ELEN 4720: Machine Learning for Signals, Information and Data

Columbia University, Spring 2020

Homework 1: Due Monday, October 3, 2022 by 4:59PM

Please read these instructions to ensure you receive full credit on your homework. Submit the written portion of your homework as a *single* PDF file through Courseworks (less than 5MB). In addition to your PDF write-up, submit all code written by you in their original extensions through Courseworks (e.g., .m, .r, .py, .c). Any coding language is acceptable, but do not submit notebooks. Also, do not wrap your files in .rar, .zip, .tar and do not submit your write-up in .doc or other file type. When resubmitting homeworks, please be sure to resubmit *all files*. Your grade will be based on the contents of one PDF file and the original source code. Additional files will be ignored. We will not run your code, so everything you are asked to show should be put in the PDF file. Show all work for full credit.

Problem 1 (written)

Consider a regression problem, in which we observe pairs $(x_1, y_1), \ldots, (x_n, y_n)$, with $x_i \in \mathbb{R}^d$ and $y_i \in \mathbb{R}$. We seek w such that

$$x_i w \approx y_i,$$

by *least squares*:

$$w_{\text{LS}} = \arg\min_w \mathcal{L}_{\text{LS}}(w) \equiv \sum_{i=1}^{n} \left(x_i^T w - y_i\right)^2. \tag{1}$$

In lecture, we argued that

$$\nabla \mathcal{L}_{LS}(w) = 2X^T(Xw - y), \tag{2}$$

where

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} \in \mathbb{R}^{n \times d}, \qquad \text{and} \qquad y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \in \mathbb{R}^n. \tag{3}$$

We used this to show that whenever $X^T X$ has full rank d

$$w_{\text{LS}} = (X^T X)^{-1} X^T y. \tag{4}$$

Please answer the following questions:

1.1. Number of samples for uniqueness: How large does n need to be compared to d for it to be *possible* for $X^T X$ to have full rank d? When this condition is not satisfied, how should we modify the formulation (1)?

$$X^T X = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \end{bmatrix} \begin{bmatrix} 1 & x_1 \\ 1 & x_2 \\ \vdots & \vdots \\ 1 & x_n \end{bmatrix} = \begin{bmatrix} n & \sum x_i \\ \sum x_i & \sum x_i^2 \end{bmatrix}$$

1-D $[1 \times n][n \times 1]$

D-dim. $[d \times n][n \times d]$

$\downarrow$

$[d \times d]$

$X^T X$ is full rank if rank $= d$, is invertible, and $\neq 0$

if $X \in \mathbb{R}^{n \times d}$ then $X^T \in \mathbb{R}^{d \times n}$

$$X^T X w_{LS} = X^T y$$

$X^T X = X_1 X_1^T$ rank of 1 if 1 data point.

Have n data points

$X^T X$ bounded by number of data points

$$X^T X = \sum_{i=1}^{n} X_i X_i^T$$

$\text{rank}(X^T X) \leq n$

Need $n \geq d$ for $X^T X$ to be full rank

Condition not satisfied: if two inputs are perfectly correlated (eg. $x_2 = 3x_1$).

Then $X^T X$ is singular and the least squares coefficients $\vec{w}$ are not uniquely defined.

If n is not $\gg d$ (say, $n > 10d$) then solution is not unique and need to add regularization to constrain the model parameters.

$$w = \text{argmin} \|y - Xw\|^2 + \lambda g(w)$$

One regularizer is $g(w) = \|w\|^2$, called ridge regression, which penalizes large w.

The regularization parameter λ generates a tradeoff between the first and second terms (loss + regularizer).

Section 5.2.3 Hasties textbook.

1.2. Unbiased estimates: Suppose that $y = Xw_{\text{true}} + z$, with $z \sim \mu$ a random (noise) vector with probability distribution μ. Under what conditions on z, is the least squares solution *unbiased*, i.e.,

$$\mathbb{E}\Big[w_{\text{LS}}\Big] = w_{\text{true}}? \tag{5}$$

Note: in lecture we showed that if $z \sim_{\text{iid}} \mathcal{N}(0, \sigma^2)$, w_{LS} is unbiased. However, this holds more broadly – please make your conditions as broad as possible for full credit!

Lemma: If $y = Xw + z$ where X is nonrandom and X^TX is invertible and $E(z) = 0$, then $E(w_{LS}) = w_{true}$

Proof: $w_{LS} = (X^TX)^{-1}X^Ty = (X^TX)^{-1}(X^TXw + z) = w + (X^TX)^{-1}X^Tz$

$$E(w_{LS}) = w + (X^TX)^{-1}X^TE(z)$$
$$= w$$

The least squares estimates w_{LS} are unbiased for w as long as z has mean zero. Does not require normally distributed errors. Does not even make assumptions about $var(z)$. To study $var(w_{LS})$ we will need assumptions on $var(w)$ but not its mean

1.3. Least squares by gradient descent: We can also compute w_{LS} iteratively, using gradient descent – namely, by choosing some initial guess w^0 and updating w as

$$w^{k+1} = w^k - t\nabla\mathcal{L}_{\text{LS}}(w^k). \tag{6}$$

Part A. Show that

$$w^{k+1} - w_{\text{LS}} = \left(I - 2tX^TX\right)\left(w^k - w_{\text{LS}}\right) \tag{7}$$

Hint: use our expressions for ∇L_{LS} and w_{LS}.

$$w^{K+1} = w^K - t\nabla\mathcal{L}_{LS}(w^K)$$

substitute $\nabla\mathcal{L}_{LS}(w^K) = 2X^T(Xw^K - y)$

$$w^{K+1} = w^K - 2tX^T(Xw^K - y)$$
$$w^{K+1} = w^K + 2tX^T(y - Xw^K)$$
$$w^{K+1} = w^K + 2t(X^Ty - X^TXw^k)$$

Substitute $X^T X w_{LS} = X^T y$

$$w^{K+1} = w^K + 2t\left(X^T X w_{LS} - X^T X w^K\right)$$

$$= w^K + 2t\left(X^T X w_{LS}\right) - 2t\left(X^T X w^K\right)$$

$$= w^K - 2t\left(X^T X w^K\right) + 2t\left(X^T X w_{LS}\right)$$

$$= w^K \underbrace{\left(I - 2t X^T X\right)}_{I} + 2t X^T X w_{LS}$$

then subtract w_{LS} from both sides

$$w^{K+1} - w_{LS} = w^K\left(I - 2t X^T X\right) + 2t X^T X w_{LS} - w_{LS}$$

$$w^{K+1} - w_{LS} = w^K\left(I - 2t X^T X\right) - w_{LS}\left(I - 2t X^T X\right)$$

$$\boxed{w^{K+1} - w_{LS} = \left(w^K - w_{LS}\right)\left(I - 2t X^T X\right)}$$

Part B. Let $\lambda_1 \geq \cdots \geq \lambda_n$ denote the eigenvalues of the symmetric matrix $I - 2tX^T X$. Argue that if t is chosen such that $|\lambda_i| < 1$ for all i, then as $k \to \infty$, $w^k \to w_{\text{LS}}$.

Hint: You can use the following inequality. For a symmetric matrix M with eigenvalues $\lambda_1, \ldots \lambda_n$, and a vector x,

$$\|Mx\|_2 \leq \left(\max_i |\lambda_i|\right)\|x\|_2.$$

t chosen such that $|\lambda_i| < 1$ for all i $\rightarrow$ full column rank assumption

to ensure that it is strictly better than -1 we need $t < \frac{1}{\lambda_{max}(X^T X)}$

$$\|Mx\|_2 \leq \lambda_{max} X^T X$$

$$\|(I - 2tX^T X)x\|_2 \leq \lambda_{max} X^T X$$

$$w_{LS} = (X^T X)^{-1} X^T y$$

$$w_{LS} = w_{RR} = (X^T X + \lambda I)^{-1} X^T y$$

Problem 2 (written)

As in Problem 1, you have data (x_i, y_i) for $i = 1, \ldots, n$, where $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$. Here, we apply ridge regression, setting

$$w_{\mathrm{RR}} = \arg\min_w \sum_{i=1}^{n} (x_i^T w - y_i)^2 + \lambda w^T w. \tag{8}$$

Please answer the following questions:

2.1. Demonstrate that the ridge regression solution w_{RR} is given by

$$w_{\mathrm{RR}} = \left(\lambda I + X^T X\right)^{-1} X^T y. \tag{9}$$

Hint: differentiate the objective function in (8) and set the derivative equal to zero.

$\underset{w}{\operatorname{argmin}} \sum_{i=1}^{n} (x_i^T w - y_i)^2 + \lambda w^T w$

least squares: $\|y - Xw\|^2$; $\lambda \|w\|^2$

$\mathcal{L} = (y - Xw)^T (y - Xw) + \lambda w^T w$

take gradient of $\mathcal{L}$ w.r.t w and set to zero

$\nabla \mathcal{L} = 2X^T X w - 2X^T y + 2\lambda w = 0$

solve for w

$w(2\lambda + 2X^T X) - 2X^T y = 0$

$w(2\lambda + 2X^T X) = 2X^T y$

$w_{RR} = 2X^T y (2\lambda + 2X^T X)^{-1}$

2.2. Consider two potential values of λ: $\lambda = 0.1$ and $\lambda = 1,000$. For which value of λ do you expect the ridge regression solution w_{RR} to exhibit greater *bias*? For which value of λ do you expect it to exhibit greater *variance*?

The regularization term $\lambda w^T w$ is imposed to penalize values in w that are large. This reduces potential high-variance predictions from least squares. As λ increases, the flexibility of the ridge regression fit decreases, leading to decreased variance but increased bias.

So $\lambda = 1,000$ leads to more bias since the regularizing term is large.

$\lambda = 0.01$ exhibits more variance, since the regularizing term is made very small and basically follows the least squares solution.

The least squares solution has increased variance but decreased bias.

2.3. Describe a procedure for choosing λ in practice.

In practice, bias is related to a model failing to fit the training dataset and variance is related to a model failing to fit the testing data set.

An optimal value of λ finds the minimum bias and variance despite the tradeoff. One way to do this is through cross validation.

1. Split data into K roughly equal groups
2. Train the model on K-1 groups and predict the held-out K^{th} group.
3. Repeat K times, holding out each group once. Calculate prediction error of the fitted model when predicting the K^{th} part of the data. Typical choices of K are 5 or 10. The case $K = N$ is known as leave-one-out cross val. The cross validation estimate of prediction error is:

$$CV(\hat{f}) = \frac{1}{N} \sum_{i=1}^{N} L\left(y_i, \hat{f}^{-\kappa(i)}(x_i)\right)$$

where $\kappa\{1,\ldots,N\} \longmapsto \{1,\ldots,K\}$

4. After all the iterations are done, the model will be trained each time using a different fold. To get the cross-validation score, take the average of all the individual hold out scores.
5. This cross-validation score will have used a specific λ on the data. Steps 1-4 should be repeated for various λ. The optimal value of λ will be that with the highest cross-validation score.

Problem 3 (coding)

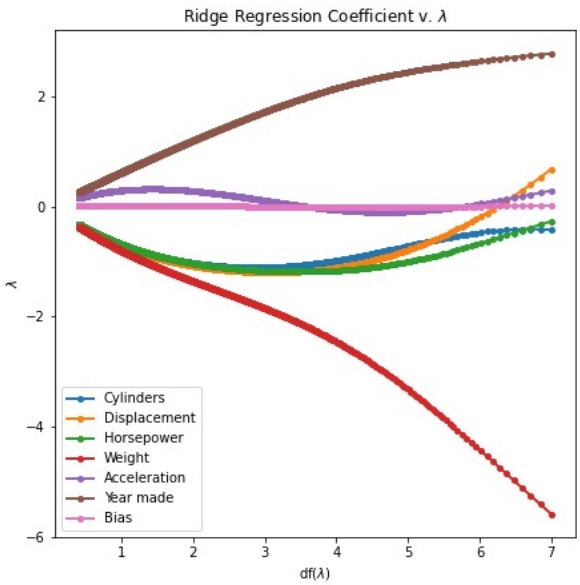
In this problem you will analyze data using the linear regression techniques we have discussed. The goal of the problem is to predict the miles per gallon a car will get using six quantities (features) about that car. The zip file containing the data can be found on Courseworks.[1] The data is broken into training and testing sets. Each row in both "X" files contain six features for a single car (plus a 1 in the 7th dimension) and the same row in the corresponding "y" file contains the miles per gallon for that car.

Remember to submit all original source code with your homework. Put everything you are asked to show below in the PDF file.

Part 1. Using the training data only, write code to solve the ridge regression problem

$$\mathcal{L} = \lambda\|w\|^2 + \sum_{i=1}^{350} \|y_i - x_i^T w\|^2.$$

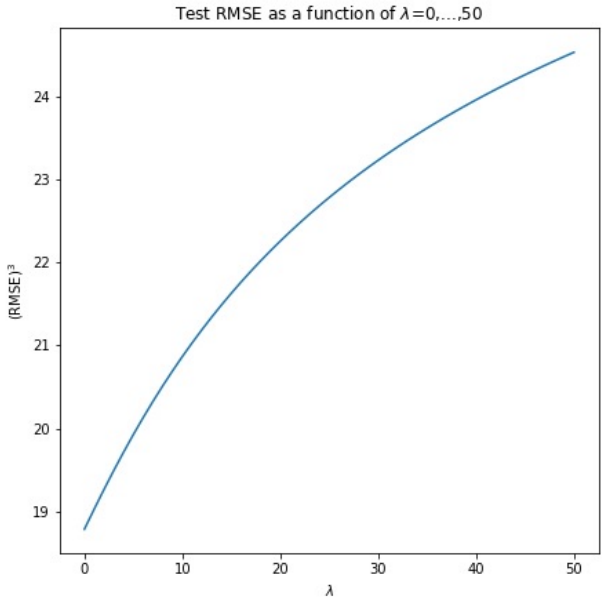
(a) For $\lambda = 0, 1, 2, 3, \ldots, 5000$, solve for w_{RR}. (Notice that when $\lambda = 0$, $w_{\text{RR}} = w_{\text{LS}}$.) In one figure, plot the 7 values in w_{RR} as a function of $df(\lambda)$. You will need to call a built in SVD function to do this as discussed in the slides. Be sure to label your 7 curves by their dimension in x.[2]



(b) Two dimensions clearly stand out over the others. Which ones are they and what information can we get from this?

"Year made" and "weight" stand out which means these two features are most predictive in determining the value of y (miles per gallon) for the car. This is also intuitive since newer cars are more fuel efficient and increasing weight will require more gas/energy to move the car.

(c) For $\lambda = 0, \ldots, 50$, predict all 42 test cases. Plot the root mean squared error (RMSE)^3 on the test set as a function of λ—*not* as a function of $df(\lambda)$. What does this figure tell you when choosing λ for this problem (and when choosing between ridge regression and least squares)?



With increasing λ, RMSE also increases. Referring to the ridge regression equation: $\mathcal{L} = \sum_{i=1}^{n} \|y_i - x_i^T w\|^2 + \lambda \|w\|^2$

λ controls the weight of the penalty term $\|w\|^2$. When λ is zero, the equation becomes a least squares estimate: $\mathcal{L} = \sum \|y_i - x_i^T w\|^2$.

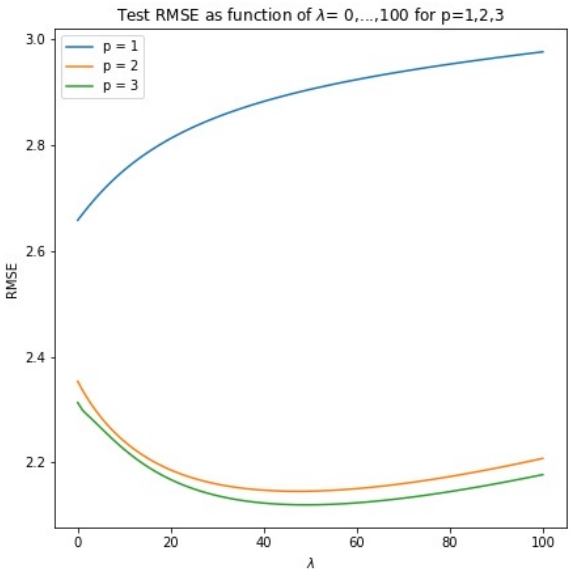
Adding a penalty term reduces the variance and increases the bias.

From the plot in part (b) the features most correlated with the predicted variable are most apparent with high bias / increasing λ. Therefore ridge regression / high λ is favored.

However, this plot (c) shows that a $\lambda = 0$ has the lowest error, and therefore a low λ or least squares should be used.

Part 2. Modify your code to learn a pth-order polynomial regression model for $p = 1, 2, 3$. (You've already done $p = 1$ above.) For this implementation use the method discussed in the slides. Also, be sure to standardize each additional dimension of your data.

(d) In one figure, plot the test RMSE as a function of $\lambda = 0, \ldots, 100$ for $p = 1, 2, 3$. Based on this plot, which value of p should you choose and why? How does your assessment of the ideal value of λ change for this problem?



As shown also in plot from (c), a λ value of 0 should be chosen for $p=1$ since this is the lowest error. However when $p=2$, the error decreases with increasing values of λ, with a valley in the RMSE graph at around $\lambda = 40$. For $p=3$, the lowest error is at around $\lambda = 50$. Therefore, the ideal value for λ depends on the p^{th}-order of the polynomial regression model. It appears from this graph (d) that increasing p will decrease the RMSE. But from the data given, $p=3$ should be used since it has the lowest error, and a λ value of ~ 50.