CS 446 / ECE 449 — Homework 4

sehsu2, dcku2 (partner)

Version 1.0

Instructions.

- Homework is due **Tuesday, April 4, at noon CST**; no late homework accepted.
- Everyone must submit individually at gradescope under `hw4` and `hw4code`.
- The "written" submission at `hw4` **must be typed**, and submitted in any format gradescope accepts (to be safe, submit a PDF). You may use LaTeX, Markdown, Google Docs, MS Word, whatever you like; but it must be typed!
- When submitting at `hw4`, gradescope will ask you to mark out boxes/pages around each of your answers; please do this precisely!
- Please make sure your NetID is clear and large on the first page of the homework.
- Your solution **must** be written in your own words. Please see the course webpage for full academic integrity information. Briefly, you may have high-level discussions with at most 3 classmates, whose NetIDs you should place on the first page of your solutions, and you should cite any external reference you use; despite all this, your solution must be written in your own words.
- We reserve the right to reduce the auto-graded score for `hw4code` if we detect funny business (e.g., your solution lacks any algorithm and hard-codes answers you obtained from someone else, or simply via trial-and-error with the autograder).
- Coding problems come with suggested "library routines"; we include these to reduce your time fishing around APIs, but you are free to use other APIs.
- When submitting to `hw4code`, only upload the two python files `hw4.py` and `hw4_utils.py`. Don't upload a zip file or additional files.

Version history.

1. Initial version.

1. Linear Regression/SVD.

Throughout this problem let $\boldsymbol{X}$ be the $n \times d$ matrix with the feature vectors $(\boldsymbol{x}_i)_{i=1}^n$ as its rows. Suppose we have the singular value decomposition $\boldsymbol{X} = \sum_{i=1}^r s_i \boldsymbol{u}_i \boldsymbol{v}_i^\top$.

Solution.

(a) The empirical risk minimization problem with squared loss is given by:

$$\boldsymbol{R}(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{d} \sum_{j=1}^{n_i} (\boldsymbol{w}^\top \boldsymbol{x}_i - y_{i_j})^2.$$

Substituting the training set, we get:

$$\boldsymbol{R}(\boldsymbol{w}) = \frac{1}{n} \sum_{i=1}^{d} \sum_{j=1}^{n_i} (\boldsymbol{w}^\top \boldsymbol{e}_i - y_{i_j})^2.$$

Setting the gradient with respect to $\boldsymbol{w}$ equal to zero, we get:

$$\frac{\partial \boldsymbol{R}(\boldsymbol{w})}{\partial \boldsymbol{w}} = \frac{1}{n} \sum_{i=1}^{d} \sum_{j=1}^{n_i} 2(\boldsymbol{w}^\top \boldsymbol{e}_i - y_{i_j}) \boldsymbol{e}_i = 0.$$

To minimize the empirical risk, the gradient is equal to the zero vector as $\boldsymbol{w}^\top \boldsymbol{e}_i = w_i$:

$$\sum_{i=1}^{d} \sum_{j=1}^{n_i} (w_i - y_{i_j}) \boldsymbol{e}_i = 0.$$

For each $k \in (1, \ldots, d)$ of the gradient, we obtain:

$$\sum_{j=1}^{n_k} (w_k - y_{k_j}) = 0.$$

The above equation is satisfied when:

$$w_k = \frac{1}{n_k} \sum_{j=1}^{n_k} y_{k_j}.$$

(b) We will use the SVD and compute the empirical risk. Since $\boldsymbol{y} = \sum_{i=1}^r a_i \boldsymbol{u}_i$ and a_i are some random constants, the SVD form of $\boldsymbol{X}$ is $\boldsymbol{X} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top$, and the pseudoinverse of $\boldsymbol{X}$ is obtained from $\boldsymbol{X}^+ = \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{U}^\top$:

$$\hat{\boldsymbol{w}}_{\text{ols}} = \boldsymbol{X}^+ \boldsymbol{y} = \boldsymbol{V}\boldsymbol{\Sigma}^+\boldsymbol{U}^\top \sum_{i=1}^{r} a_i \boldsymbol{u}_i = \boldsymbol{V}\boldsymbol{\Sigma}^\top \sum_{i=1}^{r} a_i \boldsymbol{e}_i$$

$$\boldsymbol{X}\hat{\boldsymbol{w}}_{\text{ols}} = \boldsymbol{X}\boldsymbol{X}^+ \boldsymbol{y} = \boldsymbol{U}\boldsymbol{\Sigma}\boldsymbol{V}^\top \boldsymbol{V} \sum_{i=1}^{r} a_i \boldsymbol{e}_i = \boldsymbol{U} \sum_{i=1}^{r} a_i \boldsymbol{e}_i = \sum_{i=1}^{r} a_i \boldsymbol{u}_i$$

Since $\boldsymbol{X}\hat{\boldsymbol{w}}_{\text{ols}} = \boldsymbol{y}$, this implies that the empirical risk is zero when the label vector $\boldsymbol{y}$ is a linear combination of the $\{\boldsymbol{u}_i\}_{i=1}^r$. On the other hand, the label vector $\boldsymbol{y}$ is not a linear combination of the $\{\boldsymbol{u}_i\}_{i=1}^r$ if the empirical risk is nonzero. The empirical risk is shown as:

$$\boldsymbol{R}(\hat{\boldsymbol{w}}_{\text{ols}}) = \frac{1}{n} ||\boldsymbol{X}\hat{\boldsymbol{w}}_{\text{ols}} - \boldsymbol{y}||^2$$

The above equation shows that the empirical risk must be nonzero because $\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} \neq \boldsymbol{y}$ and we know that $|\boldsymbol{P}_{\boldsymbol{X}}\boldsymbol{y} - \boldsymbol{y}|^2 > 0$.

(c) If $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible, all of its eigenvalues are nonzero. This implies that all the singular values of $\boldsymbol{X}$ are nonzero. Since the number of nonzero singular values equals the rank of $\boldsymbol{X}$, the rank of $\boldsymbol{X}$ is d. Therefore, $(\boldsymbol{x}_i)_{i=1}^n$ spans $\mathbb{R}^d$. If $(\boldsymbol{x}_i)_{i=1}^n$ spans $\mathbb{R}^d$, the rank of $\boldsymbol{X}$ is d. This means that $\boldsymbol{X}$ has d nonzero singular values. Since the squared singular values are the eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}$, all eigenvalues of $\boldsymbol{X}^\top \boldsymbol{X}$ are nonzero. Thus, $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible. Consequently, $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible if and only if $(\boldsymbol{x}_i)_{i=1}^n$ spans $\mathbb{R}^d$. This characterizes when linear regression has a unique solution due to the normal equation.

(d) Let's consider a 3×2 matrix $\boldsymbol{X}$ given by:

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

Now, let's compute $\boldsymbol{X}^\top \boldsymbol{X}$ and $\boldsymbol{X}\boldsymbol{X}^\top$:

$$\boldsymbol{X}^\top \boldsymbol{X} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

$$\boldsymbol{X}\boldsymbol{X}^\top = \begin{bmatrix} 1 & 0 \\ 0 & 1 \\ 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix}$$

To check whether a matrix is invertible or not, we calculate its determinant to see if it is nonzero. If the determinant is nonzero, then it is invertible. Otherwise, it is not invertible.

$$det(\boldsymbol{X}^\top \boldsymbol{X}) = 1$$

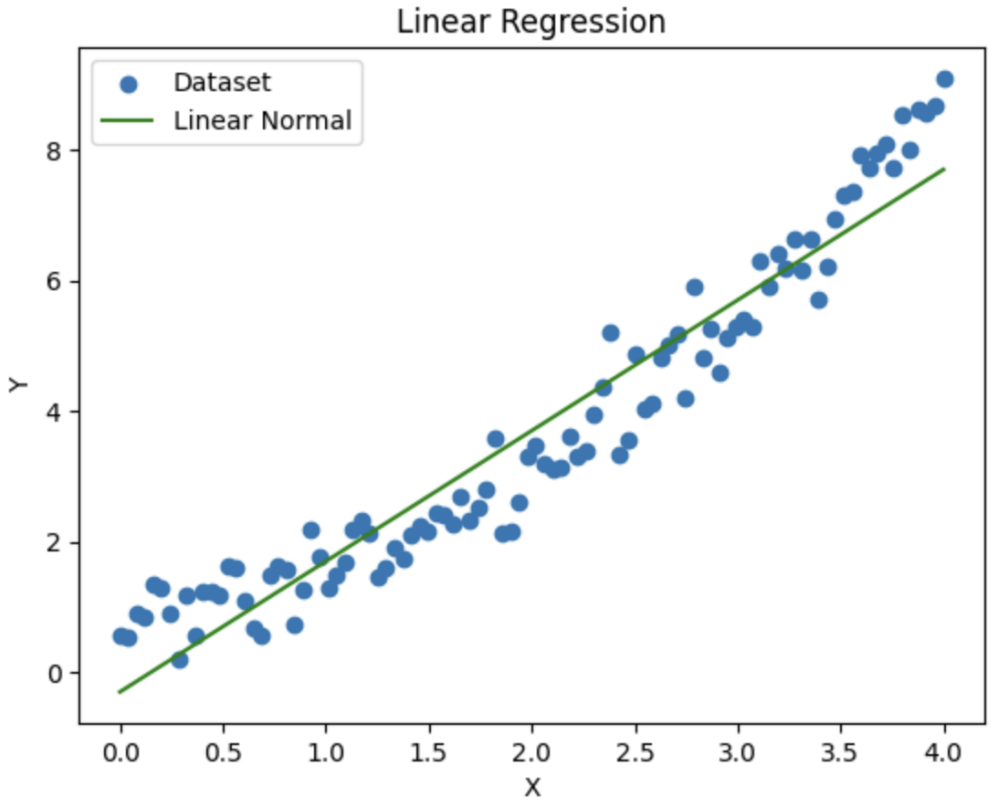
$$det(\boldsymbol{X}\boldsymbol{X}^\top) = 0$$

The above equations show that $\boldsymbol{X}^\top \boldsymbol{X}$ is invertible and $\boldsymbol{X}\boldsymbol{X}^\top$ is not invertible.

2. Linear Regression.

Recall that the empirical risk in the linear regression method is defined as $\widehat{\mathcal{R}}(\boldsymbol{w}) := \frac{1}{2n}\sum_{i=1}^{n}(\boldsymbol{w}^\top \boldsymbol{x}_i - y_i)^2$, where $\boldsymbol{x}_i \in \mathbb{R}^d$ is a data point and y_i is an associated label.

(c)



3. Polynomial Regression.

In Problem 3 you constructed a linear model $\boldsymbol{w}^\top \boldsymbol{x} = \sum_{i=1}^{d} x_i w_i$. In this problem you will use the same setup as in the previous problem, but enhance your linear model by doing a quadratic expansion of the features. Namely, you will construct a new linear model $f_{\boldsymbol{w}}$ with parameters

$$(w_0, w_{01}, \ldots, w_{0d}, w_{11}, w_{12}, \ldots, w_{1d}, w_{22}, w_{23}, \ldots, w_{2d}, \ldots, w_{dd})^\top,$$

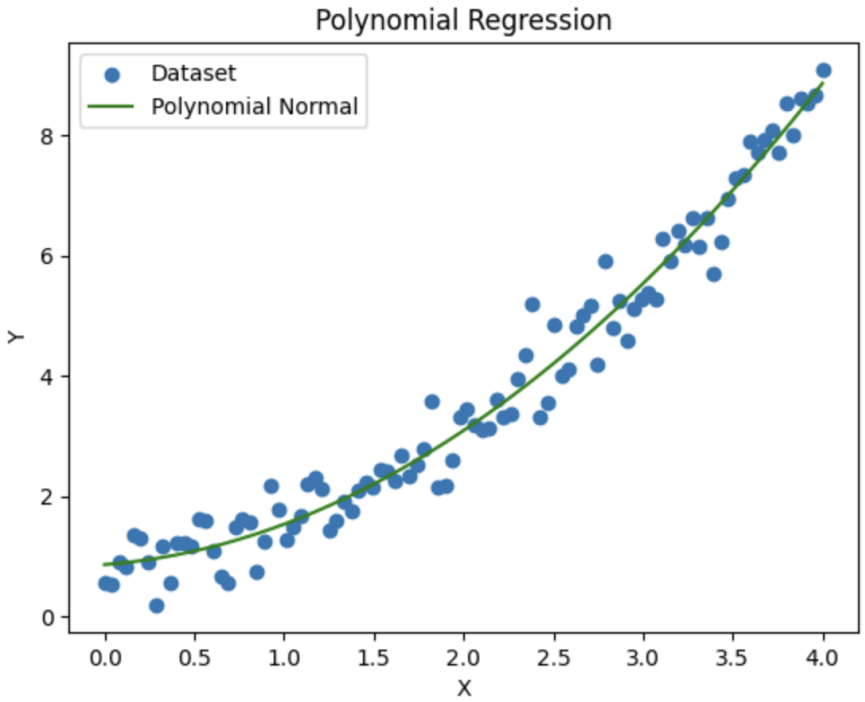
defined by

$$f_{\boldsymbol{w}}(x) = \boldsymbol{w}^\top \phi(\boldsymbol{x}) = w_0 + \sum_{i=1}^{d} w_{0i} x_i + \sum_{i \le j}^{d} w_{ij} x_i x_j.$$

(a)

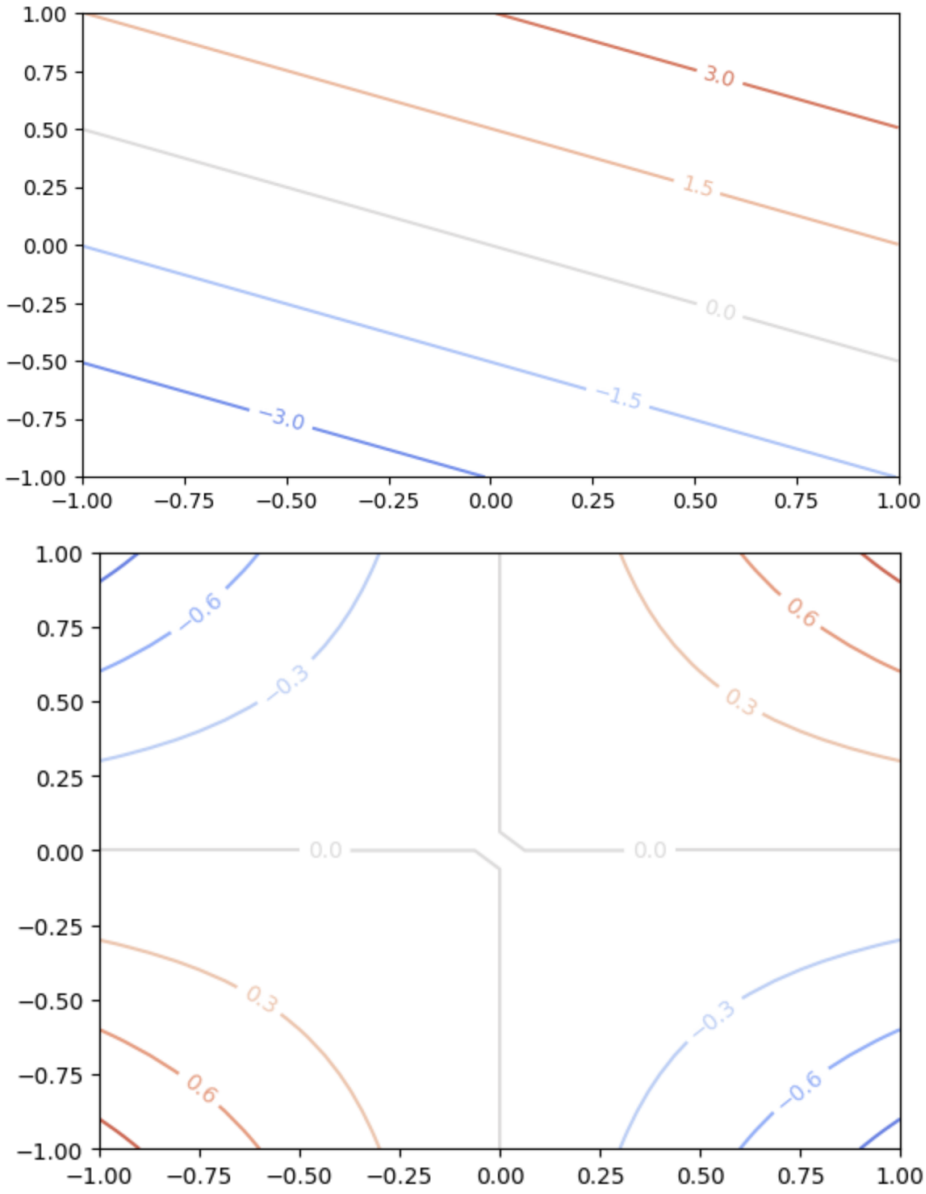
$$\phi(\boldsymbol{x}) = \begin{bmatrix} 1 \\ x_1 \\ x_2 \\ x_3 \\ x_1 x_1 \\ x_1 x_2 \\ x_1 x_3 \\ x_2 x_2 \\ x_2 x_3 \\ x_3 x_3 \end{bmatrix}$$

(d)



The polynomial model approximates the data better compared to the linear model. According to the polynomial regression figure, the polynomial normal seems to fit the dataset better.

(e)



According to the figures, a linear model (top figure) cannot correctly classify all points, as the problem is not linearly separable. A polynomial model (bottom figure), such as a second-degree polynomial model, can correctly classify all points because it can create a non-linear decision boundary.