*(\mathbf{X}^T\mathbf{A}\mathbf{X} + \lambda\mathbf{I}) = \mathbf{X}^T\mathbf{A}\mathbf{y}$$

Finally, if $(\mathbf{X}^T\mathbf{A}\mathbf{X} + \lambda\ \mathbf{I})$ is linearly independent i.e. invertible, then a minimum exists:

$$\mathbf{w}^* = (\mathbf{X}^T\mathbf{A}\mathbf{X} + \lambda\mathbf{I})^{-1}(\mathbf{X}^T\mathbf{A}\mathbf{y})$$

2.2 Loss Value Plot and Explanation

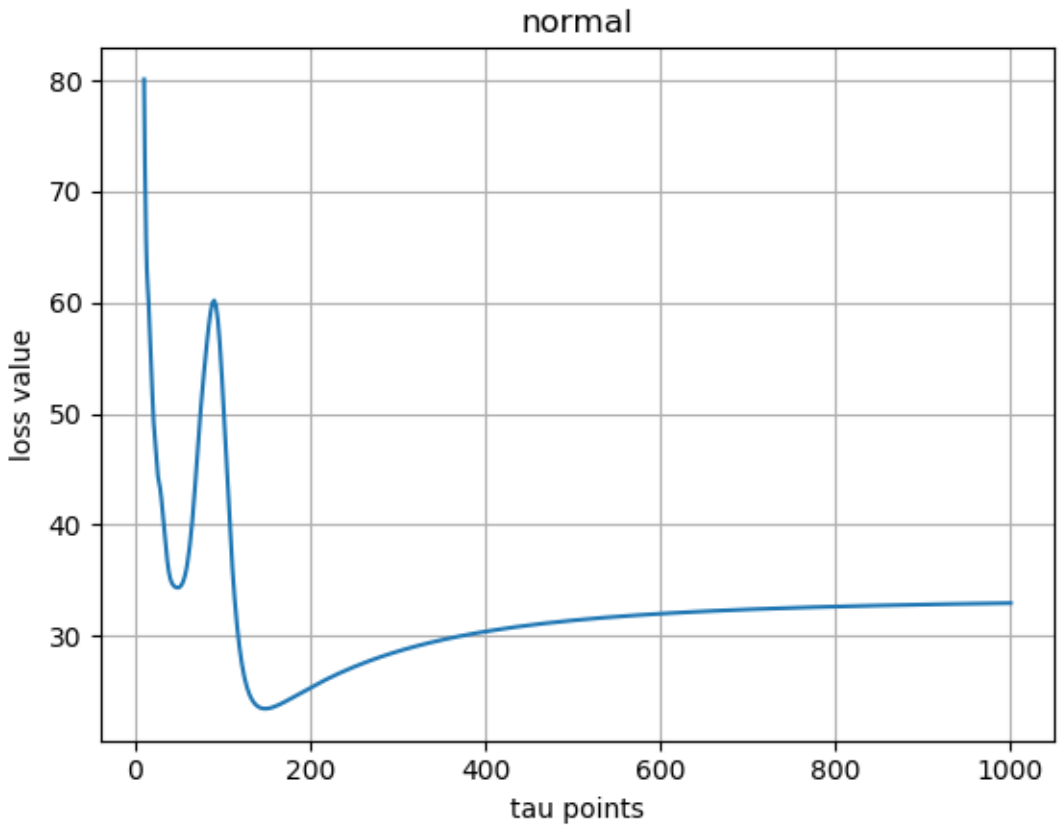

Figure 2: Plot of the tau value data vs. loss values. The first part of this graph is actually very common in physics when investigating the effects of distance on force or potential energy, and can be considered a physical rate law with a "well of lowest potential", here seen as the tau of minimized loss. There is also a local peak before exponential increase

Min Loss: 23.42

2.3 Behaviour at Extremes

- $\tau \to 0$: As the graph above shows, the losses grow exponentially large, and so as τ tends to 0 the loss value becomes infinite. This is intuitively reasonable simply due to the fact that τ^2 is a denominator term and so a small tau will result in extremely large weights that will influence the loss value results.

- $\tau \to \infty$: As the graph above shows, as τ tends to infinity, it essentially stabilizes and nigh flatlines around a particular value (here, about 32.08) for the loss. This is not the minimum loss, which is lower (seen in the dip), but makes tau irrelevant and minimizes the residual weight influences, meaning this is likely the natural loss of the A-less calculation with associated regularization term defined through λ's presence.

3 Question 3

3.1 Derivation 1: Expectation of Minibatch Sample Mean

Assume a **simple random sample.** Let $I = \{a_1...a_m\}$ be a batch of data points a_i of size m. Define the mean of all data points $a_1...a_N$ as

$$\mu = \frac{1}{n}(a_1 + ... + a_N) = \frac{1}{N}\sum_{i=1}^{N} a_i$$

Define random variables $X_i...X_m$ for each $a_1...a_m \in I$. $X_i = \frac{1}{n}$ for independent X_i, because each a_i is assumed to be **uniformly distributed**, and as a result, $\mathbf{E}(X_i) = \mu$ because the expected likelihood of picking up a_i as a value in our sample (aka data point) is the mean value over the entire population, which is μ as defined above.

Then, define $X_I = X_1 + ... + X_m = \sum_{i=1}^{m} X_i$, and the sample mean $\mu' = \frac{1}{m}\sum_{i=m}^{m} X_i$

$$\mathbf{E}(\mu') = \mathbf{E}(\frac{1}{m}\sum_{i=m}^{m} a_i) = \mathbf{E}(\frac{1}{m}\sum_{i=m}^{m} X_i)$$

$$\mathbf{E}(\frac{1}{m}\sum_{i=m}^{m} X_i) = \mathbf{E}(\frac{1}{m}X_1 + ... + \frac{1}{m}X_m)$$

The key idea here: **linearity of expectation**, because **E** is a **linear function**. We know that since I is a *simple random sample*, each a_i's selection is dependent, making the random variables dependent. However, **linearity of expectation holds regardless on if the constitutent random variables are independent or dependent**, thus, we can establish the following for the above:

$$\mathbf{E}(\frac{1}{m}X_1 + ... + \frac{1}{m}X_i) = \mathbf{E}(\frac{1}{m}X_1) + ... + \mathbf{E}(\frac{1}{m}X_m)$$

$$\mathbf{E}(\frac{1}{m}X_1 + ... + \frac{1}{m}X_i) = \frac{1}{m}\mathbf{E}(X_1) + ... + \frac{1}{m}\mathbf{E}(X_m)$$

The latter holding again because **E** is a linear function. Then, using the above definition of expectation on X_i:

$$\mathbf{E}(\frac{1}{m}X_1 + ... + \frac{1}{m}X_i) = \frac{1}{m}\mu_1 + ... + \frac{1}{m}\mu_m$$

Where each $\mu_i = \mu_j$, therefore:

$$\mathbf{E}(\frac{1}{m}X_1 + ... + \frac{1}{m}X_i) = \frac{m}{m}\mu$$

$$\mathbf{E}(\frac{1}{m}X_1 + ... + \frac{1}{m}X_i) = \mu$$

$$\mathbf{E}(\mu') = \mu$$

$$\mathbf{E}(\frac{1}{m}\sum_{i=m}^{m} a_i) = \frac{1}{N}\sum_{i=1}^{N} a_i$$

Which proves the claim.

3.2 Derivation 2: Expectation of Gradient of Minibatch Loss

Consider $\mathbf{E}(\nabla L_I(\mathbf{x}, y, \theta))$. Because L_I is a *discretized linear sum of smooth loss functions*, we can apply the **Leibniz integral rule**, by treating expectation as a (discretized) integration (aka, a sum of terms) that allows for the following gradient exchange:

$$\mathbf{E}(\nabla L_I(\mathbf{x}, y, \theta)) = \nabla(\mathbf{E}(L_I(\mathbf{x}, y, \theta)))$$

Then,

$$\mathbf{E}(\nabla L_I(\mathbf{x}, y, \theta)) = \nabla(\mathbf{E}(L_I(\mathbf{x}, y, \theta)))$$

$$\mathbf{E}(\nabla L_I(\mathbf{x}, y, \theta)) = \nabla(\mathbf{E}(\frac{1}{m}\sum_{i \in I}\mathcal{L}(\mathbf{x}^i, y^i, \theta)))$$

And by our above result in 3.1, including linearity of expectation,

$$\mathbf{E}(\nabla L_I(\mathbf{x}, y, \theta)) = \nabla(\mathbf{E}(\frac{1}{m}\sum_{i \in I}\mathcal{L}(\mathbf{x}^i, y^i, \theta)))$$

$$\mathbf{E}(\nabla L_I(\mathbf{x}, y, \theta)) = \nabla(\mathbf{E}(\frac{1}{m}\mathcal{L}(\mathbf{x}^1, y^1, \theta) + ... + \frac{1}{m}\mathcal{L}(\mathbf{x}^m, y^m, \theta)))$$

$$\mathbf{E}(\nabla L_I(\mathbf{x}, y, \theta)) = \nabla(\mathbf{E}(\frac{1}{m}\mathcal{L}(\mathbf{x}^1, y^1, \theta)) + ... + \mathbf{E}(\frac{1}{m}\mathcal{L}(\mathbf{x}^m, y^m, \theta)))$$

$$\mathbf{E}(\nabla L_I(\mathbf{x}, y, \theta)) = \nabla(\frac{1}{m}\mathbf{E}(\mathcal{L}(\mathbf{x}^1, y^1, \theta)) + ... + \frac{1}{m}\mathbf{E}(\mathcal{L}(\mathbf{x}^m, y^m, \theta)))$$

$$\mathbf{E}(\nabla L_I(\mathbf{x}, y, \theta)) = \nabla(\frac{1}{m}(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\mathbf{x}^i, y^i, \theta)) + ... + \frac{1}{m}(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\mathbf{x}^i, y^i, \theta)))$$

$$\mathbf{E}(\nabla L_I(\mathbf{x}, y, \theta)) = \nabla(\frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(\mathbf{x}^i, y^i, \theta))$$

$$\mathbf{E}(\nabla L_I(\mathbf{x}, y, \theta)) = \nabla L(\mathbf{x}, y, \theta)$$

Which proves the result.

3.3 Relevance of Expectation of Gradient

The expected gradient of the minibatch loss will be equal to the gradient of the entire loss function, i.e. minibatch loss behaviour is expected to be identical to global data loss behaviour.

3.4 Linear Regression Gradient Derivation

Recall:

$$\nabla L(\mathbf{x}, y, \theta) = \nabla(\frac{1}{n}\sum_{i=1}^{n}(y^i - \mathbf{w}^T\mathbf{x})^2)$$

Then:

$$\nabla L(\mathbf{x}, y, \theta) = \frac{1}{n}\nabla(\sum_{i=1}^{n}(y^i - \mathbf{w}^T\mathbf{x})^2)$$

$$\nabla L(\mathbf{x}, y, \theta) = \frac{1}{n}(\sum_{i=1}^{n}2(y^i - \mathbf{w}^T\mathbf{x})\mathbf{w}^T)$$

$$\nabla L(\mathbf{x}, y, \theta) = \frac{2}{n}(\sum_{i=1}^{n}(y^i - \mathbf{w}^T\mathbf{x}^i))\mathbf{w}^T$$

The below experimentation and cosine similarity will show that this is correct:

3.5 Experimentation and Cosine Similarity

- **Cosine similarity**: 1.0
- **Squared distance metric**: 5.04

The **cosine similarity** is a more accurate measurement of exactness. The cosine similarity defines the "angle" between vectors, with 1.0 matching exactness, and making the vector scale irrelevant, focusing primarily on orientation. The squared distance metric looks specifically at the vector square distance, excluding orientation. Because we are looking at an approximation of the overall gradient loss, we care more about orientation (how close the values are to each other, even if the vector scale is different) because the batch sample sizes affect the "scale" of these vectors.

3.6 Weight Variance Plot and Explanation

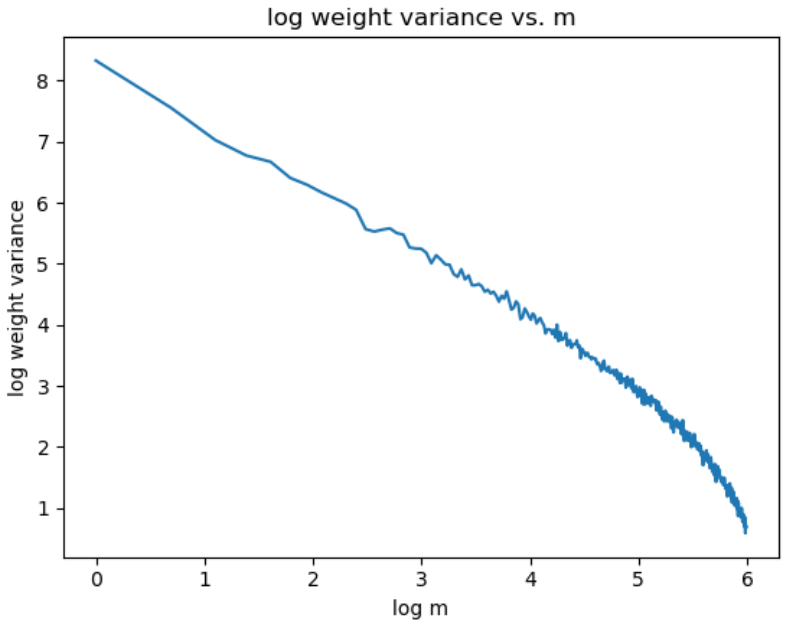

Figure 3: Plots the size of the minibatch against the weight variance for a specific weight (in this case w_1 which is CRIM). All other graphs look essentially identical.

This intuitively is reasonable. As the sample size increases, the variance in the gradient values and thus weight terms decreases because the approximation grows more exact as it reaches the size of the population itself.