Homework #2

18-461/661: Intro to ML for Engineers
Prof. Gauri Joshi and Prof. Carlee Joe-Wong
Due: Monday, Feb. 10, 2020 at 8:59pm PT/11:59pm ET

Please remember to show your work for all problems and to write down the names of any students that you collaborate with. The full collaboration and grading policies are available on the course website: https://www.andrew.cmu.edu/course/18-661/.

Your solutions should be uploaded to Gradescope (https://www.gradescope.com/) in PDF format by the deadline. We will not accept hardcopies. If you choose to hand-write your solutions, please make sure the uploaded copies are legible. Gradescope will ask you to identify which page(s) contain your solutions to which problems, so make sure you leave enough time to finish this before the deadline. We will give you a 30-minute grace period to upload your solutions in case of technical problems.

1 How do hormones impact your growth? *[20 points]*

Two groups of scientists perform a series of experiments to measure the impact of different doses of a set of p hormones (represented by $\mathbf{x} \in \mathbb{R}^p$) on animals' rate of growth, which is represented as $y \in \mathbb{R}$. Scientists believe that the relation between the growth and the hormones can be assumed to be linear, i.e., $y = \mathbf{x}^\top \mathbf{w}$ where $\mathbf{w}$ is a vector of parameters that may be different for different species. The groups performed experiments to estimate $\mathbf{w}$ for two different species:

(i) The first group experimented with *rabbits* to find $\mathbf{w}_R$ and obtained n pairs of measurements $(\mathbf{x}_{R,i}, y_{R,i})$ for $i = 1, \ldots, n$.

(ii) The second group experimented with *mice* to find $\mathbf{w}_M$ and obtained m pairs of measurements $(\mathbf{x}_{M,i}, y_{M,i})$ for $i = 1, \ldots, m$.

Questions:

(a) *[3 points]* Write down the *residual sum of squares* functions $\text{RSS}^{(R)}(\mathbf{w}_R)$ and $\text{RSS}^{(M)}(\mathbf{w}_M)$ for the rabbits and mice experiment groups respectively.

(b) *[5 points]* The scientists initially wanted to find one set of parameter values $\mathbf{w}$ that fits both species' datasets, instead of finding separate $\mathbf{w}_R$ and $\mathbf{w}_M$. Find an expression for $\widehat{\mathbf{w}}$ that minimizes the total residual sum: $\text{RSS}^{(R)}(\mathbf{w}) + \text{RSS}^{(M)}(\mathbf{w})$.

(c) *[6 points]* Scientists found that using one $\mathbf{w}$ is a little inaccurate. They want to find $\mathbf{w}_R$ and $\mathbf{w}_M$ separately on each data set, but they still believe that $\mathbf{w}_R$ and $\mathbf{w}_M$ should be similar but not necessarily the same. They want to enforce this prior knowledge by adding a regularizer to the loss function as follows:(Note: We are using L2 regularization below)

$$RSS^{(R)}(\mathbf{w}_R) + RSS^{(M)}(\mathbf{w}_M) + \frac{\lambda}{2}||\mathbf{w}_R - \mathbf{w}_M||_2^2. \tag{1}$$

Find $\widehat{\mathbf{w}}_R$ and $\widehat{\mathbf{w}}_M$, the parameter vectors that minimize this new loss function with the regularizer. Note that full credit will be awarded for writing down the normal equations in terms of $\mathbf{w}_R$ and $\mathbf{w}_M$ and explaining in words what subsequent steps are required.

(d) *[3 points]* How will $||\widehat{\mathbf{w}}_R - \widehat{\mathbf{w}}_M||$ change as we increase λ?

(e) *[3 points]* How will $\frac{||\widehat{\mathbf{w}}_R||}{||\widehat{\mathbf{w}}||}$ and $\frac{||\widehat{\mathbf{w}}_M||}{||\widehat{\mathbf{w}}||}$ change as λ goes to infinity?(Remember that $\widehat{\mathbf{w}}$ is the weight vector that minimizes the combined RSS in part (b).)

2 Hoeffding's inequality and confidence of MLE estimate *[20 points]*

Let us go back to the Dogecoin example from Lecture 2's slides. Peter found a Dogecoin in the corridor in front of HH 1107. As it is worthless, Peter decided to use it only for choosing the first team to play in the soccer game he is refereeing next weekend. As a diligent referee, Peter wants to make sure that the coin is fair. In other words, the probabilities of heads and tails are both 0.5.

We will assume that each coin flip is independent and denote the probability of heads as p. Peter wants to estimate p from his experiments of flipping the coin n times.

(a) **MLE estimate:** *[4 points]* The diligent Peter flips the coin $n = 1000$ times for his experiment, and gets 490 heads and 510 tails. What is his MLE estimate of p?

(b) **Hoeffding's inequality:** *[4 points]* Peter is not sure if his experiment of 1000 coin flips is enough to claim whether the coin is fair. To see whether 1000 flips are enough, he decides to use Hoeffding's inequality, which provides a bound on how fast the sample mean approaches the true mean.

Let $X_1, \cdots, X_n$ be i.i.d. random variables whose possible values lie in the bounded interval $[a, b]$. In our Dogecoin example, each X_i represents the outcome of the ith coin flip, which we represent as "1" if the outcome is heads and "0" if it is tails. Thus, each $X_i \sim$ Bernoulli(p), a random variable that is 1 (heads) with probability p and 0 (tails) otherwise.

We denote the sample mean of X as the average of the X_i values: $\overline{X} = \frac{1}{n}(X_1 + \cdots + X_n)$. Then, Hoeffding's inequality states the following:

$$\Pr(|\overline{X} - \mathbb{E}[X]| \geq \epsilon) \leq 2\exp\left(-\frac{2n\epsilon^2}{(b-a)^2}\right). \tag{2}$$

Rewrite Hoeffding's inequality (2) for the case $X \sim$ Bernoulli(p).

(c) **Confidence Interval:** *[4 points]* Using the inequality you obtained in part (b), find the probability that $\widehat{p}_{\text{MLE}}$ is between $p - 0.05$ and $p + 0.05$ after 1000 coin flips. Then, with how much percentage confidence can Peter claim that p lies in $(0.44, 0.54)$?

(d) **Effect of n on Confidence:** *[4 points]* Find the probability of p lying in $(0.44, 0.54)$ for $n = 500$ and $n = 10000$.

(e) **Tighter Confidence Interval:** *[4 points]* What is the minimum number of coin flips needed for Peter to claim that he estimated the true value of p within ± 0.01 error (that is, with 99% confidence)?

3 Ridge Regression *[20 points]*

3.1 Least Squares estimator *[5 points]*

We are given three training data points as follows:

$$(\boldsymbol{x}_1 = (2, -1), y_1 = 3)$$

$$(\boldsymbol{x}_2 = (4, -2), y_2 = 1)$$

$$(\boldsymbol{x}_3 = (6, -3), y_3 = 7)$$

We want to fit a simple linear model:

$$y = w_1 x[1] + w_2 x[2] = \boldsymbol{x}^\top \boldsymbol{w},$$

where $x[1]$ and $x[2]$ denote the first and the second dimensions of the feature vector $\boldsymbol{x}$. Note that there is no bias term. Compute the weight vector $\hat{\boldsymbol{w}}$ that minimizes the residual sum of squares error, $\sum_{i=1}^{3}(y_i - \boldsymbol{x}_i^\top \boldsymbol{w})^2$, for the given data points. If there is more than one solution, please find all such vectors.

3.2 Implementation of the Ridge Regression Estimator *[5 points]*

In order to demonstrate the ridge regression estimation, we use a data example prepared by Liebmann et al. (2009) (https://www.ncbi.nlm.nih.gov/pubmed/19427473). The matrix Y contains the concentration of glucose and ethanol (in g/L) for $n = 166$ alcoholic fermentation mashes of different feedstock (rye, wheat and corn). These are the two dependent variables. There are 235 covariates in X, which contain the first derivatives of near infrared spectroscopy (NIR) absorbance values at $1115 - 2285$ nm. In this problem, we will predict the glucose concentration for the given covariates. The training dataset (files: Ytraining.csv, Xtraining.csv) consists of 126 observations. The dataset is further divided into a validation set (files: Yvalidation.csv, Xvalidation.csv) and a testing set (files: Ytesting.csv, Xtesting.csv); each contains 20 observations. (Note: There is no need to include bias term)

In this sub-question, you will implement the Ridge Regression estimator. Your code must have the following methods:

- **fit($\mathbf{X}_n, \mathbf{y}_n, \lambda$)**: which fits the model given a value for the regularization parameter.
- **predict(X, w)**: which predicts the values of the dependent variable for a new set of covariates using the learned model.

Please use the baseline python file provided to write your code, and attach it to your solution pdf.

3.3 Evaluation *[10 points]*

For the given dataset:

- Plot in the same figure, the learned coefficients $\mathbf{w}$ with respect to the regularization parameter λ, when it ranges from 0.001 to 2 with a step of 0.002.What do you observe?
- Plot on the y-axis the RMSE (Root Mean Squared Error) of the learned model on the validation set with respect to the regularization parameter. Find the regularization parameter λ^* which achieves the minimum RMSE.
- For the λ^* found above, plot the predicted versus the real value of the glucose concentration, when the model is evaluated on the testing dataset. That is, for the 20 testing points plot the true values of glucose concentration (on x-axis) vs. predicted values of glucose concentration (on y-axis).

4 Regression with Regularization *[10 points]*

You are asked to use regularized linear regression to predict the target $y \in \mathbb{R}$ from the eight-dimensional feature vector $\mathbf{x} \in \mathbb{R}^8$. You define the model $y = \mathbf{w}^T\mathbf{x}$ and then you recall from class the following three objective functions(Note that the bias term is included within each expression as a part of x_i):

$$\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 \tag{3}$$

$$\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda \sum_{j=1}^{8} (\mathbf{w}_j)^2 \tag{4}$$

$$\min_{\mathbf{w}} \sum_{i=1}^{n} (y_i - \mathbf{w}^T \mathbf{x}_i)^2 + \lambda \sum_{j=1}^{8} |\mathbf{w}_j| \tag{5}$$

(a) *[3 points]* Identify the regularization terms (if any) in each objective function above and state the type of regularization used for each term.

(b) *[2 points]* As we increase the value of λ in objective (4), the bias would:

(i) increase
(ii) decrease
(iii) remain unaffected

(c) *[2 points]* As we increase the value of λ in objective (5), the variance would:

(i) increase
(ii) decrease
(iii) remain unaffected

(d) *[3 points]* The following table contains the weights learned for all three objective functions (not in any particular order):

–	ColA	ColB	ColC
$\mathbf{w}_1$	0.60	0.38	0.50
$\mathbf{w}_2$	0.30	0.23	0.20
$\mathbf{w}_3$	-0.10	-0.02	0.00
$\mathbf{w}_4$	0.20	0.15	0.09
$\mathbf{w}_5$	0.30	0.21	0.00
$\mathbf{w}_6$	0.20	0.03	0.00
$\mathbf{w}_7$	0.02	0.04	0.00
$\mathbf{w}_8$	0.26	0.12	0.05

Determine which column of weight values (A,B or C) corresponds to each objective. Briefly explain your reasoning.

- Objective 3
- Objective 4
- Objective 5

5 Naive Bayes Classifier *[10 points]*

Consider the following data set with three features X_1,X_2,X_3, and label Y.

X_1	X_2	X_3	Y
1	0	1	0
0	0	1	0
1	1	0	0
0	0	0	1
1	1	1	1
1	0	0	1

(a) *[3 points]* Suppose that we wish to fit a Naive Bayes model to the data, i.e $P(Y|X_1, ...X_n) \propto P(Y)\prod_{i=1}^{n} P(X_i|Y)$. Calculate the probability that $Y = 0$ given $X = (0, 1, 1)$. Show your work.

(b) *[3 points]* Given the same data set, apply Laplace Smoothing with $k = 1$. Calculate the probability that $Y = 1$ given $X = (0, 1, 0)$.

(c) *[4 points]* Now apply Laplace Smoothing with $k = 5$. What is the probability that $Y = 1$ given $X = (0, 1, 0)$? How does this probability compare with your answer to part (b)? Briefly explain your answer.

6 Polynomial Regression *[20 points]*

6.1 Derivation *[5 points]*

In the class, we have learned that by using non-linear basis functions, we can fit a non-linear model as if it is a linear model. In this problem, we are given a training data set $\mathcal{D}$ that consists of n points, (x_i, y_i) for $i = 1, \cdots, n$, and we consider a polynomial regression problem as follows:

$$y_i = w_0 + w_1 x_i + \cdots w_k x_i^k. \tag{6}$$

x_i and y_i are both real numbers.

(a) What is the basis function $\boldsymbol{\phi}(x)$ for this problem?

(b) Can you express residual sum of squares error in terms of $\boldsymbol{\phi}(x)$?

(c) What is $\mathbf{w} = \begin{bmatrix} w_0, & w_1, & \cdots, & w_k \end{bmatrix}^\top$ that minimizes the residual sum of squares error?

6.2 Implementation *[5 points]*

Now we want to implement polynomial regression on 40 data points in `poly_reg_data.csv`. Implement the following functions:

- `fit(X_train, Y_train, k)`: This fits a polynomial model on the training dataset given the polynomial degree `k`, and outputs the weight terms `w`.
- `validate(X_val, Y_val, w)`: This computes the RSS error on the validation set given the fitted weight vector `w`.

As with Section 3.3, please remember to attach your code to the solution pdf.

6.3 Evaluation *[10 points]*

Now, divide randomly the dataset into two parts: training set (25 points) and validation set (15 points). By varying k from 1 to 10, obtain the RSS error on the validation set.

- Plot training error (RSS error on the training set) versus k. From the plot, which k gives you the minimum training error?
- Plot validation error versus k. Plot which k gives you the minimum validation error?
- If the optimum k values for training error and validation error are different, can you explain why?
- For each $k = 1, 3, 5, 10$, create a scatter plot of the training data points. On the same plot for each k, draw a line for the fitted polynomial.