STAT 339 - HOMEWORK 1B (LINEAR REGRESSIONS)

PENG GU, LIAM AXON, XIAOYUN GONG

Exercise 1

part(b). We applied our own OLS solver to the "womens100.csv" data set, with the results shown below:

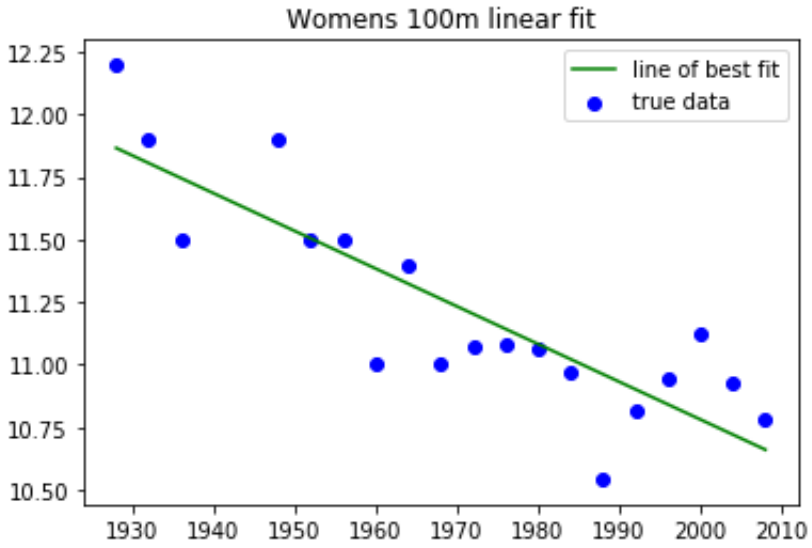

Figure 1. Graph of "womens100.csv" best fit

The line shown on the graph is $y = w_0 + w_1 x$ where $w_0 = 40.09$ and $w_1 = -0.015$.

We compared this to the line given in the textbook, which has $w_0 = 40.92$ and $w_1 = -0.015$.

These values are very close, and tell us that our regression solver is correctly finding a line of best fit.

Date: February 26, 2020.

part(c). Our algorithm predicted a value for the 2012 Olympics of 10.647, and it predicted a value for the 2016 Olympics of 10.593s. The actual racetime for 2012 is 10.75s, and the actual racetime for 2016 is 10.71s. These are both a little high of the actual value, but the difference is not prominent.

In fact, the squared prediction error for 2012 is 0.0106 and the squared prediction error for 2016 is 0.0136.

part(g). When we applied our polynomial regression solver to the "synthdata2016.csv" data set, looking for a cubic of best fit, we got the following:

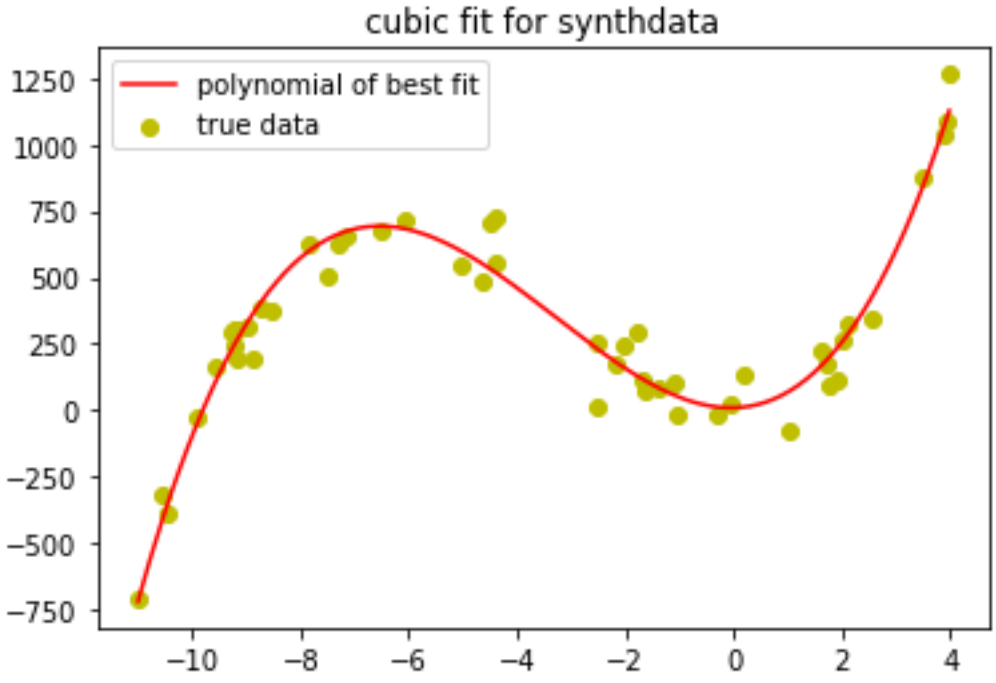

FIGURE 2. Graph of "synthdata2016.csv" best fit

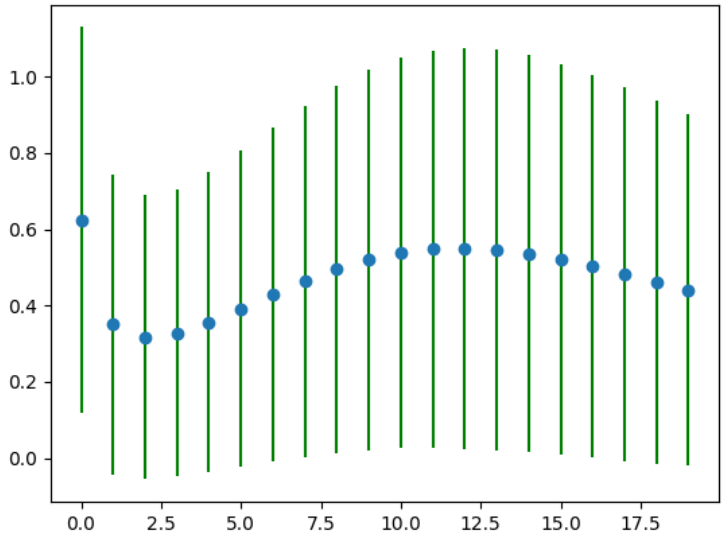

FIGURE 3. Validation error for "womens100.csv"(K=10)

EXERCISE 2

part(d). For womens100.csv, when K equals 10, optimal polynomial order according to the validation set is 2. When K equals N, optimal polynomial order according to the validation set is 2. The situation for the two K values is similar: it is hard to tell that one polynomial order is significantly better than the other due to the high standard deviation range in the validation error graph, however the training error graph makes sense since the graph can nearly perfectly fit every point when the order approaches the number of points, and possibly the original curve has polynomial around 2.5 so that's where training error was lowest.

For synthdata2016.csv, when K equals 10, optimal order is 4 according to the validation error graph. When K equals N, optimal order is 12. Both K values give very strange looking training error graphs. This might be due to the size

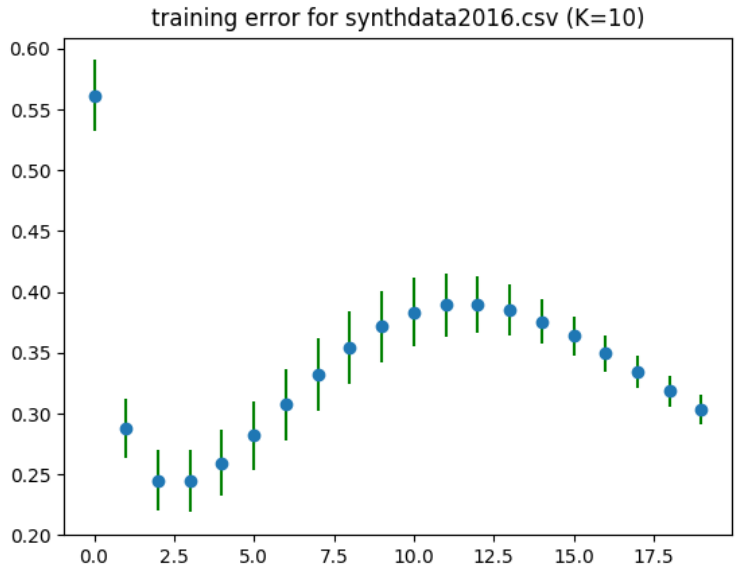

FIGURE 4. Train error for "womens100.csv"(K=10)

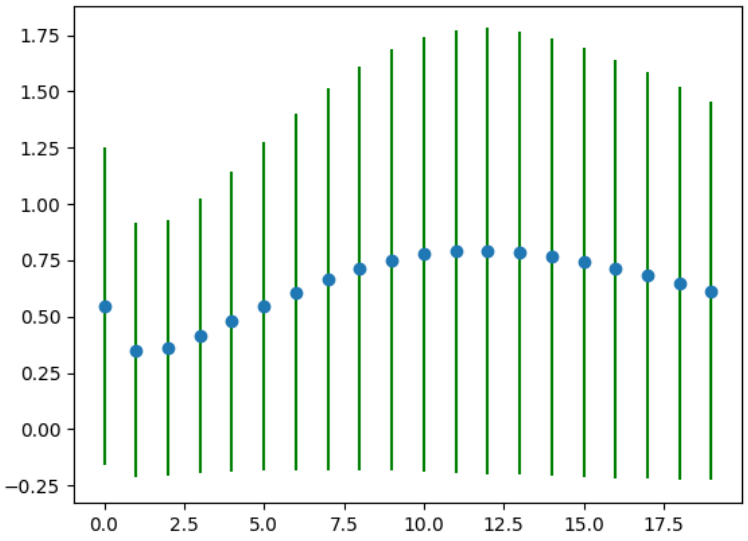

FIGURE 5. Validation error for "womens100.csv"(K=N)

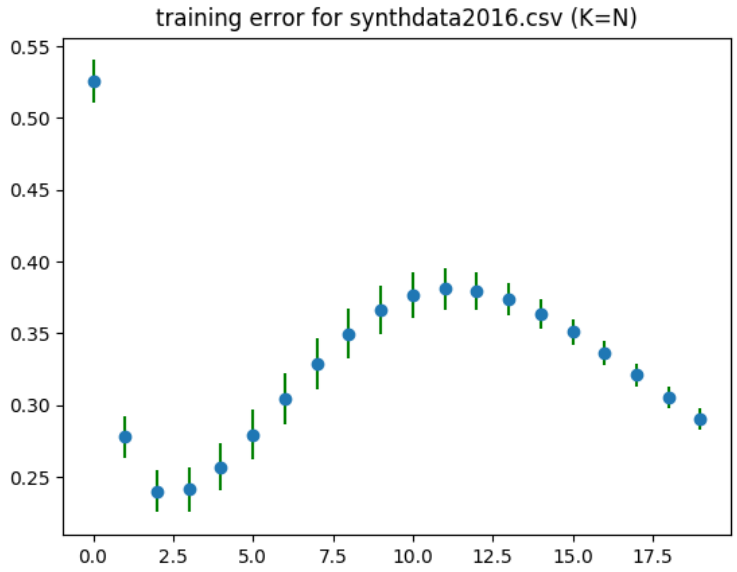

FIGURE 6. Train error for "womens100.csv" (K=N)

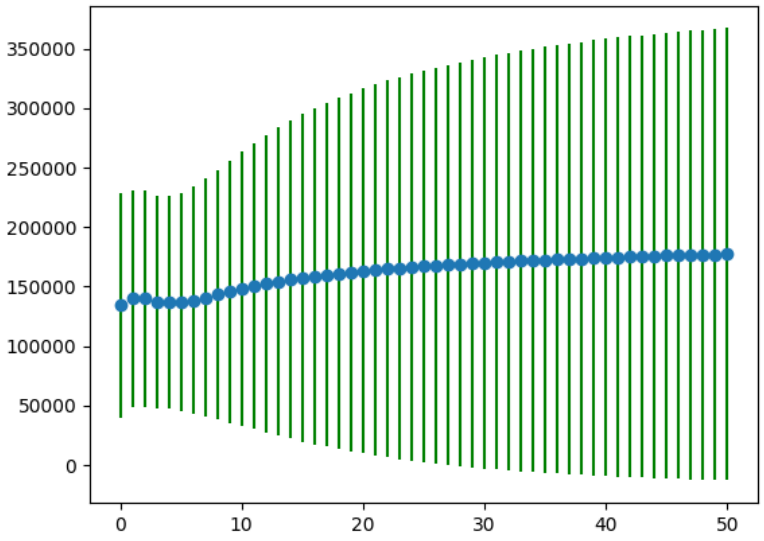

FIGURE 7. Validation error for "womens100.csv" (K=10)

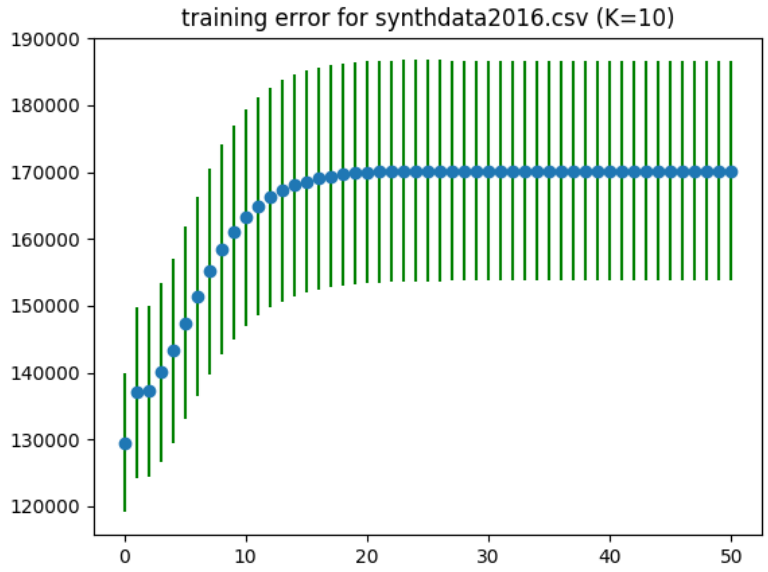

FIGURE 8. Train error for "womens100.csv"(K=10)

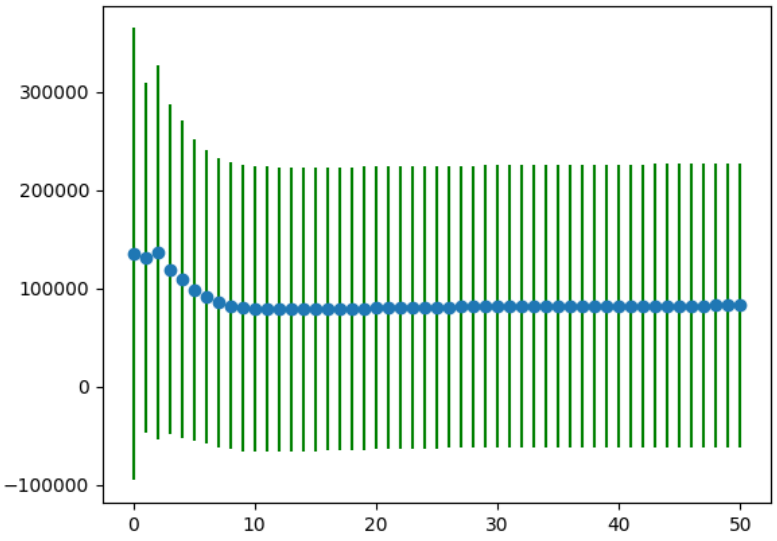

FIGURE 9. Validation error for "womens100.csv"(K=N)

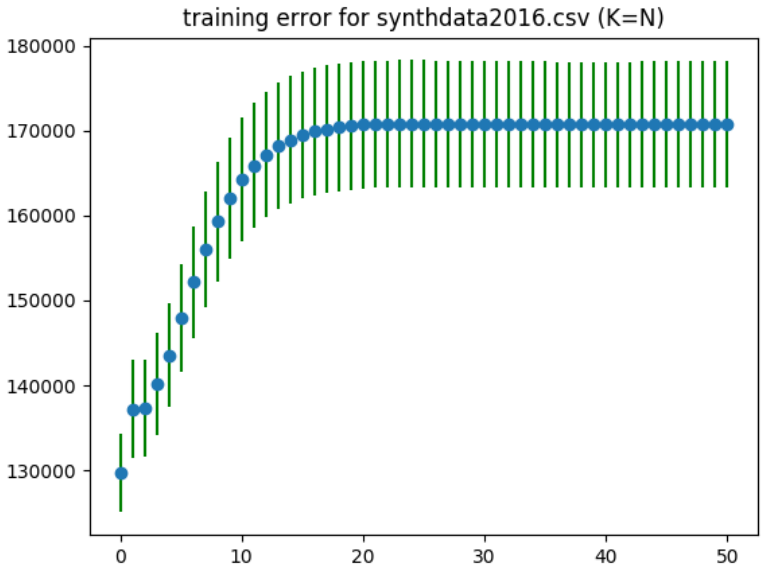

FIGURE 10. Train error for "womens100.csv"(K=N)

of the dataset. Validation error graphs are reasonable similar to those for womens100.csv.

part(e).

part(f). Grid search λ from 1 to 16 and D from 1 to 16. For synthedata2016.csv, optimal pair is $\lambda = 5$ and $D = 1$. Error for OLS is 0.05985, error for ridge regression is 0.062429. For womens100.csv, optimal pair is $\lambda = 2$ and $D = 1$. Error for OLS is 0.0002554 and error for ridge regression is 0.0030068. It seems that ridge regression requires more fine tuning in hyperparameteres, probably λ and D with shorter intervals within a larger range. With bad hyperparameters it is easily outperformed by OLS.

Exercise 3

part(a).

Proof. Let $f, g : \mathbb{R}^N \to \mathbb{R}^M$, $\mathbf{v} \in \mathbb{R}^N$, and $a = f(\mathbf{v})^T g(\mathbf{v}) \in \mathbb{R}$.

$$\begin{aligned}
\frac{da}{d\mathbf{v}} &= \frac{d}{d\mathbf{v}}(f(\mathbf{v})^T g(\mathbf{v})) \\
&= \frac{d}{d\mathbf{v}} \sum_{i=1}^{M} (f_i(\mathbf{v}) g_i(\mathbf{v})) \\
&= \sum_{i=1}^{M} \frac{d}{d\mathbf{v}} (f_i(\mathbf{v}) g_i(\mathbf{v})) \\
&= \sum_{i=1}^{M} (f_i(\mathbf{v}) \frac{dg_i}{d\mathbf{v}} + g_i(\mathbf{v}) \frac{df_i}{d\mathbf{v}}) \\
&= \sum_{i=1}^{M} f_i(\mathbf{v}) \frac{dg_i}{d\mathbf{v}} + \sum_{i=1}^{M} g_i(\mathbf{v}) \frac{df_i}{d\mathbf{v}} \\
&= f(\mathbf{v})^T \frac{dg}{d\mathbf{v}} + g(\mathbf{v})^T \frac{df}{d\mathbf{v}}
\end{aligned}$$

□

part(b).

Proof. Let $g : \mathbb{R}^N \to \mathbb{R}^N$, $g(\mathbf{v}) = \mathbf{A}\mathbf{v}$, and let $f = \mathbf{I}$.

$$\begin{aligned}
\frac{d}{d\mathbf{v}} \mathbf{v}^T \mathbf{A} \mathbf{v} = \frac{d}{d\mathbf{v}} f(\mathbf{v})^T g(\mathbf{v}) &= f(\mathbf{v})^T \frac{dg}{d\mathbf{v}} + g(\mathbf{v})^T \frac{df}{d\mathbf{v}} \\
&= f(\mathbf{v})^T \mathbf{A} + g(\mathbf{v})^T \mathbf{I} \\
&= \mathbf{v}^T \mathbf{A} + (\mathbf{A}\mathbf{v})^T \\
&= \mathbf{v}^T \mathbf{A} + \mathbf{v}^T \mathbf{A}^T \\
&= \mathbf{v}^T \mathbf{A} + \mathbf{v}^T \mathbf{A} \\
&= 2\mathbf{v}^T \mathbf{A}
\end{aligned}$$

□

Exercise 4

We need to find the derivative first:

$$\begin{aligned}\frac{d\mathcal{L}}{d\mathbf{w}} &= \frac{1}{N}[(\mathbf{t}-\mathbf{X}\mathbf{w})^T\frac{d}{d\mathbf{w}}(\mathbf{A}(\mathbf{t}-\mathbf{X}\mathbf{w})) + (\mathbf{A}(\mathbf{t}-\mathbf{X}\mathbf{w}))^T\frac{d}{d\mathbf{w}}(\mathbf{t}-\mathbf{X}\mathbf{w})] \\ &= \frac{1}{N}[(\mathbf{t}-\mathbf{X}\mathbf{w})^T(-\mathbf{A}\mathbf{X}) + (\mathbf{t}-\mathbf{X}\mathbf{w})^T\mathbf{A}^T(-\mathbf{X})]\end{aligned}$$

(Note that $\mathbf{A}$ is symmetric, so $\mathbf{A}^T = \mathbf{A}$.)

$$\begin{aligned}&= -\frac{1}{N}[(\mathbf{t}-\mathbf{X}\mathbf{w})^T\mathbf{A}\mathbf{X} + (\mathbf{t}-\mathbf{X}\mathbf{w})^T\mathbf{A}\mathbf{X}] \\ &= -\frac{2}{N}[(\mathbf{t}-\mathbf{X}\mathbf{w})^T\mathbf{A}\mathbf{X}]\end{aligned}$$

Now, set $\frac{d\mathcal{L}}{d\mathbf{w}}$ to 0:

$$\begin{aligned}-\frac{2}{N}[(\mathbf{t}-\mathbf{X}\mathbf{w})^T\mathbf{A}\mathbf{X}] &= 0 \\ (\mathbf{t}-\mathbf{X}\mathbf{w})^T\mathbf{A}\mathbf{X} &= 0 \\ (\mathbf{t}^T-\mathbf{w}^T\mathbf{X}^T)\mathbf{A}\mathbf{X} &= 0 \\ \mathbf{t}^T\mathbf{A}\mathbf{X}-\mathbf{w}^T\mathbf{X}^T\mathbf{A}\mathbf{X} &= 0 \\ \mathbf{w}^T\mathbf{X}^T\mathbf{A}\mathbf{X} &= \mathbf{t}^T\mathbf{A}\mathbf{X} \\ \mathbf{w}^T &= \mathbf{t}^T\mathbf{A}\mathbf{X}(\mathbf{X}^T\mathbf{A}\mathbf{X})^{-1} \\ \mathbf{w} &= ((\mathbf{X}^T\mathbf{A}\mathbf{X})^{-1})^T\mathbf{X}^T\mathbf{A}^T\mathbf{t} \\ \mathbf{w} &= ((\mathbf{X}^T\mathbf{A}\mathbf{X})^T)^{-1}\mathbf{X}^T\mathbf{A}^T\mathbf{t}\end{aligned}$$

(Again, note that $\mathbf{A}$ is symmetric, so $\mathbf{A}^T = \mathbf{A}$.)

$$\mathbf{w} = (\mathbf{X}^T\mathbf{A}\mathbf{X})^{-1}\mathbf{X}^T\mathbf{A}\mathbf{t}$$