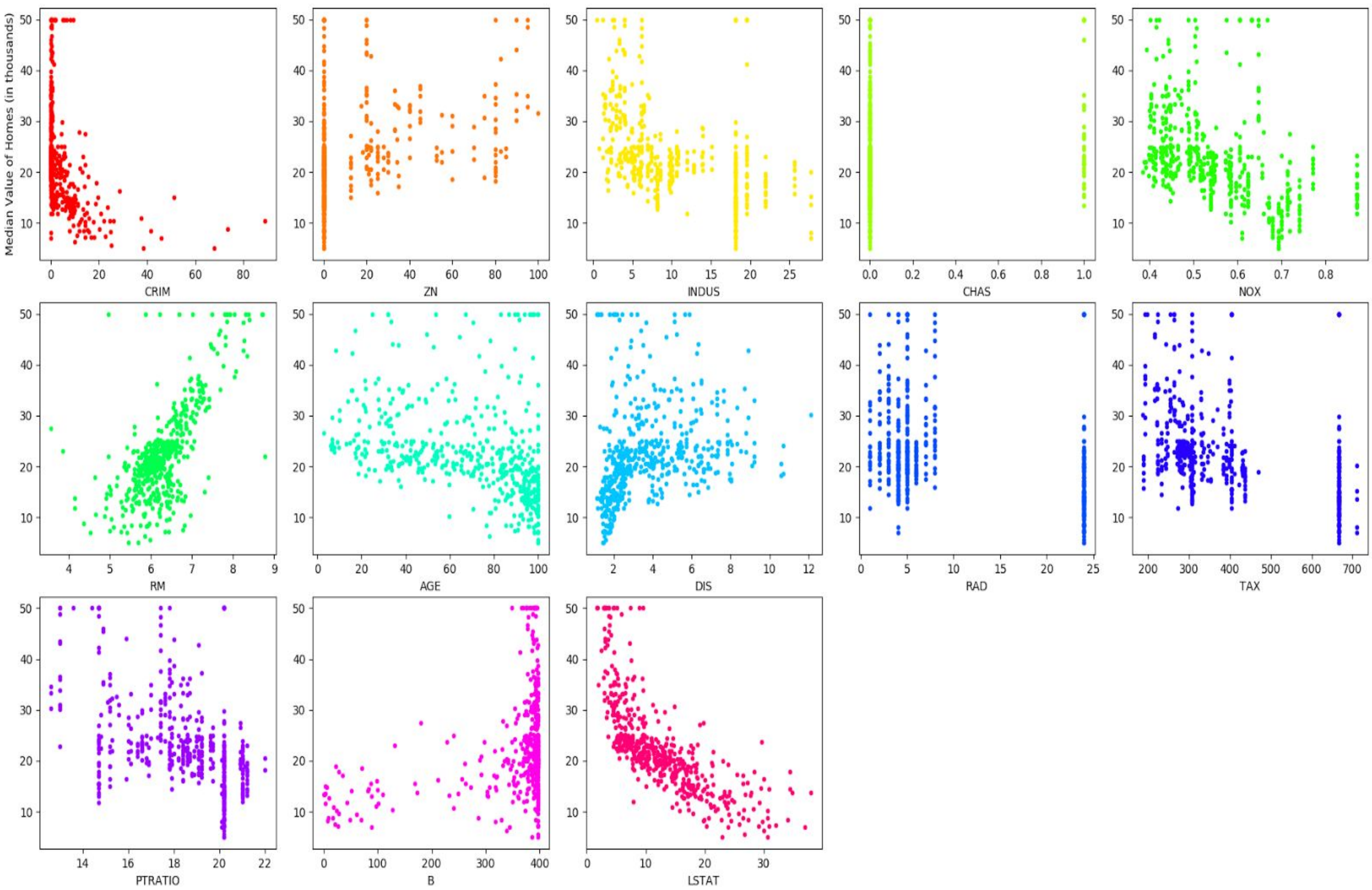
Ramon Sibello (999753253)
CSC411 - Assignment 1

1. The Boston Dataset contains 506 data points, of 14 dimensions, 13 features and a target. The target is the Median value of owner-occupied homes in $1000’s. The 13 features are as below:

   CRIM - per capita crime rate by town
   ZN - proportion of residential land zoned for lots over 25,000 sq.ft.
   INDUS - proportion of non-retail business acres per town.
   CHAS - Charles River dummy variable (1 if tract bounds river; 0 otherwise)
   NOX - nitric oxides concentration (parts per 10 million)
   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 per $10,000
   PTRATIO - pupil-teacher ratio by town
   B - 1000(Bk - 0.63)^2 where Bk is the proportion of blacks by town
   LSTAT - % lower status of the population



Feature	w
CRIM	-0.109486042009
ZN	0.0412084752285
INDUS	0.0107855924311
CHAS	1.93188004673
NOX	-17.7129537219
RM	3.28026748679
AGE	0.00428588684161
DIS	-1.37898240738
RAD	0.368367464429
TAX	-0.0155248993208
PTRATIO	-0.890566399962
B	0.00887265141401
LSTAT	-0.554260263396

The positive sign of the INDUS feature indicates a positive correlation between it and the target (Median value of homes). That is, a higher “proportion of non-retail business acres per town” tends to result in a higher “median value of homes”. If we assume that “non-retail business” mean industrial businesses, then a higher proportion would indicate a more prosperous town, increasing the median value. However, the weight value (w) is small for the INDUS feature, suggesting it doesn’t have a large effect on the median value of homes.

Mean Squared Error	16.4860293731
Mean Absolute Error	3.02310960593
Root Mean Squared Error	4.06029917286

The Mean Absolute Error gives insight into the average error for the predictions, without punishing far outliers severely. Root Mean Squared Error gives insight into the similarity between the two results, regardless of differences between any two specific elements, and the size of the set.

The most significant feature are CHAS, NOX, RM, and DIS, having the highest weights on the linear model, which makes sense based on their meaning.

2.

$$L(w) = (y-\hat{y})A(y-\hat{y}) + \lambda w^T w \qquad \text{with } \hat{y} = Xw$$

$$= (y^T A - \hat{y}^T A)(y - \hat{y}) + \lambda w^T w$$

$$= y^T A y - y^T A \hat{y} - \hat{y}^T A y + \hat{y}^T A \hat{y} + \lambda w^T w$$

$$= y^T A y - y^T A X w - w^T X^T A y + w^T X^T A X w + \lambda w^T w$$

$$= y^T A y + w^T X^T A X w - 2 w^T X^T A y + \lambda w^T w$$

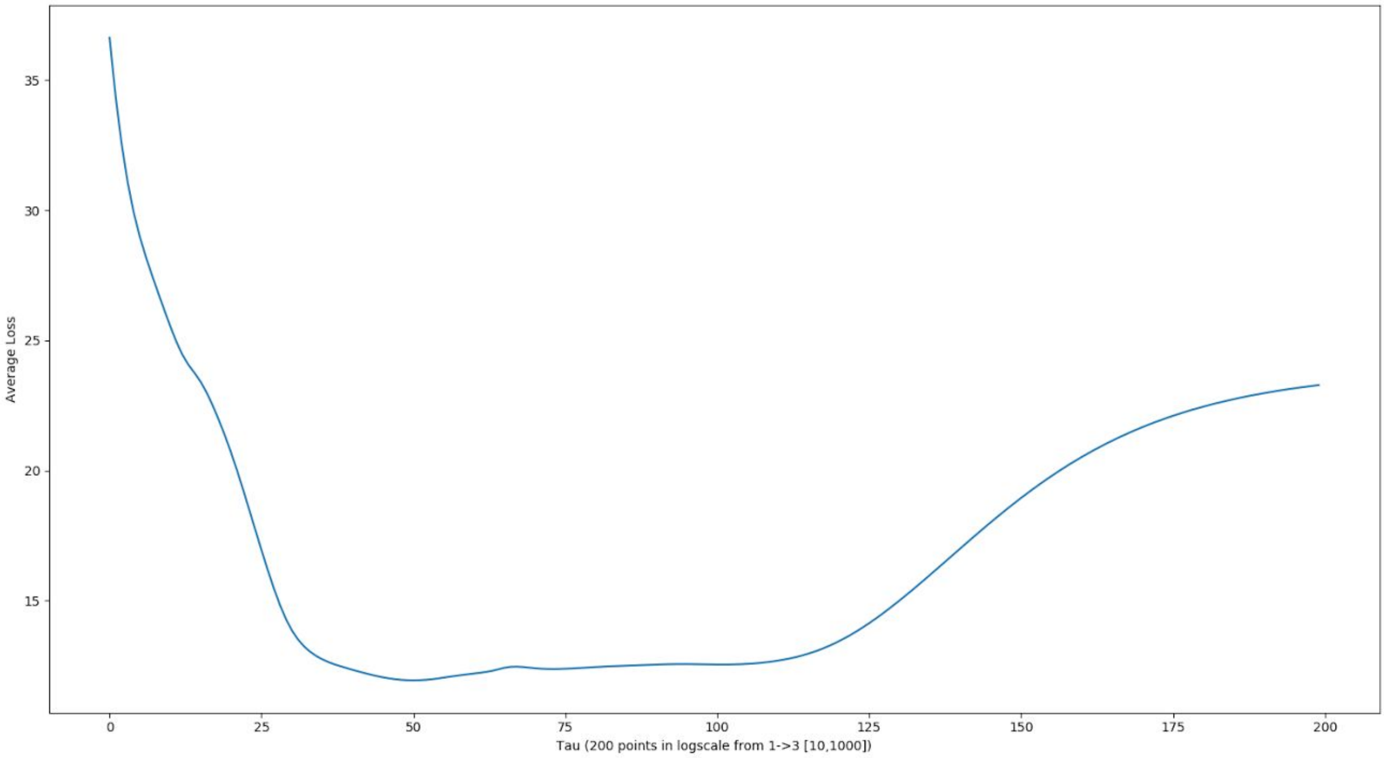
$$\nabla L(w) = 2X^T A X w - 2X^T A y + 2\lambda w$$

Set $\nabla L(w) = 0$, then:

$$2X^T A X w^* + 2\lambda w^* = 2X^T A y$$

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

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



As tau approaches infinity, the Average Loss increases, and as tau approaches zero the Average Loss increases sharply, presumably to infinity.

$\boxed{3}$ 1. Need to show that $\mathbb{E}_I\left[\frac{1}{m}\sum_{i\in I} a_i\right] = \frac{1}{n}\sum_{i=1}^{n} a_i$ (1)

we have $\mathbb{E}_x[f] = \sum_x p(x) f(x)$, then:

$$\mathbb{E}_I\left[\frac{1}{m}\sum_{i\in I} a_i\right] = \sum_I p(I)\frac{1}{m}\sum_{i\in I} a_i = \frac{1}{m}\sum_I p(I)\sum_{i\in I} a_i$$

$p(I) = \frac{m}{n}$ since we choose I randomly, then:

$$= \frac{1}{m}\sum_I \frac{m}{n}\sum_{i\in I} a_i = \frac{1}{m}\cdot\frac{m}{n}\sum_I\sum_{i\in I} a_i = \frac{1}{n}\sum_I\sum_{i\in I} a_i$$

we take I from set of n without replacement, so $\sum_I \sum_{i\in I} = \sum_{i=1}^{n}$

then we have $\frac{1}{n}\sum_I\sum_{i\in I} a_i = \frac{1}{n}\sum_{i=1}^{n} a_i$ ∎

2. Need to show that $\mathbb{E}_I[\nabla L_I(x,y,\theta)] = \nabla L(x,y\theta)$

we have $L_I(x,y,\theta) = \frac{1}{m}\sum_{i\in I} \ell(x^{(i)}, y^{(i)}, \theta)$, then:

$$\nabla L_I(x,y,\theta) = \frac{1}{m}\sum_{i\in I}\nabla \ell(x^{(i)}, y^{(i)}, \theta) \overset{(2)}{=} \frac{1}{m}\sum_{i\in I} a_i, \text{ letting } \nabla \ell(x^{(i)}, y^{(i)}, \theta)$$

Then, with (1) and (2) we have $\mathbb{E}_I[\nabla L_I(x,y,\theta)] = \nabla L(x,y,\theta)$ ∎

3. This shows that sample uniformly drawn from a set without replacement is an unbiased estimator, and can be used to estimate the entire set.

4. $L(X,y,\theta) = \frac{1}{n}\sum_{i=1}^{n}(y^{(i)} - w^T x^{(i)})^2$, $L(w) = \frac{1}{n}\|y-\hat{y}\|^2 = \frac{1}{n}\|y - Xw\|^2$

$$= \frac{1}{n}(y^Ty + w^TX^TXw - 2w^TX^Ty)$$

$$\nabla L(w^*) = \frac{2}{n}X^TXw^* - \frac{2}{n}Xy$$

Squared Distance Metric	3240214.56476
Cosine Similarity	0.999998418657

Cosine Similarity is a more meaningful metric because it measures the similarity between the vector’s directions, as opposed to the magnitude, which becomes irrelevant once it is used to compute the optimum **w** values.