# Homework 1

MSA 8150 - Machine Learning for Analytics (Spring 2020)
Instructor: Alireza Aghasi
Due date: Friday, Feb 7th (1 PM)

January 27, 2020

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.** (a) Consider the following system of two equations, where $x$ and $y$ are the unknowns:

$$\begin{cases} a_1x + b_1y = c_1 \\ a_2x + b_2y = c_2 \end{cases}. \tag{1}$$

Show that if $a_1b_2 - a_2b_1 \neq 0$, then simultanuosly solving the system above for $x$ and $y$ yields

$$x = \frac{c_1b_2 - b_1c_2}{a_1b_2 - a_2b_1}, \qquad y = \frac{a_1c_2 - c_1a_2}{a_1b_2 - a_2b_1}.$$

(b) The goal is to find the optimal values of $\beta_1$ and $\beta_2$ which fit the general model

$$y = \beta_1 g(x) + \beta_2 f(x) \tag{2}$$

to the data points $(x_1, y_1), \cdots, (x_n, y_n)$. Here $f(x) \in \mathbb{R}$ and $g(x) \in \mathbb{R}$ are some arbitrary functions of $x$. Use the result in part (a) to mathematically show that these optimal values are

$$\hat{\beta}_1 = \frac{\left(\sum_{i=1}^n y_i g_i\right)\left(\sum_{i=1}^n f_i^2\right) - \left(\sum_{i=1}^n y_i f_i\right)\left(\sum_{i=1}^n g_i f_i\right)}{\left(\sum_{i=1}^n g_i^2\right)\left(\sum_{i=1}^n f_i^2\right) - \left(\sum_{i=1}^n f_i g_i\right)^2}, \tag{3}$$

$$\hat{\beta}_2 = \frac{\left(\sum_{i=1}^n y_i f_i\right)\left(\sum_{i=1}^n g_i^2\right) - \left(\sum_{i=1}^n y_i g_i\right)\left(\sum_{i=1}^n g_i f_i\right)}{\left(\sum_{i=1}^n g_i^2\right)\left(\sum_{i=1}^n f_i^2\right) - \left(\sum_{i=1}^n f_i g_i\right)^2}. \tag{4}$$

For brevity we used the notations $g_i = g(x_i)$ and $f_i = f(x_i)$.

**Hint:** In the class we went through the exercise of showing that to fit a simple linear model

$$y = \beta_0 + \beta_1 x$$

to the data points $(x_1, y_1), \cdots, (x_n, y_n)$, the optimal choice of parameters are

$$\hat{\beta}_1 = \frac{\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^n (x_i - \bar{x})^2}, \quad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}.$$

You should be able to follow a similar process to derive the equations (3) and (4).

(c) Now for the model defined in (2), assume that $g(x) = 0$, which reduces our model to $y = \beta f(x)$. Show that fitting this simplified model to the data points $(x_1, y_1), \cdots, (x_n, y_n)$, yields the following expression for the optimal value of $\beta$:

$$\hat{\beta} = \frac{\sum_{i=1}^n f_i y_i}{\sum_{i=1}^n f_i^2}. \tag{5}$$

(d) Is it possible to derive (5) by simply setting $g(x_i) = g_i = 0$ in (4)?

(e) In part (c) you showed that assuming that our data follows the model $y = \beta f(x) + \epsilon$, where $\epsilon$ is zero mean iid Gaussian noise, then the best estimate for $\hat{\beta}$ is obtained via (5). Using the basic properties of the expectation mentioned in the class, prove that $\hat{\beta}$ is an unbiased estimate of $\beta$. Basically you would need to show that

$$\mathbb{E}(\hat{\beta}) = \beta.$$

(f) Assume that for the true regression function we have $g(x) = 0$ and $f(x) = x^2$, meaning that our observations are in the form of $y_i = \beta x_i^2 + \epsilon_i$. Moreover, assume that $\epsilon_i$ represents i.i.d zero-mean noise, that is $\mathbb{E}\epsilon_i = 0$, $var(\epsilon_i) = \sigma^2$. Use (5) to derive the optimal fit $\hat{\beta}$ for the model $y = \beta x^2$, and then using the properties mentioned in the class about the variance of the sum of independent random variables, show that:

$$var(\hat{\beta}) = \frac{\sigma^2}{\sum_{i=1}^n x_i^4}.$$

(g) Using the outcome of part (e), is it true that for the model $y = \beta x^2$, having samples with larger $x$ can reduce the uncertainty in evaluating $\hat{\beta}$? (uncertainty has a similar interpretation as the variance here)

**Q2.** The goal of this question is calculating the body fat using the Brozek's equation. In this problem the first column (`brozek`) is the response variable and all other columns are treated as the features (we may not use all the features, please only use the ones stated in each question). To start, read the data file fat.csv in the homework folder (containing 252 samples) and split it into two sets. Set 1 includes the first 200 rows of the data (do not count the row associated with the feature/response names), and set 2, which includes the last 52 rows of the data. Name the first set `train` and the second set `test`.

(a) As a first modeling attempt, consider a linear model using all the 17 features, that is

$$\texttt{brozek} = \beta_0 + \beta_1 \texttt{siri} + \ldots + \beta_{17} \texttt{wrist} \tag{6}$$

report the fitted parameters, the 95% confidence interval for each estimated parameter and the p-values. What is the $R^2$ value, and based on that how good do you see the model fitting the training data?

(b) Based on $\alpha = 0.05$ and the calculated p-values, which features seem problematic?

(c) Calculate the prediction error using your test file. If $\hat{y}_i$ is the prediction and $y_i$ is the actual brozek value in the test file, the error for this model is calculate as follows:

$$e_1 = \sqrt{\sum_{i=1}^{52} (y_i - \hat{y}_i)^2}$$

(d) Now consider a second model which uses the features indicated in part (a), with the only difference that density is replaced by inverse density. In other words, your model has the the following parametric form:

$$\texttt{brozek} = \beta_0 + \beta_1 \texttt{siri} + \frac{\beta_2}{\texttt{density}} + \beta_3 \texttt{age} \ldots + \beta_{17} \texttt{wrist}. \tag{7}$$

Repeat the steps in part (a) and report the values.

(e) Use a similar formulation in part (c) and calculate the test error (we will call this test error $e_2$ which corresponds to our second model).

(f) Now consider a third model which only uses `siri` and `density` as the features, but follows a parametric formulation as:

$$\texttt{brozek} = \beta_0 + \beta_1 \texttt{siri} + \beta_2 \texttt{siri}^2 + \frac{\beta_3}{\texttt{density}} + \beta_4 \texttt{density}. \tag{8}$$

Repeat the steps in part (a) and report the values.

(g) Based on $\alpha = 0.05$ and the calculated p-values, which features seem problematic?

(h) Repeat part (c) for this model and call the error $e_3$.

(i) Based on the values $e_1$, $e_2$ and $e_3$, and the model formulations, which model would you pick and why (state two reasons)?

Please hand in your code along with comprehensive response to each part of the question.

**Q3.** In the previous question we observed some redundancy in the features. We would like to try some feature selection heuristic in this question. Consider the same dataset as question 2 (fat.csv), where `brozek` is the response variable and the other 17 columns are the model features. Follow this steps below.

- Form an extended version of the dataset, by appending two more columns. One column corresponding to $\texttt{siri}^2$ and one column corresponding to $\frac{1}{\texttt{density}}$. Your extended dataset should now have 20 columns, where the first column is `brozek` and used as the response variable, 17 columns identical to the original fat.csv data set, and columns 19 and 20 with the values $\texttt{siri}^2$ and $\frac{1}{\texttt{density}}$, respectively. We will refer to this dataset as the *extended dataset*.
- In a similar way as question 2, split the extended dataset into two sets. Set 1 includes the first 200 rows of the data (do not count the row associated with the feature/response names), and set 2, which includes the last 52 rows of the data. Name the first set `train` and the second set `test`.

(a) Use the training data to fit a model of the following form

$$\texttt{brozek} = \beta_0 + \beta_1\texttt{siri} + \ldots + \beta_{17}\texttt{wrist} + \beta_{18}\texttt{siri}^2 + \frac{\beta_{19}}{\texttt{density}} \tag{9}$$

report the fitted parameters, the 95% confidence interval for each estimated parameter and the p-values. What is the $R^2$ value?

(b) Use the test data to calculate the test error (similar to the formulation in part (c) of the previous question), and call it $e_{full}$.

(c) Let's run a heuristic scheme to perform feature selection (the method is called backward selection and described on page 79 of your textbook, also on the slides). Start with the full model (the model containing all 19 features of the extended dataset) and drop the feature with the highest p-value (or the second largest if the largest p-value is for the intercept), then redo the modeling and drop the next feature with the highest p-value, and continue dropping until all p-values are small and you are left with a set of important features. Implement this approach and stop when all p-values are below 0.03. Which features are selected as the most important ones when your code stops?

(d) Apply the model developed in part (c) to the test data and call the error $e_{sel}$.

(e) Compare $e_{full}$ and $e_{sel}$. Does the feature selection scheme seem to reduce overfitting?

(f) Compare $e_{sel}$ with $e_3$ from part (h) of question 2. In terms of the test accuracy does your feature selection scheme seem to find the best model?

Please hand in your code along with comprehensive response to each part of the question.