EE 660 **Homework 1 (Week 3)** Posted: Tues., 9/6/2022
Jenkins Due: Tues., 9/13/2022, 11:59 PM PDT

Reminder: Policy on collaboration on all homework assignments [from syllabus, p.8]:

> Collaboration on techniques for solving homework assignments and computer problems is allowed, and can be helpful; however, each student is expected to work out, code, and write up his or her own solution. Use of other solutions to homework assignments or computer problems, from any source including other students, before the assignment is turned in, is not permitted.

Please turn in your Homework 1 solution by uploading 2 files to the Homework 1 (Week 3) assignment dropbox in D2L, as follows. This is required.

(1) **a single pdf file of your solutions / answers to all the homework problems.** Please note:
   (a) Your work can be handwritten or typeset. If handwritten, just scan it in, or take a picture with your smartphone and use a scan app to convert it to pdf; the result should look like a document-scanner result, not a photograph. Please do not upload pictures in native (picture) format; the quality will be very sensitive to lighting, and might not entirely readable.
   (b) Please check the pdf for readability before uploading, and keep the file size reasonable (less than 5 MB).

(2) **a second pdf file that contains all your computer code** (for Problem 1, and for Problem 4 if you used a computer). This must be machine readable (not a scan, not a screenshot), and in a single file.

Thank you for cooperating; our grading methods depend on submissions as described above.

1. *Comparison of linear regression using least squares, ridge, and lasso.*

Comment: for this homework problem, you may use Python libraries NumPy, sklearn, pandas, matplotlib, etc.

After learning the regression part and different regularizations, Bob is interested in trying them out right away! He starts with the linear regression problem: given that the feature vector $\boldsymbol{x}$ and the observation y have a linear relationship $y = \boldsymbol{w}^{\boldsymbol{T}}\boldsymbol{x} + w_0 + n$, estimate the weight vector $\boldsymbol{w} = [w_1, w_2, \dots]$ and the bias w_0 from multiple data points. For simplicity in writing, we can augment the feature space and now parameters to be estimated can be written as $\boldsymbol{w} = [w_0, w_1, w_2, \dots]$. Here n is the observation noise on the output labels y. Bob starts to collect some samples to generate his dataset. He does the collection for several times and gets several datasets with different numbers of samples:

	number of training samples (N_{tr})	number of testing samples
Dataset1	5	1000

Dataset2	50	1000
Dataset3	500	1000

Could you help him out on analysis on Datasets 1, 2, 3 above?

(a) Given that the dimension of features is 9 (before augmentation), estimate the $\boldsymbol{w}$ and try three regularization settings: [no regularization, l_1 regularization, l_2 regularization] and report the corresponding statistics. For each regularization setting to try, you need to search for a good regularization coefficient λ over the range $-10 \leq \log_2 \lambda \leq 10$ with step size of 0.5 for $\log_2 \lambda$, and use MSE (mean squared error) on the validation set to choose the best one. During the parameter search, you need to do 5-fold cross validation on each parameter value you try.
Tip: after finding the best value of λ, use that value for one final training run using all N_{tr} training data points (nothing held out as a validation set), to get the weight vector and training MSE.

i. Fill all your numerical results into the following table. (Each dataset should have a different table. So for this question you'll have 3 tables.)

ii. Based on statistics on all datasets, answer the following questions:

1. Comparison of test MSE with no regularizer, l_1 regularizer, and l_2 regularizer for a given N_{tr} (your answer might also depend on N_{tr})
2. Does each regularizer lower the corresponding norm of $\boldsymbol{w}$? by very much? Please explain. Why are these answers different depending on N_{tr}?
3. Observe and explain the dependence of sparsity on regression method, and on different values of N_{tr} and λ.

	Model selection			Performance	
	Best param $\log_2 \lambda$	Mean of MSE	Std of MSE	MSE on train	MSE on test
Least square	-	-	-		
	$\boldsymbol{w}$	(show your estimated w)			
		$l_1(w) =$	$l_2(w) =$	Spars=	
LASSO					
	$\boldsymbol{w}$	(show your estimated w)			
		$l_1(w) =$	$l_2(w) =$	Spars=	
Ridge					
	$\boldsymbol{w}$	(show your estimated w)			
		$l_1(w) =$	$l_2(w) =$	Spars=	

Caption for statistics in the table:

- Best param λ: the regularization coefficient you choose using cross validation.
- Mean of MSE: the averaged MSE of the 5-fold cross validation process for your chosen λ.

- Std of MSE: the standard deviation of MSE of the 5-fold cross validation process for your chosen λ.
- $l_1(w)$: l_1 norm of $\boldsymbol{w}$
- $l_2(w)$: l_2 norm of $\boldsymbol{w}$
- Spars: Sparsity, i.e., the number of zeros in the augmented weight vector

(b) Bob learned that l_1 regularization could lead to more sparsity, and he really wants to visualize this. So he collects Datasets 4-9, all for 2-dimensional (before augmentation) features:

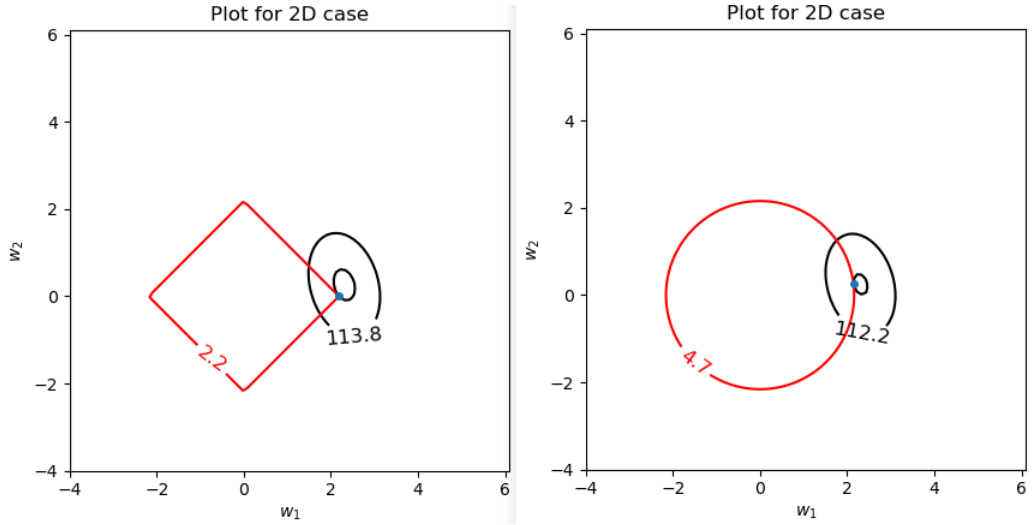
	number of training samples (N_{tr})	number of testing samples
Dataset4	5	1000
Dataset5	15	1000
Dataset6	50	1000
Dataset7	5	1000
Dataset8	15	1000
Dataset9	50	1000

He tries them out and finds some expected and some unexpected results.

i. Repeat (a)(i) for all new datasets. (You'll have 6 tables.)

ii. For each dataset, draw the following plot in the 2D space w_2 vs. w_1 with $w_0 =$ your estimated w_0: (1) draw the curve of 'MSE = training MSE of your estimated $\boldsymbol{w}$ and 'MSE=10+training_MSE of your estimated $\boldsymbol{w}$; (2) draw the curve for $\|w\|_{l1} =$ the l_1 norm of your estimated $\boldsymbol{w}$. Repeat this plot drawing for ridge regression results, except for (2) draw the curve for $\|w\|_{l2} =$ the l_2 norm of your estimated $\boldsymbol{w}$. (therefore you have 2 plots for each dataset. An example is shown below.)

iii. Based on the statistics and plots, answer the following questions:
   1. Observe and explain how the plots relate to sparsity.
   2. Can you explain how much effect the regularizer has, from looking at the plots (i.e., how different the regularized performance (MSE) is from the unregularized performance)
   3. Observe and explain how Lasso has a different effect with the "special case" datasets than the other datasets



Hint: please refer to the example.py code file in the homework folder on how to generate such plots.

2. *Estimating σ^2 in linear regression.* You are given a Gaussian model:

$$p(y|\underline{x},\underline{\theta}) = N(y|\underline{w}^T\underline{x},\sigma^2)$$

and a dataset with *N* data points.

(a) Find the MLE of the variance σ^2, for a given constant $\underline{w} = \underline{\hat{w}}$.

(b) Is the assumption for part (a) that $\underline{w} = \underline{\hat{w}}$ is a constant of σ^2, reasonable? Justify your answer. (**Hint:** consider the MLE solution for $\underline{\hat{w}}$.)

3. *Nonlinear ridge regression*. Suppose we use a basis function expansion $\underline{\phi}(\underline{x})$ to make ridge regression nonlinear in $\underline{x}$. Thus the model is:

$$p(y|\underline{x},\underline{\theta}) = N\left(y|\underline{w}^T\underline{\phi}(\underline{x}),\sigma^2\right)$$

and the prior on $\underline{w}$ is a Gaussian as stated in lecture.

(a) Let $\underline{\underline{\Phi}}$ be the data matrix. Give the objective function $J(\underline{w},\mathcal{D})$ in terms of $\underline{\underline{\Phi}}$, $\underline{w}$, $\underline{y}$, and λ .

(b) Give the solution for $\underline{\hat{w}}$ in terms of the same quantities. Briefly justify, or derive, your result.

4. *Comparison of loss functions* in logistic regression (log exponential loss), perceptron, and mean-squared error criterion functions, for classification.

For this problem you may use python built-in functions, NumPy, and matplotlib. Note that $\tilde{y}_i$ is binary with $\tilde{y}_i \in \{-1,+1\}$.

Throughout this problem, let $s_i = \tilde{y}_i \underline{w}^T \underline{x}_i$.

(a) For logistic regression based on MLE, the loss function is (Lecture 5, Eq. 22):

$$E_i^{(lr)} = \ln\left[1 + \exp\left\{-\tilde{y}_i \underline{w}^T \underline{x}_i\right\}\right] .$$

Plot $E_i^{(lr)}$ *vs.* s_i, twice: once for $-10 \le s_i \le 10$, and again for only $-2 \le s_i \le 2$ so that more detail can be viewed near $s_i = 0$. (2 plots total.)

(b) For 2-class linear perceptron learning (from EE 559), the objective function is:

$$J(\underline{w}) = -\sum_{i=1}^{N} \left[\!\left[\tilde{y}_i \underline{w}^T \underline{x}_i \le 0 \right]\!\right] \tilde{y}_i \underline{w}^T \underline{x}_i = \sum_{i=1}^{N} E_i^{(p)}$$

Give an expression for the loss function $E_i^{(p)}$ in terms of s_i. Plot $E_i^{(p)}$ *vs.* s_i, twice: once for $-10 \le s_i \le 10$, and again for only $-2 \le s_i \le 2$ so that more detail can be viewed near $s_i = 0$. (2 plots total.)

(c) For the MSE objective function in a 2-class linear classification problem, the MSE can be written:

$$MSE = \frac{1}{N} \sum_{i=1}^{N} \left[\underline{w}^T \underline{x}_i - b_i \right]^2 = \sum_{i=1}^{N} E_i^{(mse)} \ .$$

in which b_i is the target value for data point *i*. Let the target value be $b_i = \tilde{y}_i \quad \forall i.$

Write the loss function $E_i^{(mse)}$ in terms of s_i. (**Hint:** first insert $\tilde{y}_i$ into the above expression for MSE, in an appropriate place, where it has no effect on the MSE result.) Plot $E_i^{(mse)}$ *vs.* s_i, twice: once for $-10 \le s_i \le 10$, and again for only $-2 \le s_i \le 2$ so that more detail can be viewed near $s_i = 0$. (2 plots total.)

(d) Compare the plots of (a), (b), and (c) above. Describe how these 3 loss functions contribute differently to the objective function. (For example, compare the loss functions for correctly classified data points that are near the decision boundary, and that are far from the decision boundary; likewise, compare the loss functions for incorrectly classified data points that are near the decision boundary, and that are far from the decision boundary.)