# Homework 4

AI 539 - Machine Learning for Non-AI Majors
Instructor: Alireza Aghasi
Due date: See Canvas

May 6, 2024

Please revise the homework guidelines reviewed in the first lecture. Specifically note that:

- Start working on the homework early
- Late homework **is not accepted** and will receive zero credit.
- Each student must write up and turn in their own solutions
- **(IMPORTANT)** If you solve a question together with another colleague, each need to write up your own solution and **need to list the name of people who you discussed the problem with on the first page of the material turned in**
- The homework should be manageable given the material lectured in the class. The long questions are to help clarifying the problem.

**Q1.** In the class we mentioned that for linear models, the LOOCV test error estimate can be obtained as a closed form expression, and there is no need to perform $n$ linear fits for a data set of size $n$. The goal of this question is for you to obtain that closed-form expression, for a class of linear models. We focus on the simple linear model.

We have a dataset as $(x_1, y_1), (x_2, y_2), \ldots, (x_n, y_n)$. Consider fitting a model of the form $y = \beta_0 + \beta_1 x$ to this set. By setting up the RSS and minimizing it, we remember that the fitted parameters are

$$\hat{\beta}_1 = \frac{S_{xy} - n\bar{x}\bar{y}}{S_{xx} - n\bar{x}^2}, \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

where:

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i, \quad \bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i, \quad S_{xy} = \sum_{i=1}^{n} x_i y_i, \quad S_{xx} = \sum_{i=1}^{n} x_i^2.$$

**Assumption 1.** To have an easier derivation, throughout this question assume that, the feature part of the data are centered, or basically $\bar{x} = 0$.

(a) Under Assumption 1, show that

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \qquad \hat{\beta}_0 = \bar{y}. \tag{1}$$

(b) Let's pick $j$ to be a fixed index between 1 and $n$. Now split our reference data into two sets:

$$\text{set } 1 : (x_1, y_1), (x_2, y_2), \ldots, (x_{j-1}, y_{j-1}), (x_{j+1}, y_{j+1}), \ldots, (x_n, y_n)$$
$$\text{set } 2 : (x_j, y_j).$$

Basically set 1 is the reference set without the $j$-th sample, and set 2 only contains the $j$-th sample. Notice, that when $x_j$ is removed the data is no more centered. We define the new means by

$$\overline{x^{(-j)}} = \frac{1}{n-1}\sum_{i\neq j}^{n} x_i, \quad \overline{y^{(-j)}} = \frac{1}{n-1}\sum_{i\neq j}^{n} y_i,$$

where the superscript is to emphasize the dependency on the excluded index $j$. Show that under Assumption 1, the new means for $x$ and $y$ are

$$\overline{x^{(-j)}} = -\frac{x_j}{n-1}, \quad \overline{y^{(-j)}} = \frac{n\bar{y} - y_j}{n-1}. \tag{2}$$

(c) Now this time, consider fitting a model of the form $y = \beta_0 + \beta_1 x$ to set 1. Show that under Assumption 1, the fitted value of $\beta_0$ and $\beta_1$ denoted by $\hat{\beta}_0^{(-j)}$ and $\hat{\beta}_1^{(-j)}$ can be acquired through

$$\hat{\beta}_1^{(-j)} = \frac{\sum_{i\neq j} x_i y_i - (n-1)\overline{x^{(-j)}}\ \overline{y^{(-j)}}}{\sum_{i\neq j} x_i^2 - (n-1)\overline{x^{(-j)}}^2}, \qquad \hat{\beta}_0^{(-j)} = \frac{n\bar{y} - y_j}{n-1} + \hat{\beta}_1^{(-j)}\frac{x_j}{n-1}.$$

(Hint: you do not need to minimize the RSS from scratch, the derivation should not be hard by a simple trick).

(d) By simplifying the results in part(c) show that

$$\hat{\beta}_1^{(-j)} = \frac{S_{xy} - \frac{n}{n-1}x_j(y_j - \bar{y})}{S_{xx} - \frac{n}{n-1}x_j^2}, \quad \text{and} \quad y_j - \hat{\beta}_0^{(-j)} = \frac{n}{n-1}(y_j - \bar{y}) - \frac{1}{n-1}\hat{\beta}_1^{(-j)} x_j.$$

(e) Now we want to test our model against set 2. For this purpose all you need to do is finding the squared difference between $y_j$ and the fitted model at $x_j$, which is $MSE_j = (y_j - \hat{\beta}_0^{(-j)} - \hat{\beta}_1^{(-j)} x_j)^2$. Using the result of part (d), perform the calculation and show that

$$MSE_j = \left( \frac{y_j - \hat{y}_j}{1 - h_j} \right)^2,$$

where

$$\hat{y}_j = \hat{\beta}_0 + \hat{\beta}_1 x_j, \quad \text{and} \quad h_j = \frac{1}{n} + \frac{x_j^2}{S_{xx}}.$$

This clearly implies that

$$CV_n = \frac{1}{n} \sum_{j=1}^{n} MSE_j = \frac{1}{n} \sum_{j=1}^{n} \left( \frac{y_j - \hat{y}_j}{1 - h_j} \right)^2. \tag{3}$$

(f) In the homework folder you have access to the data file `SimpleReg.csv`. The data contains a feature column $x$ and a response column $y$. Read the data, then center the $x$ data, and fit a linear model in the form of $y = \beta_0 + \beta_1 x$. Now use an R or Python program to calculate the LOOCV $CV_n$, as we did in the class (if you use R, pick the first element of `delta`). Also write a code that calculates the $CV_n$ using equation (3). You should see that the two methods produce identical results. You may also be surprised with how faster your customized code is, compared to the R `cv.glm` function!

**Q2.** In HW3 you did an exercise on performing a classification on the `HeartData.csv` data file. We recently learned how to apply the LOOCV and K-Fold CV to regression problems. In this homework we would like to apply the LOOCV and K-Fold CV to the logistic regression, LDA and QDA models used in HW3. Using 10-fold CV and LOOCV fit the models and report the classification accuracy for the 3 models (logistic regression, LDA and QDA). For this question use *num* as the response variable and all the other variables as features.

**Q3.** Using some basic calculus we can show that if we want to fit the simple model

$$y = \beta + \beta x,$$

to some data points $(x_1, y_1), \cdots, (x_n, y_n)$, the optimal choice of $\beta$ is (**you do not need to show this**)

$$\hat{\beta} = \frac{\sum_{i=1}^{n} (1 + x_i) y_i}{\sum_{i=1}^{n} (1 + x_i)^2}. \tag{4}$$

(a) Now consider a Ridge regression problem which requires minimizing

$$RSS_{Ridge} = \sum_{i=1}^{n} (y_i - \beta - \beta x_i)^2 + \lambda \beta^2.$$

Show that in this case the optimal selection of $\beta$ is

$$\hat{\beta}_R = \frac{\sum_{i=1}^{n} (1 + x_i) y_i}{\lambda + \sum_{i=1}^{n} (1 + x_i)^2}. \tag{5}$$

(b) If the true regression function is in the form of $f(x) = \beta + \beta x$ and we measure the noisy observations $y = \beta + \beta x + \epsilon$, where $\epsilon$ is a random variable with $\mathbb{E}\epsilon = 0$, $var(\epsilon) = \sigma^2$, show that

$$var(\hat{\beta}_R) = \frac{\sum_{i=1}^{n} (x_i + 1)^2}{\left( \lambda + \sum_{i=1}^{n} (x_i + 1)^2 \right)^2} \sigma^2.$$

(c) Going through basic steps (**you do not need to show this**), we can show that for the optimal value of $\beta$ in (4), $var(\hat{\beta})$ can be calculated as

$$var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^{n}(x_i + 1)^2},$$

show that for all $\lambda > 0$:

$$var(\hat{\beta}_R) \leq var(\hat{\beta}).$$

**Q4.** In this question we analyze the data file `Fertility.csv` which is available in the homework folder, and try to build a model for fertility. In this dataset Fertility, the first column, is the response variable, and the other variables are potential predictors. We will use several different statistical modeling techniques. The data set contains 47 rows (samples), split the data into training and test sets. Set the first 30 rows to training samples and the rows 31 through 47 as the test samples.

(a) Fit a linear model on the training set, and report the test error (MSE) obtained.

(b) Fit a Ridge regression model on the training set, with $\lambda$ chosen by cross-validation on a dense grid similar to the example solved in the class. Report the test error obtained.

(c) Fit a LASSO model on the training set, with $\lambda$ chosen by cross-validation on a dense grid similar to the example solved in the class. Report the test error obtained, along with the number of non-zero coefficient estimates.

(d) Compare the results of (a), (b), and (c). Which one seems to outperform the others for this specific setup?