

CM146, Winter 2020  
 Problem Set 2: Perceptron and regression  
 Due February 8, 2020 at 11:59 pm

## 1 Perceptron [2 pts]

Design (specify  $\theta$  for) a two-input perceptron (with an additional bias or offset term) that computes the following boolean functions. Assume  $T = 1$  and  $F = -1$ . If a valid perceptron exists, show that it is not unique by designing another valid perceptron (with a different hyperplane, not simply through normalization). If no perceptron exists, state why. **Solution:** For  $\mathbf{x} = (1, x_1, x_2)$ , a valid perceptron defined by  $\theta = (\theta_0, \theta_1, \theta_2)$  requires that  $y = \text{sign}(\theta^T \mathbf{x})$ .

| $x_1$ | $x_2$ | $y$ |
|-------|-------|-----|
| -1    | -1    | -1  |
| -1    | +1    | -1  |
| +1    | -1    | -1  |
| +1    | +1    | +1  |

(a) OR

| $x_1$ | $x_2$ | $y$ |
|-------|-------|-----|
| -1    | -1    | -1  |
| -1    | +1    | +1  |
| +1    | -1    | +1  |
| +1    | +1    | -1  |

(b) XOR

(a) (1 pts) OR **Solution:** One possible solution is  $\theta = (-1, 1, 1)$ . Another possible solution is  $\theta = (-1, 0.5, 0.6)$ .

(b) (1 pts) XOR **Solution:** No solution exists because the data is not linearly separable.

## 2 Logistic Regression [10 pts]

Consider the objective function that we minimize in logistic regression:

$$J(\theta) = - \sum_{n=1}^N [y_n \log h_{\theta}(\mathbf{x}_n) + (1 - y_n) \log (1 - h_{\theta}(\mathbf{x}_n))]$$

**Solution:** Recall that we have  $\sigma'(z) = \sigma(z)(1 - \sigma(z))$ , and thus for  $h_{\theta}(\mathbf{x}) = \sigma(\theta^T \mathbf{x})$ , we have:

$$\begin{aligned}
 \frac{\partial h_{\theta}(\mathbf{x})}{\partial \theta_k} &= \frac{\partial \sigma(\theta^T \mathbf{x})}{\partial \theta_k} \\
 &= \frac{\partial \sigma(\theta^T \mathbf{x})}{\partial \theta^T \mathbf{x}} \frac{\partial (\theta^T \mathbf{x})}{\partial \theta_k} \\
 &= \sigma(\theta^T \mathbf{x})(1 - \sigma(\theta^T \mathbf{x})) \frac{\partial (\sum_j \theta_j x_j)}{\partial \theta_k} \\
 &= \sigma(\theta^T \mathbf{x})(1 - \sigma(\theta^T \mathbf{x})) x_k \\
 &= h_{\theta}(\mathbf{x})(1 - h_{\theta}(\mathbf{x})) x_k
 \end{aligned} \tag{1}$$

---

Parts of this assignment are adapted from course material by Andrew Ng (Stanford), Jenna Wiens (UMich) and Jessica Wu (Harvey Mudd).

This latter fact is very useful to make the following derivations.

- (a) **(2 pts)** Find the partial derivatives  $\frac{\partial J}{\partial \theta_j}$ . **Solution:**

$$\begin{aligned}
 \frac{\partial J}{\partial \theta_j} &= \frac{\partial \left( - \sum_{n=1}^N [y_n \log h_{\boldsymbol{\theta}}(\mathbf{x}_n) + (1 - y_n) \log (1 - h_{\boldsymbol{\theta}}(\mathbf{x}_n))] \right)}{\partial \theta_j} \\
 &= - \sum_{n=1}^N \frac{\partial [y_n \log h_{\boldsymbol{\theta}}(\mathbf{x}_n) + (1 - y_n) \log (1 - h_{\boldsymbol{\theta}}(\mathbf{x}_n))]}{\partial \theta_j} \\
 &= - \sum_{n=1}^N \left( \frac{\partial [y_n \log h_{\boldsymbol{\theta}}(\mathbf{x}_n)]}{\partial \theta_j} + \frac{\partial [(1 - y_n) \log (1 - h_{\boldsymbol{\theta}}(\mathbf{x}_n))]}{\partial \theta_j} \right) \\
 &= - \sum_{n=1}^N \left( y_n \frac{\partial \log h_{\boldsymbol{\theta}}(\mathbf{x}_n)}{\partial \theta_j} + (1 - y_n) \frac{\partial \log (1 - h_{\boldsymbol{\theta}}(\mathbf{x}_n))}{\partial \theta_j} \right)
 \end{aligned}$$

To compute the partial derivatives, we use the chain rule and identity 1

$$\begin{aligned}
 \frac{\partial \log h_{\boldsymbol{\theta}}(\mathbf{x}_n)}{\partial \theta_j} &= \frac{\partial \log h_{\boldsymbol{\theta}}(\mathbf{x}_n)}{\partial h_{\boldsymbol{\theta}}(\mathbf{x}_n)} \frac{\partial h_{\boldsymbol{\theta}}(\mathbf{x}_n)}{\partial \theta_j} \\
 &= \frac{1}{h_{\boldsymbol{\theta}}(\mathbf{x}_n)} h_{\boldsymbol{\theta}}(\mathbf{x}_n) (1 - h_{\boldsymbol{\theta}}(\mathbf{x}_n)) x_j \\
 &= (1 - h_{\boldsymbol{\theta}}(\mathbf{x}_n)) x_j
 \end{aligned} \tag{2}$$

Similarly

$$\frac{\partial \log (1 - h_{\boldsymbol{\theta}}(\mathbf{x}_n))}{\partial \theta_j} = -h_{\boldsymbol{\theta}}(\mathbf{x}_n) x_j \tag{3}$$

Using Equations 2 and 3

$$\begin{aligned}
 \frac{\partial J}{\partial \theta_j} &= - \sum_{n=1}^N (y_n (1 - h_{\boldsymbol{\theta}}(\mathbf{x}_n)) x_j - (1 - y_n) h_{\boldsymbol{\theta}}(\mathbf{x}_n) x_j) \\
 &= - \sum_{n=1}^N (y_n - h_{\boldsymbol{\theta}}(\mathbf{x}_n)) x_{n,j} \\
 &= \sum_{n=1}^N (h_{\boldsymbol{\theta}}(\mathbf{x}_n) - y_n) x_{n,j}
 \end{aligned}$$

- (b) **(3 pts)** Find the partial second derivatives  $\frac{\partial^2 J}{\partial \theta_j \partial \theta_k}$  and show that the Hessian (the matrix  $\mathbf{H}$  of second derivatives with elements  $H_{jk} = \frac{\partial^2 J}{\partial \theta_j \partial \theta_k}$ ) can be written as  $\mathbf{H} = \sum_{n=1}^N h_{\boldsymbol{\theta}}(\mathbf{x}_n) (1 - h_{\boldsymbol{\theta}}(\mathbf{x}_n)) \mathbf{x}_n \mathbf{x}_n^T$ .

**Solution:**

$$\begin{aligned}
\frac{\partial^2 J}{\partial \theta_j \partial \theta_k} &= \frac{\partial}{\partial \theta_j} \left( \frac{\partial J}{\partial \theta_k} \right) \\
&= \frac{\partial}{\partial \theta_j} \left( \sum_{n=1}^N (h_{\theta}(\mathbf{x}_n) - y_n) x_{n,k} \right) \\
&= \sum_{n=1}^N \frac{\partial ((h_{\theta}(\mathbf{x}_n) - y_n) x_{n,k})}{\partial \theta_j} \\
&= \sum_{n=1}^N \frac{\partial (h_{\theta}(\mathbf{x}_n) x_{n,k})}{\partial \theta_j} \\
&= \sum_{n=1}^N \frac{\partial h_{\theta}(\mathbf{x}_n)}{\partial \theta_j} x_{n,k}
\end{aligned}$$

Using identity 1

$$\frac{\partial^2 J}{\partial \theta_j \partial \theta_k} = \sum_{n=1}^N h_{\theta}(\mathbf{x}_n) (1 - h_{\theta}(\mathbf{x}_n)) x_{n,j} x_{n,k} \quad (4)$$

$\mathbf{H}$  is defined as:

$$\mathbf{H} = \sum_{n=1}^N h_{\theta}(\mathbf{x}_n) (1 - h_{\theta}(\mathbf{x}_n)) \mathbf{x}_n \mathbf{x}_n^T$$

Entry  $j, k$  of  $\mathbf{H}$  is given by:

$$\begin{aligned}
H_{j,k} &= \sum_{n=1}^N h_{\theta}(\mathbf{x}_n) (1 - h_{\theta}(\mathbf{x}_n)) (\mathbf{x}_n \mathbf{x}_n^T)_{j,k} \\
&= \sum_{n=1}^N h_{\theta}(\mathbf{x}_n) (1 - h_{\theta}(\mathbf{x}_n)) x_{n,j} x_{n,k} \\
&= \frac{\partial^2 J}{\partial \theta_j \partial \theta_k}, \quad \text{from Equation 4}
\end{aligned}$$

Thus,  $\mathbf{H}$  is the Hessian.

- (c) **(5 pts)** Show that  $J$  is a convex function and therefore has no local minima other than the global one.

*Hint:* A function  $J$  is convex if its Hessian is positive semi-definite (PSD), written  $\mathbf{H} \succeq 0$ . A matrix is PSD if and only if

$$\mathbf{z}^T \mathbf{H} \mathbf{z} \equiv \sum_{j,k} z_j z_k H_{jk} \geq 0.$$

for all real vectors  $\mathbf{z}$ . **Solution:** So we have for the Hessian matrix  $\mathbf{H}$  (using that for  $\mathbf{X} = \mathbf{x} \mathbf{x}^T$  if and only if  $X_{ij} = x_i x_j$ ):

$$\mathbf{H} = \sum_{n=1}^N h_{\theta}(\mathbf{x}_n) (1 - h_{\theta}(\mathbf{x}_n)) \mathbf{x}_n \mathbf{x}_n^T.$$

And to prove  $\mathbf{H}$  is positive semidefinite, we show  $\mathbf{z}^T \mathbf{H} \mathbf{z} \geq 0$  for all real vectors  $\mathbf{z}$ .

$$\begin{aligned} \mathbf{z}^T \mathbf{H} \mathbf{z} &= \mathbf{z}^T \left( \sum_{n=1}^N h_{\theta}(\mathbf{x}_n) (1 - h_{\theta}(\mathbf{x}_n)) \mathbf{x}_n \mathbf{x}_n^T \right) \mathbf{z} \\ &= \sum_{n=1}^N h_{\theta}(\mathbf{x}_n) (1 - h_{\theta}(\mathbf{x}_n)) \mathbf{z}^T \mathbf{x}_n \mathbf{x}_n^T \mathbf{z} \\ &= \sum_{n=1}^N h_{\theta}(\mathbf{x}_n) (1 - h_{\theta}(\mathbf{x}_n)) (\mathbf{z}^T \mathbf{x}_n)^2 \\ &\geq 0 \end{aligned}$$

with the last inequality holding because  $0 \leq h(\mathbf{x}) \leq 1$ , which implies that  $h_{\theta}(\mathbf{x}_n) (1 - h_{\theta}(\mathbf{x}_n)) \geq 0$ , and  $(\mathbf{z}^T \mathbf{x}_n)^2 \geq 0$ .

### 3 Maximum Likelihood Estimation [15 pts]

Suppose we observe the values of  $n$  independent random variables  $X_1, \dots, X_n$  drawn from the same Bernoulli distribution with parameter  $\theta$ <sup>1</sup>. In other words, for each  $X_i$ , we know that

$$P(X_i = 1) = \theta \quad \text{and} \quad P(X_i = 0) = 1 - \theta.$$

Our goal is to estimate the value of  $\theta$  from these observed values of  $X_1$  through  $X_n$ .

For any hypothetical value  $\hat{\theta}$ , we can compute the probability of observing the outcome  $X_1, \dots, X_n$  if the true parameter value  $\theta$  were equal to  $\hat{\theta}$ . This probability of the observed data is often called the *likelihood*, and the function  $L(\theta)$  that maps each  $\theta$  to the corresponding likelihood is called the *likelihood function*. A natural way to estimate the unknown parameter  $\theta$  is to choose the  $\theta$  that maximizes the likelihood function. Formally,

$$\hat{\theta}_{MLE} = \arg \max_{\theta} L(\theta).$$

- (a) **(3 pts)** Write a formula for the likelihood function,  $L(\theta) = P(X_1, \dots, X_n; \theta)$ . Your function should depend on the random variables  $X_1, \dots, X_n$  and the hypothetical parameter  $\theta$ . Does the likelihood function depend on the order in which the random variables are observed ?

**Solution:** Since the  $X_i$  are independent, we have

$$\begin{aligned} L(\theta) &= P_{\theta}(X_1, \dots, X_n) \\ &= \prod_{i=1}^n P_{\theta}(X_i) \\ &= \prod_{i=1}^n \theta^{X_i} (1 - \theta)^{1 - X_i} \\ &= \theta^{\#\{X_i=1\}} (1 - \theta)^{\#\{X_i=0\}}, \end{aligned}$$

---

<sup>1</sup>This is a common assumption for sampling data. So we will denote this assumption as iid, short for Independent and Identically Distributed, meaning that each random variable has the same distribution and is drawn independent of all the other random variables

where  $\#\{\cdot\}$  counts the number of  $X_i$  for which the condition in braces holds true. In the third line, we used the trick  $X_i = \mathbb{I}\{X_i = 1\}$ . The likelihood function does not depend on the order of the data.

Common mistakes:

- For this problem, many people wrote the definition of the likelihood function  $L(\hat{\theta}) = P_{\hat{\theta}}(X_1, \dots, X_n) = \prod_{i=1}^n P_{\hat{\theta}}(X_i)$ . Since you knew the distribution of the  $X_i$ , we expected you to write out an explicit formula.

- (b) **(6 pts)** Since the log function is increasing, the  $\theta$  that maximizes the *log likelihood*  $\ell(\theta) = \log(L(\theta))$  is the same as the  $\theta$  that maximizes the likelihood. Find  $\ell(\theta)$  and its first and second derivatives, and use these to find a closed-form formula for the MLE. **Solution:** Introduce the shorthand  $n_1 = \#\{X_i = 1\}$  and  $n_0 = \#\{X_i = 0\}$ . Using the properties of the log function, we can rewrite  $\ell$  as follows:

$$\begin{aligned}\ell(\theta) &= \log(\theta^{n_1}(1-\theta)^{n_0}) \\ &= n_1 \log(\theta) + n_0 \log(1-\theta).\end{aligned}$$

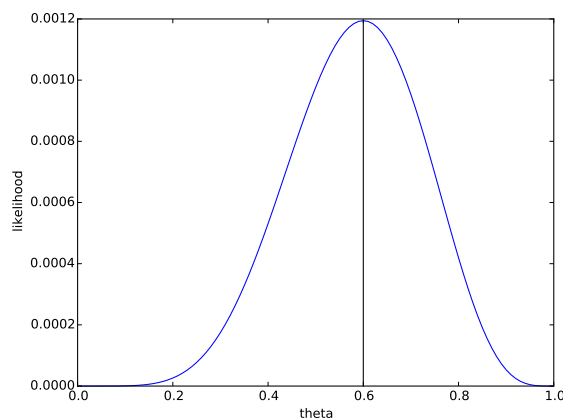
The first and second derivatives of  $\ell$  are given by

$$\ell'(\theta) = \frac{n_1}{\theta} - \frac{n_0}{1-\theta} \quad \text{and} \quad \ell''(\theta) = -\left(\frac{n_1}{\theta^2} + \frac{n_0}{(1-\theta)^2}\right).$$

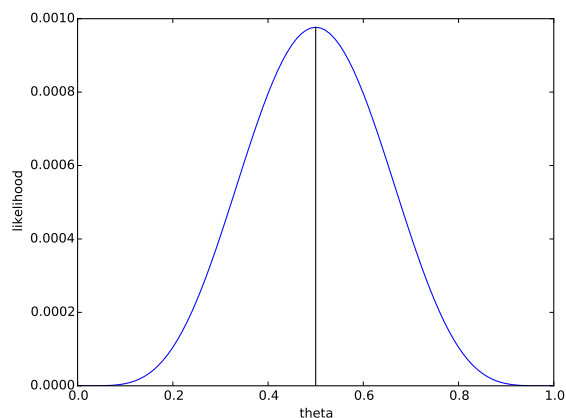
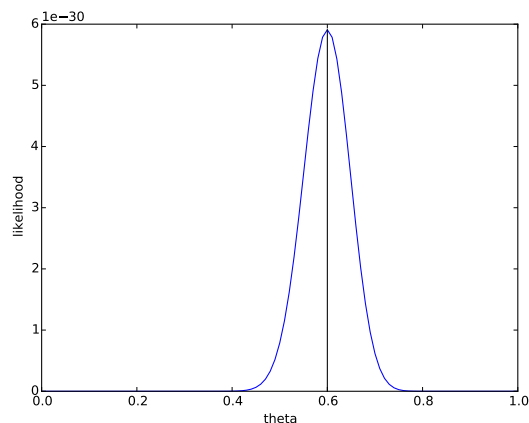
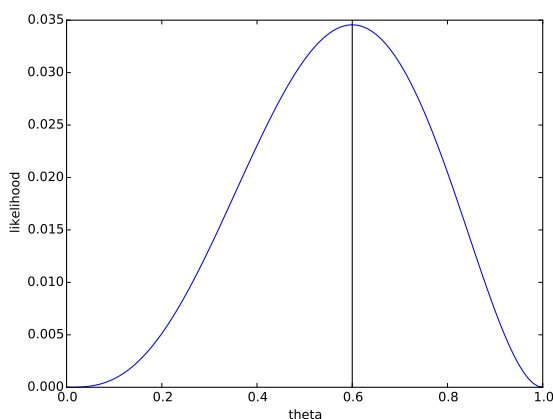
Since the second derivative is always negative, the function  $\ell$  is concave, and we can find its maximizer by solving  $\ell'(\theta) = 0$ . The solution to this equation is obtained by straight-forward algebra and is given by

$$\hat{\theta}_{MLE} = \frac{n_1}{n_1 + n_0}.$$

- (c) **(3 pts)** Suppose that  $n = 10$  and the data set contains six 1s and four 0s. Write a short program `likelihood.py` that plots the likelihood function of this data for each value of  $\hat{\theta}$  in  $\{0, 0.01, 0.02, \dots, 1.0\}$  (use `np.linspace(...)` to generate this spacing). For the plot, the x-axis should be  $\theta$  and the y-axis  $L(\theta)$ . Scale your y-axis so that you can see some variation in its value. Include the plot in your writeup (there is no need to submit your code). Estimate  $\hat{\theta}_{MLE}$  by marking on the x-axis the value of  $\hat{\theta}$  that maximizes the likelihood. Does the answer agree with the closed form answer? **Solution:**



- (d) **(3 pts)** Create three more likelihood plots: one where  $n = 5$  and the data set contains three 1s and two 0s; one where  $n = 100$  and the data set contains sixty 1s and forty 0s; and one where  $n = 10$  and there are five 1s and five 0s. Include these plots in your writeup, and describe how the likelihood functions and maximum likelihood estimates compare for the different data sets. **Solution:**



The MLE is equal to the proportion of 1s observed in the data, so for the first three plots the MLE is always at 0.6 and for the last plot it is at 0.5. As the number of samples  $n$  increases, the likelihood function gets more peaked at its maximum value, and the values it takes on decrease.

## 4 Implementation: Polynomial Regression [20 pts]

In this exercise, you will work through linear and polynomial regression. Our data consists of inputs  $x_n \in \mathbb{R}$  and outputs  $y_n \in \mathbb{R}, n \in \{1, \dots, N\}$ , which are related through a target function  $y = f(x)$ . Your goal is to learn a linear predictor  $h_\theta(x)$  that best approximates  $f(x)$ . But this time, rather than using `scikit-learn`, we will further open the “black-box”, and you will implement the regression model!

---

code and data

- code : `regression.py`
  - data : `regression_train.csv, regression_test.csv`
- 

This is likely the first time that many of you are working with `numpy` and matrix operations within a programming environment. For the uninitiated, you may find it useful to work through a `numpy` tutorial first.<sup>2</sup> Here are some things to keep in mind as you complete this problem:

- If you are seeing many errors at runtime, inspect your matrix operations to make sure that you are adding and multiplying matrices of compatible dimensions. Printing the dimensions of variables with the `X.shape` command will help you debug.
- When working with `numpy` arrays, remember that `numpy` interprets the `*` operator as element-wise multiplication. This is a common source of size incompatibility errors. If you want matrix multiplication, you need to use the `dot` function in Python. For example, `A*B` does element-wise multiplication while `dot(A,B)` does a matrix multiply.
- Be careful when handling `numpy` vectors (rank-1 arrays): the vector shapes  $1 \times N$ ,  $N \times 1$ , and  $N$  are all different things. For these dimensions, we follow the the conventions of `scikit-learn`’s `LinearRegression` class<sup>3</sup>. Most importantly, unless otherwise indicated (in the code documentation), both column and row vectors are rank-1 arrays of shape  $N$ , not rank-2 arrays of shape  $N \times 1$  or shape  $1 \times N$ .

### Visualization [1 pts]

As we learned last week, it is often useful to understand the data through visualizations. For this data set, you can use a scatter plot to visualize the data since it has only two properties to plot ( $x$  and  $y$ ).

- (a) **(1 pts)** Visualize the training and test data using the `plot_data(...)` function. What do you observe? For example, can you make an educated guess on the effectiveness of linear regression in predicting the data?

**Solution:** Neither data set shows a linear relationship. We would expect simple linear regression to be a poor predictor, but polynomial regression might work.

### Linear Regression [12 pts]

---

<sup>2</sup>Try out SciPy’s tutorial (<https://docs.scipy.org/doc/numpy/user/quickstart.html>), or use your favorite search engine to find an alternative.

<sup>3</sup>[http://scikit-learn.org/stable/modules/generated/sklearn.linear\\_model.LinearRegression.html](http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LinearRegression.html)

Recall that linear regression attempts to minimize the objective function

$$J(\boldsymbol{\theta}) = \sum_{n=1}^N (h_{\boldsymbol{\theta}}(\mathbf{x}_n) - y_n)^2.$$

In this problem, we will use the matrix-vector form where

$$\mathbf{y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_N \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} \mathbf{x}_1^T \\ \mathbf{x}_2^T \\ \vdots \\ \mathbf{x}_N^T \end{pmatrix}, \quad \boldsymbol{\theta} = \begin{pmatrix} \theta_0 \\ \theta_1 \\ \theta_2 \\ \vdots \\ \theta_D \end{pmatrix}$$

and each instance  $\mathbf{x}_n = (1, x_{n,1}, \dots, x_{n,D})^T$ .

In this instance, the number of input features  $D = 1$ .

Rather than working with this fully generalized, multivariate case, let us start by considering a simple linear regression model:

$$h_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^T \mathbf{x} = \theta_0 + \theta_1 x_1$$

`regression.py` contains the skeleton code for the class `PolynomialRegression`. Objects of this class can be instantiated as `model = PolynomialRegression(m)` where  $m$  is the degree of the polynomial feature vector where the feature vector for instance  $n$ ,  $(1, x_{n,1}, x_{n,1}^2, \dots, x_{n,1}^m)^T$ . Setting  $m = 1$  instantiates an object where the feature vector for instance  $n$ ,  $(1, x_{n,1})^T$ .

- (b) **(1 pts)** Note that to take into account the intercept term ( $\theta_0$ ), we can add an additional “feature” to each instance and set it to one, e.g.  $x_{i,0} = 1$ . This is equivalent to adding an additional first column to  $\mathbf{X}$  and setting it to all ones.

Modify `PolynomialRegression.generate_polynomial_features(...)` to create the matrix  $\mathbf{X}$  for a simple linear model. **Solution:**

- (c) **(1 pts)** Before tackling the harder problem of training the regression model, complete `PolynomialRegression.predict(...)` to predict  $\mathbf{y}$  from  $\mathbf{X}$  and  $\boldsymbol{\theta}$ . **Solution:**

- (d) **(5 pts)** One way to solve linear regression is through gradient descent (GD).

Recall that the parameters of our model are the  $\theta_j$  values. These are the values we will adjust to minimize  $J(\boldsymbol{\theta})$ . In gradient descent, each iteration performs the update

$$\theta_j \leftarrow \theta_j - 2\alpha \sum_{n=1}^N (h_{\boldsymbol{\theta}}(\mathbf{x}_n) - y_n) x_{n,j} \quad (\text{simultaneously update } \theta_j \text{ for all } j).$$

With each step of gradient descent, we expect our updated parameters  $\theta_j$  to come closer to the parameters that will achieve the lowest value of  $J(\boldsymbol{\theta})$ .

- **(1 pts)** As we perform gradient descent, it is helpful to monitor the convergence by computing the cost, *i.e.*, the value of the objective function  $J$ . Complete `PolynomialRegression.cost(...)` to calculate  $J(\boldsymbol{\theta})$ .

If you have implemented everything correctly, then the following code snippet should return 40.234.



```

train_data = load_data('regression_train.csv')
model = PolynomialRegression()
model.coef_ = np.zeros(2)
model.cost(train_data.X, train_data.y)

```

**Solution:** See solutions code.

- **(2 pts)** Next, implement the gradient descent step in `PolynomialRegression.fit_GD(...)`. The loop structure has been written for you, and you only need to supply the updates to  $\theta$  and the new predictions  $\hat{y} = h_{\theta}(x)$  within each iteration.

We will use the following specifications for the gradient descent algorithm:

- We run the algorithm for 10,000 iterations.
- We terminate the algorithm earlier if the value of the objective function is unchanged across consecutive iterations.
- We will use a fixed step size.

**Solution:**

- **(2 pts)** So far, you have used a default learning rate (or step size) of  $\eta = 0.01$ . Try different  $\eta = 10^{-4}, 10^{-3}, 10^{-2}, 0.0407$ , and make a table of the coefficients, number of iterations until convergence (this number will be 10,000 if the algorithm did not converge in a smaller number of iterations) and the final value of the objective function. How do the coefficients compare? How quickly does each algorithm converge? **Solution:** See Table 1. If you choose a learning rate that is too large, the algorithm diverges. If you choose a learning rate that is too small, the coefficients are closer to the closed-form solution, but the algorithm takes a very long time to converge.

| $\eta$    | $\theta$                                      | iterations | cost                  |
|-----------|---|------------|-----------------------|
| $10^{-4}$ | [2.2704; -2.4606]                             | 10,000     | 4.0864                |
| $10^{-3}$ | [2.4464; -2.8163]                             | 7,077      | 3.91258               |
| $10^{-2}$ | [2.4464; -2.8163]                             | 766        | 3.91258               |
| 0.0407    | $[-9.4 \times 10^{18}; -4.65 \times 10^{18}]$ | 10,000     | $2.71 \times 10^{39}$ |

Table 1: Optimal coefficients, number of iterations and cost function for varying step sizes.

- (e) **(4 pts)** In class, we learned that the closed-form solution to linear regression is

$$\theta = (X^T X)^{-1} X^T y.$$

Using this formula, you will get an exact solution in one calculation: there is no “loop until convergence” like in gradient descent.

- **(2 pts)** Implement the closed-form solution `PolynomialRegression.fit(...)`. **Solution:**
- **(2 pts)** What is the closed-form solution? How do the coefficients and the cost compare to those obtained by GD? How quickly does the algorithm run compared to GD? **Solution:** The coefficients are [2.4464; -2.8164] and the cost is 3.912576. For GD with smaller step sizes, the coefficients are closer to the closed-form solution. In this particular instance, the closed-form solution is faster than gradient descent.

- (f) **(1 pts)** Finally, set a learning rate  $\eta$  for GD that is a function of  $k$  (the number of iterations) (use  $\eta_k = \frac{1}{1+k}$ ) and converges to the same solution yielded by the closed-form optimization (minus possible rounding errors). Update `PolynomialRegression.fit_GD(...)` with your proposed learning rate. How long does it take the algorithm to converge with your proposed learning rate? **Solution:**  $\eta_k = \frac{1}{1+k}$ , where  $k$  is the number of iterations of the outer loop converges to the same solution yielded by the closed form optimization. It takes 1511 iterations for the algorithm to converge to  $[2.4464; -2.8163]$ , with this learning rate. There are multiple other possible solutions.

## Polynomial Regression[7 pts]

Now let us consider the more complicated case of polynomial regression, where our hypothesis is

$$h_{\theta}(\mathbf{x}) = \theta^T \phi(\mathbf{x}) = \theta_0 + \theta_1 x + \theta_2 x^2 + \dots + \theta^m x^m.$$

- (g) **(1 pts)** Recall that polynomial regression can be considered as an extension of linear regression in which we replace our input matrix  $\mathbf{X}$  with

$$\Phi = \begin{pmatrix} \phi(x_1)^T \\ \phi(x_2)^T \\ \vdots \\ \phi(x_N)^T \end{pmatrix},$$

where  $\phi(x)$  is a function such that  $\phi_j(x) = x^j$  for  $j = 0, \dots, m$ .

Update `PolynomialRegression.generate_polynomial_features(...)` to create an  $m + 1$  dimensional feature vector for each instance. **Solution:**

- (h) **(2 pts)** Given  $N$  training instances, it is always possible to obtain a “perfect fit” (a fit in which all the data points are exactly predicted) by setting the degree of the regression to  $N - 1$ . Of course, we would expect such a fit to generalize poorly. In the remainder of this problem, you will investigate the problem of overfitting as a function of the degree of the polynomial,  $m$ . To measure overfitting, we will use the Root-Mean-Square (RMS) error, defined as

$$E_{RMS} = \sqrt{J(\theta)/N},$$

where  $N$  is the number of instances.<sup>4</sup>

Why do you think we might prefer RMSE as a metric over  $J(\theta)$ ? **Solution:** RMSE has multiple advantages: (1) it is more comparable across different data set sizes since it normalizes by the number of instances; (2) it represents the sample standard deviation of the residuals and is therefore on the same scale as the predictions (compared to the the MSE, which represents the sample variance).

Implement `PolynomialRegression.rms_error(...)`. **Solution:**

- (i) **(4 pts)** For  $m = 0, \dots, 10$ , use the closed-form solver to determine the best-fit polynomial regression model on the training data, and with this model, calculate the RMSE on both the

---

<sup>4</sup>Note that the RMSE as defined is a biased estimator. To obtain an unbiased estimator, we would have to divide by  $n - k$ , where  $k$  is the number of parameters fitted (including the constant), so here,  $k = m + 1$ .

training data and the test data. Generate a plot depicting how RMSE varies with model complexity (polynomial degree) – you should generate a single plot with both training and test error, and include this plot in your writeup. Which degree polynomial would you say best fits the data? Was there evidence of under/overfitting the data? Use your plot to justify your answer. **Solution:** See Figure 1. 4<sup>th</sup>, 5<sup>th</sup>, and 6<sup>th</sup> degree polynomials fit the data best. For  $m > 8$ , there is clear evidence of overfitting – the test error increases greatly but training error decreases. For  $m < 3$ , there is evidence of underfitting – the training error is very high.

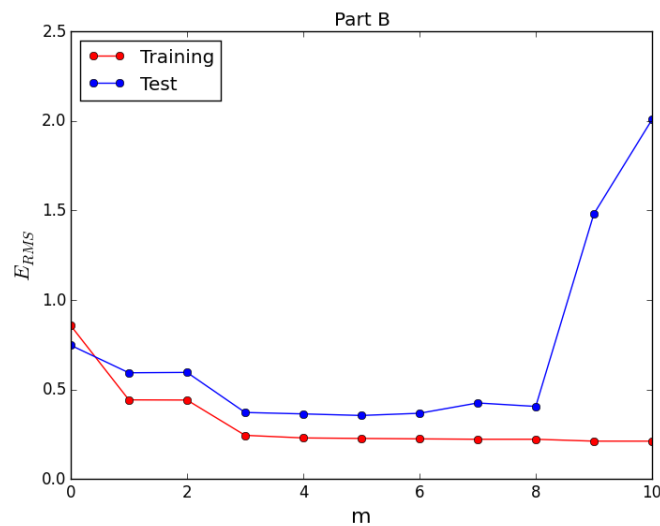


Figure 1: Training and testing error as a function of model complexity.