Andrew Elysee (aelys2)
CS446

Homework #4

1.

a. The empirical risk with squared loss is:

$$R(w) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{2}(w^T x_i - y_i)^2$$
$$0 = \nabla R(w) = \frac{1}{n}\sum_{i=1}^{n} x_i(w^T x_i - y_i) = \frac{1}{n}\sum_{i=1}^{n} x_i w^T x_i - \frac{1}{n}\sum_{i=1}^{n} x_i y_i$$
$$\frac{1}{n}\sum_{i=1}^{n} x_i y_i = \frac{1}{n}\sum_{i=1}^{n} x_i w^T x_i$$

$$R(w) = \frac{1}{2n}\|Xw - y\|^2$$
$$0 = \nabla R(w) = \frac{1}{n}X^T(Xw - y)$$
$$0 = X^T X w - X^T y$$

Since we know that our x_i are represented by the standard basis vectors we can substitute our x_i for e_i

$$0 = \sum_{i=1}^{d}\sum_{j=1}^{n_i} e_i^T e_i w_i - e_i^T y$$
$$\sum_{i=1}^{d}\sum_{j=1}^{n_i} e_i^T e_i w = \sum_{i=1}^{d}\sum_{j=1}^{n_i} e_i^T y_{i,j}$$

Since e_i represents an standard basis vector $e_i^t e_i$ would be a $d \times d$ matrix with zeros everywhere except at (i, i) where it would equal 1. This means we can say that $e_i^T e_i w = w_i$. Keep in mind that w_i is technically a length d vector with 0's everywhere except at position i where it is equal to w_i.

$$\sum_{i=1}^{d}\sum_{j=1}^{n_i} w_i = \sum_{i=1}^{d}\sum_{j=1}^{n_i} e_i^T y_{i,j}$$

Since e_i^T is a length d standard basis vector and $y_{i,j}$ is a scalar, let $y_{ij} = e_i^T y_{i,j}$, the resulting vector of multiplying the two. At this point we can also remove outer summation to focus on each individual component of w.

$$\sum_{i=1}^{d}\sum_{j=1}^{n_i} w_i = \sum_{i=1}^{d}\sum_{j=1}^{n_i} y_{ij}$$
$$\sum_{j=1}^{n_i} w_i = \sum_{j=1}^{n_i} y_{ij}$$
$$n_i w_i = \sum_{j=1}^{n_i} y_{ij}$$
$$w_i = \frac{1}{n_i}\sum_{j=1}^{n_i} y_{ij}$$

This is what we wanted to show since both w_i and y_{ij} are both 0 everywhere except i, we can think of them as the scalar value at ith component of each vector.

b.

$$R(w) = \frac{1}{2n}\|Xw - y\|^2 = \frac{1}{2n}\|XX^{+}y - y\|^2$$

$$R(w) = \frac{1}{2n}\left\|\sum_{j=1}^{r} s_j u_j v_j^T \sum_{i=1}^{r} \frac{1}{s_i} v_i u_i^T \sum_{k=1}^{r} a_k u_k - y\right\|^2$$

$$R(w) = \frac{1}{2n}\left\|\sum_{i,j,k} \frac{s_j a_k}{s_i} u_j v_j^T v_i u_i^T u_k - y\right\|^2$$

$$R(w) = \frac{1}{2n}\left\|\sum_{i=1}^{r} a_i u_i - y\right\|^2 = \frac{1}{2n}\left\|\sum_{i=1}^{r} a_i u_i - \sum_{i=1}^{r} a_i u_i\right\|^2 = \frac{1}{2n} \times 0 = 0$$

This is what we needed to show.

c. Suppose that X has a $rank(X) = d$ meaning that $(x_i)_{i=1}^n$ spans $\mathbb{R}^d$. That means that X has d non-zero singular values. We also know that the squares of the singular values of X are the eigenvalues of $X^T X$. Consider that $X^T X$ is an $d \times d$ because X^T is $d \times n$ and X is $n \times d$. Also, the determinant is equal to the product of the eigenvalues. Since all the eigenvalues of $X^T X$ are all non-zero, we can come to the conclusion that $X^T X$ is invertible because the determinant does not equal 0.

   Now, let's consider the case that $(x_i)_{i=1}^n$ does not span $\mathbb{R}^d$. This would mean that $rank(X) < d$, so there would be less than d non-zero singular values. So, $X^T X$ would have at least 1 eigenvalue equal to 0, meaning that it's determinant would be 0. Thus, $X^T X$ is not invertible if $(x_i)_{i=1}^n$ does not span $\mathbb{R}^d$.

d. Let X be:

$$X = \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}$$

   where the SVD is:

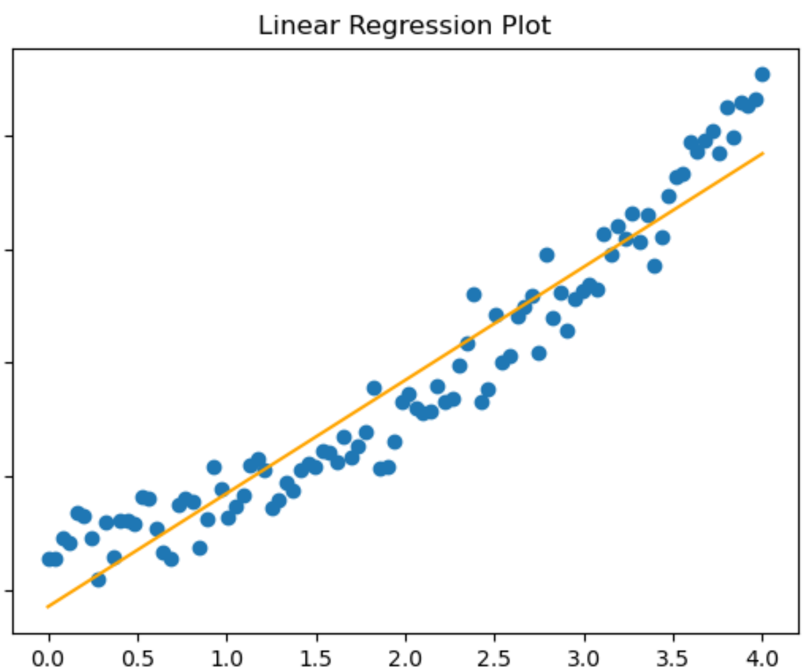
$$X = \begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}\begin{bmatrix} 2 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$X^T X = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}, \qquad XX^T = \begin{bmatrix} 4 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

   As we can clearly see, since X has only 2 singular values, the $rank(X) = 2$ meaning that it spans $\mathbb{R}^2$ but not $\mathbb{R}^3$. For this reason alone, we can conclude that $X^T X$ is invertible while XX^T is not. Furthermore, if we calculate $X^T X$ and XX^T, we see that their eigenvalues are [4,1] and [4,1,0] respectively. This means the $\det(X^T X) = 4$ making $X^T X$ invertible and the $\det(XX^T) = 0$ making XX^T not invertible.

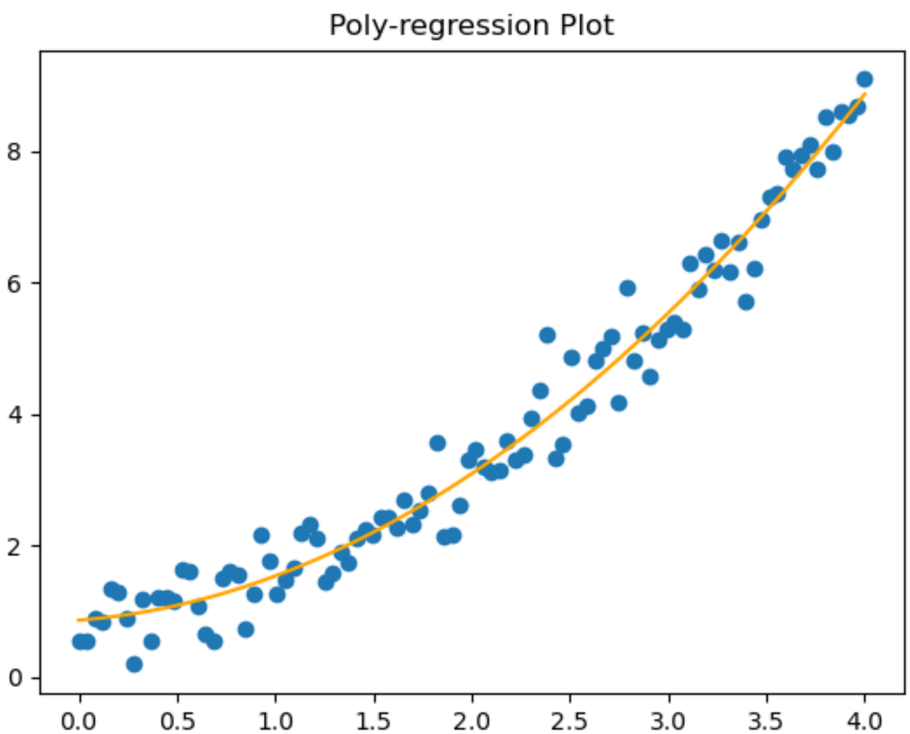
2.

a. In hw4.py
b. In hw4.py
c.



3.

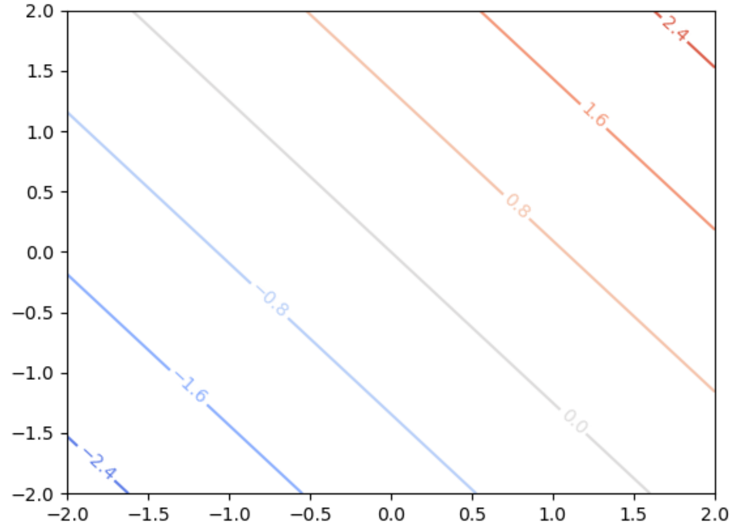
a. $\phi(x) = 1 + x_1 + x_2 + x_3 + x_1^2 + x_1x_2 + x_1x_3 + x_2^2 + x_2x_3 + x_3^2$
b. In hw4.py
c. In hw4.py
d. The Polynomial Regression gives a significantly better approximation of the data. As we can see below the data has a slight curve upwards, so the polynomial regression is more capable of staying withing the points and following that curve, while the linear regression is closer to the edge of the points in a lot of places.



e. As we can see in our plots below and our predictions, only the Polynomial Regression was able to correctly classify all our points. This makes sense because the polynomial is solving for the term x_1x_2 (which is the XOR function) while the linear regression does not.
Linear Predictions: [-1.4901e-08, 1.4901e-08, -1.3411e-07, 1.3411e-07]
Polynomial Predictions: [-1.0000, -1.0000, 1.0000, 1.0000]

<u>Plots on next page</u>

Linear Prediction Plot



Polynomial Prediction Plot