

CS 446 / ECE 449 — Homework 1

mukaiyu2

Version 1.0

Instructions.

- Homework is due **Tuesday, February 8, at noon CST**; no late homework accepted.
- Everyone must submit individually at gradescope under **hw1** and **hw1code**.
- The “written” submission at **hw1** **must be typed**, and submitted in any format gradescope accepts (to be safe, submit a PDF). You may use L^AT_EX, Markdown, Google Docs, MS Word, whatever you like; but it must be typed!
- When submitting at **hw1**, gradescope will ask you to mark out boxes 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 **hw1code** 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 **hw1code**, only upload the two python files **hw1.py** and **hw1_utils.py**. Don’t upload a zip file or additional files.

Version history.

1. Initial version.

1. Linear Regression/SVD.

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

- (a) Let the training examples $(\mathbf{x}_i)_{i=1}^n$ be the standard basis vectors \mathbf{e}_i of \mathbb{R}^d with each \mathbf{e}_i repeated $n_i > 0$ times having labels $(y_{i_j})_{j=1}^{n_i}$. That is, our training set is:

$$\bigcup_{i=1}^d \left\{ (\mathbf{e}_i, y_{i_j}) \right\}_{j=1}^{n_i},$$

where $\sum_{i=1}^d n_i = n$. Show that for a vector \mathbf{w} that minimizes the empirical risk, the components w_i of \mathbf{w} are the averages of the labels $(y_{i_j})_{j=1}^{n_i}$: $w_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{i_j}$.

Hint: Write out the expression for the empirical risk with the squared loss and set the gradient equal to zero.

Remark: This gives some intuition as to why “regression” originally meant “regression towards the mean.”

- (b) Returning to a general matrix \mathbf{X} , show that if the label vector \mathbf{y} is a linear combination of the $\{\mathbf{u}_i\}_{i=1}^r$ then there exists a \mathbf{w} for which the empirical risk is zero (meaning $\mathbf{X}\mathbf{w} = \mathbf{y}$).

Hint: Either consider the range of \mathbf{X} and use the SVD, or compute the empirical risk explicitly with $\mathbf{y} = \sum_{i=1}^r a_i \mathbf{u}_i$ for some constants a_i and $\hat{\mathbf{w}}_{\text{ols}} = \mathbf{X}^+ \mathbf{y}$.

Remark: It’s also not hard to show that if \mathbf{y} is not a linear combination of the $\{\mathbf{u}_i\}_{i=1}^r$, then the empirical risk must be nonzero.

- (c) Show that $\mathbf{X}^\top \mathbf{X}$ is invertible if and only if $(\mathbf{x}_i)_{i=1}^n$ spans \mathbb{R}^d .

Hint: Recall that the squares of the singular values of \mathbf{X} are eigenvalues of $\mathbf{X}^\top \mathbf{X}$.

Remark: This characterizes when linear regression has a unique solution due to the normal equation (note that we always have at least one solution obtained by the pseudoinverse). We would not have had a unique solution for part (a) if we had an $n_i = 0$.

- (d) Provide a matrix \mathbf{X} such that $\mathbf{X}^\top \mathbf{X}$ is invertible and $\mathbf{X}\mathbf{X}^\top$ is not. Include a formal verification of this for full points.

Hint: Use part (c). It may be helpful to think about conditions under which a matrix is not invertible.

Solution.

- (a) The empirical risk of squared loss is:

$$\hat{R}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \ell_{sq}(\mathbf{w}^\top \mathbf{x}_i, y_i) = \frac{1}{2n} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|^2$$

Whose gradient is:

$$\nabla \hat{R}(\mathbf{w}) = \mathbf{X}^\top (\mathbf{X}\mathbf{w} - \mathbf{y})$$

And when set to zero we get the normal equation:

$$\mathbf{X}^\top \mathbf{X} \mathbf{w} = \mathbf{X}^\top \mathbf{y}$$

Under the context of (a) we have:

$$\mathbf{X}_{:,i}^\top \mathbf{X}_{:,i} \mathbf{w} = \mathbf{X}_{:,i}^\top \mathbf{y}$$

Which is equal to:

$$n_i w_i = \sum_{j=1}^{n_i} y_{i_j}$$

Which is the same as:

$$w_i = \frac{1}{n_i} \sum_{j=1}^{n_i} y_{i_j}$$

(b) The empirical risk of $\hat{\mathbf{w}}_{ols}$ is:

$$\begin{aligned}\hat{R}(\mathbf{w}) &= \frac{1}{2n} \|\mathbf{X}\hat{\mathbf{w}}_{ols} - \mathbf{y}\|^2 \\ &= \frac{1}{2n} \left\| \sum_{i=1}^r s_i \mathbf{u}_i \mathbf{v}_i^T \sum_{j=1}^r \frac{1}{s_j} \mathbf{v}_j \mathbf{u}_j^T \sum_{k=1}^r a_k \mathbf{u}_k - \sum_{k=1}^r a_k \mathbf{u}_k \right\|^2 \\ &= \frac{1}{2n} \left\| \sum_{i=1}^r \sum_{j=1}^r \sum_{k=1}^r \frac{s_i a_k}{s_j} \mathbf{u}_i \mathbf{v}_i^T \mathbf{v}_j \mathbf{u}_j^T \mathbf{u}_k - \sum_{k=1}^r a_k \mathbf{u}_k \right\|^2\end{aligned}$$

And because \mathbf{u}, \mathbf{v} are orthonormal vectors, we have:

$$\begin{aligned}\hat{R}(\mathbf{w}) &= \frac{1}{2n} \left\| \sum_{i=1}^r \sum_{k=1}^r \frac{s_i a_k}{s_i} \mathbf{u}_i \mathbf{v}_i^T \mathbf{v}_i \mathbf{u}_i^T \mathbf{u}_k - \sum_{k=1}^r a_k \mathbf{u}_k \right\|^2 \\ &= \frac{1}{2n} \left\| \sum_{i=1}^r \sum_{k=1}^r a_k \mathbf{u}_i \mathbf{u}_i^T \mathbf{u}_k - \sum_{k=1}^r a_k \mathbf{u}_k \right\|^2 \\ &= \frac{1}{2n} \left\| \sum_{k=1}^r a_k \mathbf{u}_k \mathbf{u}_k^T \mathbf{u}_k - \sum_{k=1}^r a_k \mathbf{u}_k \right\|^2 \\ &= \frac{1}{2n} \left\| \sum_{k=1}^r a_k \mathbf{u}_k - \sum_{k=1}^r a_k \mathbf{u}_k \right\|^2 \\ &= 0\end{aligned}$$

So $\hat{\mathbf{w}}_{ols}$ is such a \mathbf{w} that minimizes the empirical risk to 0.

- (c) i. $\mathbf{X}^T \mathbf{X}$ is invertible \Leftrightarrow The null space of $\mathbf{X}^T \mathbf{X}$ is the set of $\vec{0}$
ii. $(\mathbf{x}_i)_{i=1}^n$ spans $\mathbb{R}^d \Leftrightarrow$ The dimension of both the row space and column space of \mathbf{X} is $d \Leftrightarrow$ The null space of \mathbf{X} is the set of $\vec{0}$

So it suffices to prove that the null space of $\mathbf{X}^T \mathbf{X}$ and \mathbf{X} are always the same subspace, it will be proved in both directions respectively:

- i. If $\mathbf{X}\mathbf{v} = \vec{0}$,

Then left multiply both side with \mathbf{X}^T we get: $\mathbf{X}^T \mathbf{X}\mathbf{v} = \mathbf{X}^T \vec{0} = \vec{0}$,
So $\forall \mathbf{v} \in Null(\mathbf{X}), \mathbf{v} \in Null(\mathbf{X}^T \mathbf{X})$

- ii. If $\mathbf{X}^T \mathbf{X}\mathbf{v} = \vec{0}$,

Then left multiply both side with \mathbf{v}^T we get:

$$\begin{aligned}\mathbf{v}^T \mathbf{X}^T \mathbf{X}\mathbf{v} &= \mathbf{v}^T \vec{0} = \vec{0} \Rightarrow \\ \mathbf{v}^T \mathbf{X}^T \mathbf{X}\mathbf{v} &= (\mathbf{X}\mathbf{v})^T \mathbf{X}\mathbf{v} = \vec{0} \Rightarrow \\ \mathbf{X}\mathbf{v} &= \vec{0},\end{aligned}$$

So $\forall \mathbf{v} \in Null(\mathbf{X}^T \mathbf{X}), \mathbf{v} \in Null(\mathbf{X})$

So $Null(\mathbf{X}^T \mathbf{X})$ and $Null(\mathbf{X})$ are always the same.

Thus: $\mathbf{X}^T \mathbf{X}$ is invertible $\Leftrightarrow (\mathbf{x}_i)_{i=1}^n$ spans \mathbb{R}^d

- (d) $\mathbf{X} = \begin{bmatrix} 1 \\ 0 \end{bmatrix}$,

Then $\mathbf{X}^T \mathbf{X} = [1]$, whose inverse is $[1]$,

But $\mathbf{X}\mathbf{X}^T = \begin{bmatrix} 1 & 0 \\ 0 & 0 \end{bmatrix}$, which is not invertible because $\begin{bmatrix} 0 \\ 1 \end{bmatrix} \neq \vec{0}$, but $\mathbf{X}\mathbf{X}^T \begin{bmatrix} 0 \\ 1 \end{bmatrix} = \vec{0}$.

2. Logistic Loss.

Throughout this problem, let $\ell_{\log}(z) := \ln(1 + \exp(-z))$ denote the logistic loss. Recall the empirical risk $\hat{\mathcal{R}}$ for logistic regression (as presented in lecture 3):

$$\hat{\mathcal{R}}_{\log}(\mathbf{w}) = \frac{1}{n} \sum_{k=1}^n \ln(1 + \exp(-y_k \mathbf{w}^\top \mathbf{x}_k)).$$

As in lecture, the corresponding empirical risk $\hat{\mathcal{R}}_{\log}$ can not in general be minimized by solving for \mathbf{w} in $\nabla \hat{\mathcal{R}}_{\log}(\mathbf{w}) = 0$, but here we will consider a special case where it is possible. Specifically, consider the setting of part (a) of the previous problem: the training examples $(\mathbf{x}_i)_{i=1}^n$ are the standard basis vectors \mathbf{e}_i of \mathbb{R}^d , each repeated respectively $n_i > 0$ times; one difference is that now $y_{i_j} \in \{+1, -1\}$, rather than being arbitrary elements of \mathbb{R} .

Crucially, throughout this problem, for each i suppose that $n_i \geq 2$, and there exist at least one positive and one negative label.

- (a) Recall that the logistic loss suggests a conditional probability model that the label of a given example is $+1$: for any \mathbf{w} and any \mathbf{x} , the corresponding conditional model $\hat{p}_{\mathbf{w}}$ is

$$\hat{p}_{\mathbf{w}}(\mathbf{x}) := \frac{1}{1 + \exp(-\mathbf{w}^\top \mathbf{x})}.$$

Show that in the above setting, consider a given i ,

$$\hat{p}_{\hat{\mathbf{w}}}(\mathbf{e}_i) := \frac{|\{j : y_{i_j} = +1\}|}{n_i},$$

where $\hat{\mathbf{w}}$ denotes the optimal solution to $\hat{\mathcal{R}}_{\log}$ (which exists in this setting).

For full points, you must include and justify a derivation.

- (b) Now suppose that for some k , $n_k \rightarrow \infty$; specifically, there is only one positive pair $(\mathbf{e}_k, +1)$, and $n_k - 1$ negative pairs $(\mathbf{e}_k, -1)$. correspondingly, let $\hat{\mathbf{w}}_n$ denote the optimal solution for each sample size.

Determine and formally prove the limiting behavior of $\|\hat{\mathbf{w}}_n\|$ and $\hat{p}_{\hat{\mathbf{w}}_n}(\mathbf{e}_k)$.

- (c) Now consider the case of a fixed finite n , but let's violate our assumption above regarding $n_i \geq 2$: specifically, suppose there exists \mathbf{e}_i with $n_i = 1$.

We can no longer compute an optimal $\hat{\mathbf{w}} \in \mathbb{R}^d$ (why not?), but based on the preceding reasoning, we can make a guess for $\hat{p}_{\hat{\mathbf{w}}}$.

What do you think it is? And, as a consequence, do you think $\hat{p}_{\hat{\mathbf{w}}}$ represents a reasonable notion of "prediction confidence"?

Grading note: Any answer to this problem which consists of at least one complete sentence will receive full credit. That said, as the use of confidences in this form is standard in machine learning, please consider spending a bit of time.

Solution.

(a)

$$\begin{aligned}
\nabla \widehat{\mathcal{R}}_{\log}(\mathbf{w}) &= -\frac{1}{n} \sum_{i=1}^n \frac{y_i}{1 + e^{y_i \mathbf{w}^T \mathbf{x}_i}} \mathbf{x}_i \\
&= -\frac{1}{n} \sum_{i=1}^d \sum_{j=1}^{n_i} \frac{y_{i_j}}{1 + e^{y_{i_j} \mathbf{w}^T \mathbf{e}_i}} \mathbf{e}_i \\
&= -\frac{1}{n} \sum_{i=1}^d \sum_{j=1}^{n_i} \frac{y_{i_j}}{1 + e^{y_{i_j} \mathbf{w}_i}} \mathbf{e}_i \\
&= -\frac{1}{n} \begin{bmatrix} \sum_{j=1}^{n_1} \frac{y_{1_j}}{1 + e^{y_{1_j} \mathbf{w}_1}} \\ \vdots \\ \sum_{j=1}^{n_d} \frac{y_{d_j}}{1 + e^{y_{d_j} \mathbf{w}_d}} \end{bmatrix}
\end{aligned}$$

So if $\hat{\mathbf{w}}$ is the optimal solution, we must have $\nabla \widehat{\mathcal{R}}_{\log}(\hat{\mathbf{w}}) = \vec{0}$, for each n_i that is:

$$\sum_{j=1}^{n_i} \frac{y_{i_j}}{1 + e^{y_{i_j} \hat{\mathbf{w}}_i}} = 0$$

For simplicity, let $\begin{cases} a_i &:= |\{j : y_{i_j} = +1\}| \\ b_i &:= |\{j : y_{i_j} = -1\}| \end{cases}$, so $a_i + b_i = n_i$, and:

$$\begin{aligned}
\sum_{j=1}^{n_i} \frac{y_{i_j}}{1 + e^{y_{i_j} \hat{\mathbf{w}}_i}} &= a_i \frac{1}{1 + e^{\hat{\mathbf{w}}_i}} + b_i \frac{-1}{1 + e^{-\hat{\mathbf{w}}_i}} = \frac{a_i - b_i e^{\hat{\mathbf{w}}_i}}{1 + e^{\hat{\mathbf{w}}_i}} = 0 \Rightarrow \\
e^{\hat{\mathbf{w}}_i} &= \frac{a_i}{b_i}
\end{aligned}$$

Because $\hat{p}_{\mathbf{w}}(\mathbf{x}) := \frac{1}{1 + \exp(-\mathbf{w}^T \mathbf{x})}$, we have:

$$\begin{aligned}
\hat{p}_{\hat{\mathbf{w}}}(\mathbf{e}_i) &= \frac{1}{1 + e^{-\hat{\mathbf{w}}_i}} \\
&= \frac{1}{1 + \frac{b_i}{a_i}} \\
&= \frac{a_i}{a_i + b_i} \\
&= \frac{|\{j : y_{i_j} = +1\}|}{n_i}
\end{aligned}$$

(b)

$$\begin{cases} \lim_{n_k \rightarrow \infty} \|\hat{\mathbf{w}}_n\| &= 0 \\ \lim_{n_k \rightarrow \infty} \hat{p}_{\hat{\mathbf{w}}_n}(\mathbf{e}_k) &= 0 \end{cases}$$

Because in this case, $\begin{cases} a_k = 1 \\ b_k \rightarrow \infty \end{cases} \Rightarrow \begin{cases} \lim_{n_k \rightarrow \infty} \hat{w}_k = \lim_{n_k \rightarrow \infty} \ln \frac{a_k}{b_k} = \lim_{n_k \rightarrow \infty} \ln \frac{1}{n_k - 1} = 0 \Rightarrow \lim_{n_k \rightarrow \infty} \|\hat{\mathbf{w}}_n\| = 0 \\ \lim_{n_k \rightarrow \infty} \hat{p}_{\hat{\mathbf{w}}_n}(\mathbf{e}_k) = \lim_{n_k \rightarrow \infty} \frac{a_k}{n_k} = \lim_{n_k \rightarrow \infty} \frac{1}{n_k} = 0 \end{cases}$

- (c) i. We can no longer compute an optimal $\hat{\mathbf{w}} \in \mathbb{R}^d$ because $\nabla \widehat{\mathcal{R}}_{\log}(\mathbf{w})$ will never reach $\vec{0}$.
ii. I think in this case when $n_k = 1$, $\hat{p}_{\hat{\mathbf{w}}}(\mathbf{e}_k)$ becomes a binary indicator of whether $y_k = +1$, indeed it does represent a reasonable notion of “prediction confidence”, because if the only feature suggests one result, then we have to believe it with 100% confidence.

3. Linear Regression.

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

- (a) Implement linear regression using gradient descent in the `linear_gd(X, Y, lrate, num_iter)` function of `hw1.py`. You are given as input a training set \mathbf{X} as an $n \times d$ tensor, training labels \mathbf{Y} as an $n \times 1$ tensor, a learning rate `lrate`, and the number of iterations of gradient descent to run `num_iter`. Using gradient descent, find parameters \mathbf{w} that minimize the empirical risk $\hat{\mathcal{R}}(\mathbf{w})$. Use $\mathbf{w} = 0$ as your initial parameters, and return your final \mathbf{w} as output. Prepend a column of ones to \mathbf{X} in order to accommodate a bias term in \mathbf{w} .

Library routines: `torch.matmul` (`@`), `torch.tensor.shape`, `torch.tensor.t`, `torch.cat`, `torch.ones`, `torch.zeros`, `torch.reshape`.

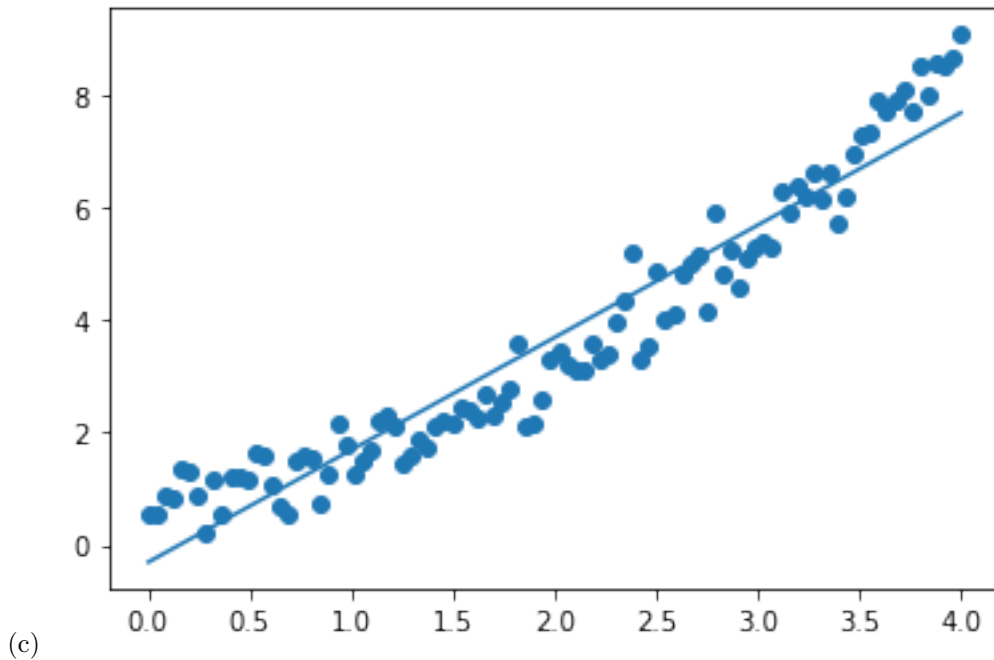
- (b) Implement linear regression by using the pseudoinverse to solve for \mathbf{w} in the `linear_normal(X, Y)` function of `hw1.py`. You are given a training set \mathbf{X} as an $n \times d$ tensor and training labels \mathbf{Y} as an $n \times 1$ tensor. Return your parameters \mathbf{w} as output. As before, make sure to accommodate a bias term by prepending ones to the training examples \mathbf{X} .

Library routines: `torch.matmul` (`@`), `torch.cat`, `torch.ones`, `torch.pinv`.

- (c) Implement the `plot_linear()` function in `hw1.py`. Use the provided function `hw1_utils.load_reg_data()` to generate a training set \mathbf{X} and training labels \mathbf{Y} . Plot the curve generated by `linear_normal()` along with the points from the data set. Return the plot as output. Include the plot in your written submission.

Library routines: `torch.matmul` (`@`), `torch.cat`, `torch.ones`, `plt.plot`, `plt.scatter`, `plt.show`, `plt.gcf` where `plt` refers to the `matplotlib.pyplot` library.

Solution.



4. Polynomial Regression.

In Problem 3 you constructed a linear model $\mathbf{w}^\top \mathbf{x} = \sum_{i=1}^d x_i w_i$. 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 $f_{\mathbf{w}}$ with parameters

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

defined by

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

Warning: If the computational complexity of your implementation is high, it may crash the autograder (try to optimize your algorithm if it does)!

- (a) Given a 3-dimensional feature vector $\mathbf{x} = (x_1, x_2, x_3)$, completely write out the quadratic expanded feature vector $\phi(\mathbf{x})$.
- (b) Implement the `poly_gd()` function in `hw1.py`. The input is in the same format as it was in Problem 3. Implement gradient descent on this training set with \mathbf{w} initialized to 0. Return \mathbf{w} as the output with terms in this exact order: bias, linear, then quadratic. For example, if $d = 3$ then you would return $(w_0, w_{01}, w_{02}, w_{03}, w_{11}, w_{12}, w_{13}, w_{22}, w_{23}, w_{33})$.

Library routines: `torch.cat`, `torch.ones`, `torch.zeros`, `torch.stack`.

Hint: You will want to prepend a column of ones to \mathbf{X} , and append to \mathbf{X} the squared features in the specified order. You can generate the squared features in the correct order (This is important! The order of the polynomial features matters for your answer to match the correct answer on GradeScope. Check the order in the problem definition above.) using a nested for loop. We don't want duplicates (e.g., $x_0 x_1$ and $x_1 x_0$ should not both be included; we should only include $x_0 x_1$ in the quadratic case).

- (c) Implement the `poly_normal` function in `hw1.py`. You are given the same data set as from part (b), but this time determine \mathbf{w} by using the pseudoinverse. Return \mathbf{w} in the same order as in part (b).

Library routines: `torch.pinverse`.

Hint: You will still need to transform the matrix \mathbf{X} in the same way as in part (b).

- (d) Implement the `plot_poly()` function in `hw1.py`. Use the provided function `hw1_utils.load_reg_data()` to generate a training set \mathbf{X} and training labels \mathbf{Y} . Plot the curve generated by `poly_normal()` along with the points from the data set. Return the plot as output and include it in your written submission. Compare and contrast this plot with the plot from Problem 3. Which model appears to approximate the data better? Justify your answer.

Library routines: `plt.plot`, `plt.scatter`, `plt.show`, `plt.gcf`.

- (e) The Minsky-Papert XOR problem is a classification problem with data set:

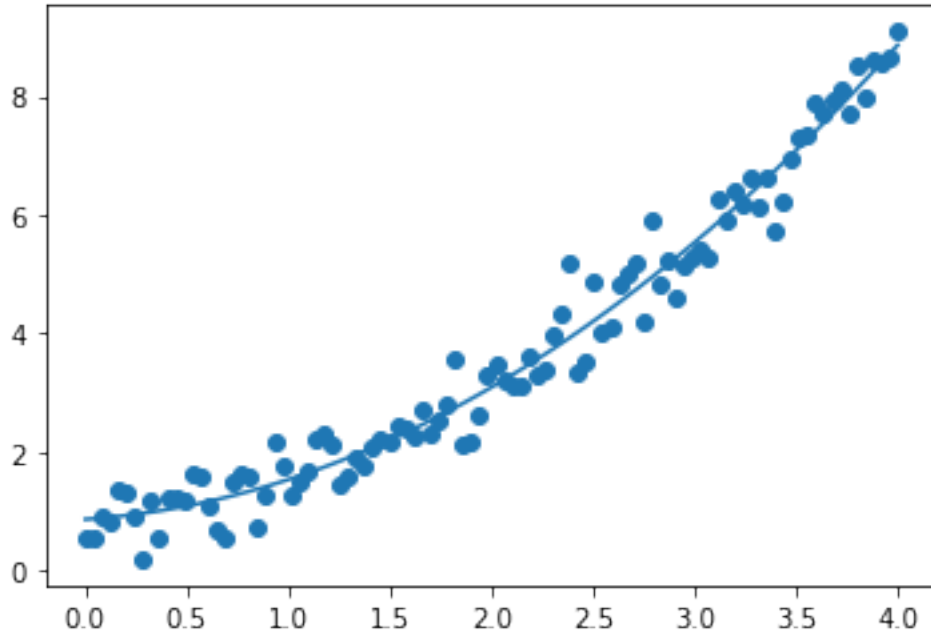
$$X = \{(-1, +1), (+1, -1), (-1, -1), (+1, +1)\}$$

where the label for a given point (x_1, x_2) is given by its product $x_1 x_2$. For example, the point $(-1, +1)$ would be given label $y = (-1)(1) = -1$. Implement the `poly_xor()` function in `hw1.py`. In this function you will load the XOR data set by calling the `hw1_utils.load_xor_data()` function, and then apply the `linear_normal()` and `poly_normal()` functions to generate predictions for the XOR points. Include a plot of contour lines that show how each model classifies points in your written submission. Return the predictions for both the linear model and the polynomial model and use `contour_plot()` in `hw1_utils.py` to help with the plot. Do both models correctly classify all points? (Note that red corresponds to larger values and blue to smaller values when using `contour_plot` with the "coolwarm" colormap).

Hint: A "Contour plot" is a way to represent a 3-dimensional surface in a 2-D figure. In this example, the data points are pinned to the figure with their features (x_1, x_2) as the coordinates in 2-D space (e.g., x and y axis); the third dimension (e.g., the predictions of the data points) is labeled on the points in the figure. The lines or curves that link the grid points with the same predictions together are called the "contours". See `contour_plot()` in `hw1_utils.py` for details.

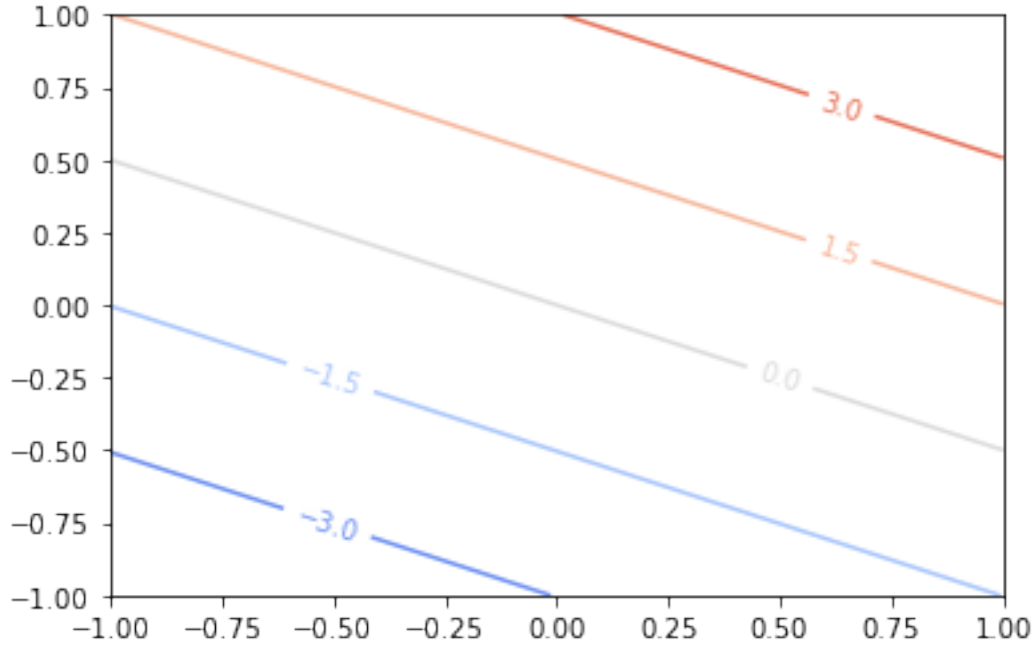
Solution.

(a) $\phi(x) = [1 \quad x_1 \quad x_2 \quad x_3 \quad x_1^2 \quad x_1x_2 \quad x_1x_3 \quad x_2^2 \quad x_2x_3 \quad x_3^2]$

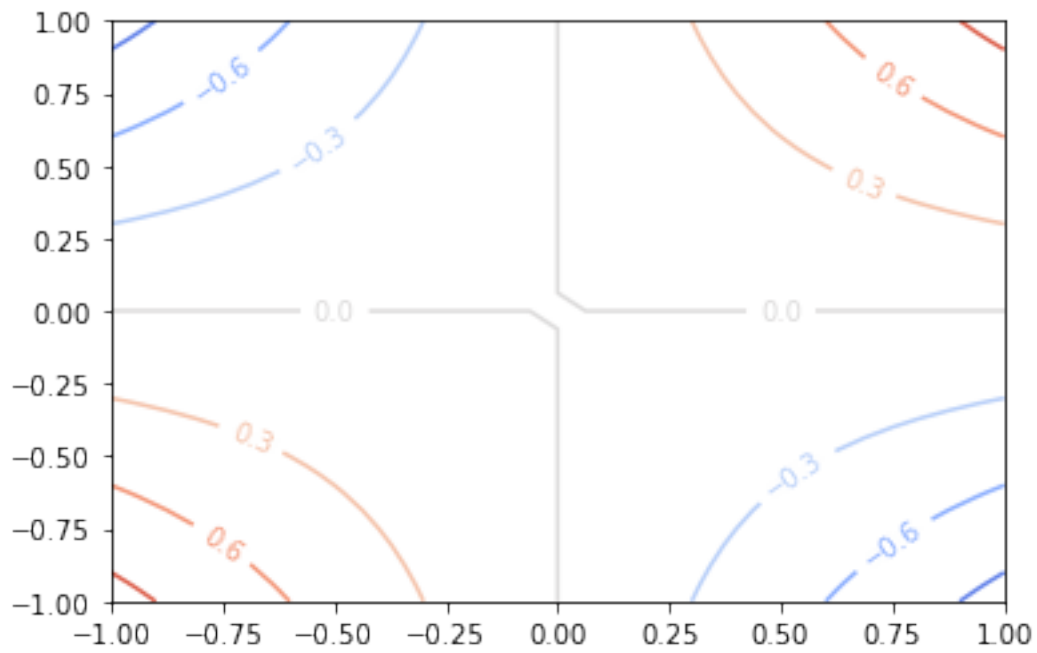


(d)

Clearly, regression with the quadratic feature better fits the data, because the empirical risk is smaller.



(e)



Naive linear regression cannot classify all points correctly, but regression with quadratic feature can.

5. Logistic Regression.

Recall the empirical risk $\hat{\mathcal{R}}$ for logistic regression (as presented in lecture 3):

$$\hat{\mathcal{R}}_{\log}(\mathbf{w}) = \frac{1}{n} \sum_{i=1}^n \ln(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i)).$$

Here you will minimize this risk using gradient descent.

- In your written submission, derive the gradient descent update rule for this empirical risk by taking the gradient. Write your answer in terms of the learning rate η , previous parameters \mathbf{w} , new parameters \mathbf{w}' , number of examples n , and training examples \mathbf{x}_i . Show all of your steps.
- Implement the `logistic()` function in `hw1.py`. You are given as input a training set `X`, training labels `Y`, a learning rate `lr`, and number of gradient updates `num_iter`. Implement gradient descent to find parameters \mathbf{w} that minimize the empirical risk $\hat{\mathcal{R}}_{\log}(\mathbf{w})$. Perform gradient descent for `num_iter` updates with a learning rate of `lr`, initializing $\mathbf{w} = 0$ and returning \mathbf{w} as output. Don't forget to prepend `X` with a column of ones.

Library routines: `torch.matmul (@)`, `torch.tensor.t`, `torch.exp`.

- Implement the `logistic_vs_ols()` function in `hw1.py`. Use `hw1_utils.load_logistic_data()` to generate a training set `X` and training labels `Y`. Run `logistic(X,Y)` from part (b) taking `X` and `Y` as input to obtain parameters \mathbf{w} (use the defaults for `num_iter` and `lr`). Also run `linear_gd(X,Y)` from Problem 3 to obtain parameters \mathbf{w} . Plot the decision boundaries for your logistic regression and least squares models along with the data `X`. Which model appears to classify the data better? Explain why you believe your choice is the better classifier for this problem.

Library routines: `torch.linspace`, `plt.scatter`, `plt.plot`, `plt.show`, `plt.gcf`.

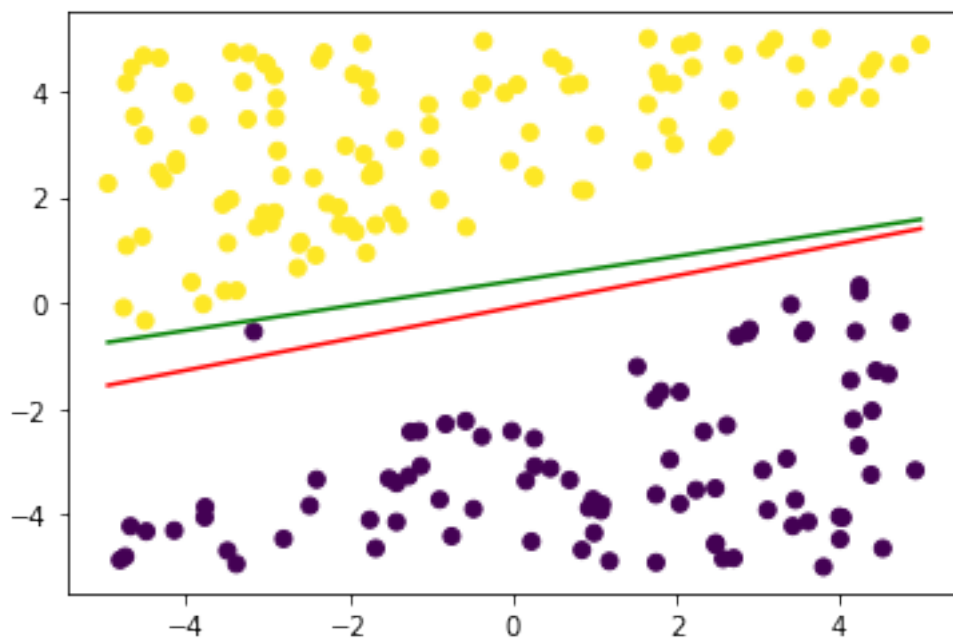
Hints:

- The positive and negative points are guaranteed to be linearly separable (though an algorithm may or may not find the optimal line to separate them).
- The “decision boundary” in the problem description refers to the set of points \mathbf{x} such that $\mathbf{w}^\top \mathbf{x} = 0$ for the chosen predictor. In this case, it suffices to plot the corresponding line.
- In order to make the two models significantly different, we recommend that you train the logistic regression with a large `num_iter` (e.g., 1,000,000 or even larger).

Solution.

(a)

$$\begin{aligned} \hat{\mathcal{R}}_{\log}(\mathbf{w}) &= \frac{1}{n} \sum_{i=1}^n \ln(1 + \exp(-y_i \mathbf{w}^\top \mathbf{x}_i)) \\ &\Downarrow \\ \nabla_{\mathbf{w}} \hat{\mathcal{R}}_{\log}(\mathbf{w}) &= \frac{1}{n} \sum_{i=1}^n \frac{-e^{-y_i \mathbf{w}^\top \mathbf{x}_i}}{1 + e^{-y_i \mathbf{w}^\top \mathbf{x}_i}} (y_i \mathbf{x}_i) \\ &= \frac{1}{n} \sum_{i=1}^n -\frac{y_i}{e^{y_i \mathbf{w}^\top \mathbf{x}_i} + 1} \mathbf{x}_i \\ &\Downarrow \\ \mathbf{w}' &= \mathbf{w} - \eta \nabla_{\mathbf{w}} \hat{\mathcal{R}}_{\log}(\mathbf{w}) \\ &= \mathbf{w} - \eta \frac{1}{n} \sum_{i=1}^n -\frac{y_i}{e^{y_i \mathbf{w}^\top \mathbf{x}_i} + 1} \mathbf{x}_i \\ &= \mathbf{w} + \frac{\eta}{n} \sum_{i=1}^n \frac{y_i}{e^{y_i \mathbf{w}^\top \mathbf{x}_i} + 1} \mathbf{x}_i \end{aligned}$$



(c)

The green line denotes the decision boundary of Logistic regression, which seems to perform well in this example, but might not be what the true model behind the data suggests. (maybe there are outliers)