Homework 1

MSA 8150 - Machine Learning for Analytics (Spring 2019)
Instructor: Alireza Aghasi
Due date: Feb 12th (7 PM)

February 2, 2019

Please revise the homework guidelines reviewed during the first session. 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 totally manageable given the material lectured in the class. The long questions are to help clarifying the problem, the majority of the solutions are short

Q1. (a) Find the optimal value of β which fits the simple model

$$y = \frac{\beta}{x} \tag{1}$$

to the data points $(x_1, y_1), \cdots, (x_n, y_n)$. Mathematically justify your answer.
Hint: In the class we went through some basic calculus to show 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 derive the optimal expression for β in (1) if you use a similar technique.

(b) In practice if the true regression function is in the form of $f(x) = \frac{\beta}{x}$ then we measure the noisy observations $y_i = \frac{\beta}{x_i} + \epsilon_i$, where ϵ_i represents i.i.d zero-mean noise, that is $\mathbb{E}\epsilon_i = 0$, $var(\epsilon_i) = \sigma^2$. Mathematically show that if $\hat{\beta}$ is the optimal parameter for part (a), then

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

Hint: Note that in this problem $x_1, \cdots x_n$ are fixed (deterministic) quantities but y_i and $\hat{\beta}$ are random quantities because of their dependence to ϵ_i.

(c) In this part you generalize the result of part (a). Find the optimal value of β which for an arbitrary function $g(x)$ fits the model

$$y = \beta g(x)$$

to the data points $(x_1, y_1), \cdots, (x_n, y_n)$. Mathematically justify your answer.

(d) Assume that the true regression function is in the form of $f(x) = \beta g(x)$ and we measure the noisy observations $y_i = \beta g(x_i) + \epsilon_i$, with the noise characteristics in part (b). Mathematically show that if $\hat{\beta}$ is the optimal parameter for part (c), then

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

(e) Explain why to reduce the uncertainty of $\hat{\beta}$ (variance is a measure of uncertainty) it is better to have samples where $g(x_i)$ is large.

Q2. Newton's law of universal gravitation states that every particle attracts every other particle in the universe with a force which is directly proportional to the product of their masses and inversely proportional to the square of the distance between their centers (reference: Wikipedia).

In the homework we provide some data related to this physical phenomenon and try to formulate an equation for the law of gravity. In the homework folder you would have access to a data file named `GravityForce.csv`. This is a data file with 3 features `MASS1` `(`m_1`),` `MASS2` `(`m_2`)` , `DISTANCE` (r) and a response column `Force` `(`F`)`. The data correspond to some measurements of each quantity and the corresponding measured force among the objects.

– Read the data file 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 40 rows of the data. Name the first set `train` and the second set `test`.

(a) Using the training data, fit a linear regression model as

$$F = \mathcal{M}_1(m_1, m_2, r) = \beta_0 + \beta_1 m_1 + \beta_2 m_2 + \beta_3 r, \tag{2}$$

report the fitted parameters, the 95% confidence interval for each estimated parameter, the p-values and the R^2 statistic. Explain what the R^2 statistic tells you.

(b) Based on the p-values you observe ($\alpha = 0.05$), do you see any significance problem with any of the features?

(c) Use the fitted model and pass the features in your test file to generate the corresponding response $\boldsymbol{F}_{test}^{pred}$ (a vector of length 40). Now compare this quantity with the original responses in your test file $\boldsymbol{F}_{test}$ by finding the Euclidean distance between them:

$$E_1 = \|\boldsymbol{F}_{test}^{pred} - \boldsymbol{F}_{test}\| = \sqrt{\sum_{i=1}^{40} \left(F_{test\ i}^{pred} - F_{test i}\right)^2}.$$

(d) Let's consider a different path and try to fit a model that relates the log-values of the features to the log-value of the response, i.e.,

$$\log F = \mathcal{M}_2(\log m_1, \log m_2, \log r) = \gamma_0 + \alpha_1 \log m_1 + \alpha_2 \log m_2 + \alpha_3 \log r, \tag{3}$$

(the log is natural logarithm) report the fitted parameters, the 95% confidence interval for each estimated parameter and the p-values.

(e) Use the fitted model $\mathcal{M}_2$ and pass the log-features in your test file to generate the corresponding response $\boldsymbol{LF}_{test}^{pred}$ (a vector of length 40). Compare the predicted response with the original response in your test file and evaluate the Euclidean distance via

$$E_2 = \|\exp(\boldsymbol{LF}_{test}^{pred}) - \boldsymbol{F}_{test}\| = \sqrt{\sum_{i=1}^{40} \left(\exp(F_{test\ i}^{pred}) - F_{test i}\right)^2}.$$

(note that we had to use an exp function to compare both quantities in the original non-log domain).

(f) Now consider a hybrid model, where

$$\log F = \mathcal{M}_3(\log q_1, \log q_2, \log r, r) = \gamma_0 + \alpha_1 \log q_1 + \alpha_2 \log q_2 + \alpha_3 \log r + \alpha_4 r, \tag{4}$$

report the fitted parameters, the 95% confidence interval for each estimated parameter and the p-values.

(g) Repeat part (e) for the fitted model $\mathcal{M}_3$ and call the resulting Euclidean distance error E_3.

(h) Compare E_1 and E_2 and E_3. Which model do you think has a better prediction power.

(i) Can you make a connection between $\mathcal{M}_2$ and the original Newton's law of gravity:

$$F = G\frac{m_1 m_2}{r^2}$$

Can you reformulate your $\mathcal{M}_2$ to a form similar to Newton's law? How similar is your reformulation to that?

Hint: When you do the reformulation you probably end up with an equation like

$$F = k_0 \frac{m_1^{k_1} m_2^{k_2}}{r^{k_3}}$$

and we would like to see if k_1 and k_2 turn out to be close to 1 and k_3 close to 2.

(j) Years after Newton, Laplace suggested an extended version of Newton's model where

$$F = G\frac{m_1 m_2}{r^2}\exp(-\alpha r) \tag{5}$$

Can you reformulate your $\mathcal{M}_3$ to a form similar to the Newton-Laplace law in (5)? How similar is your reformulation to that?

Hint: When you do the reformulation you probably end up with an equation like

$$F = k_0 \frac{m_1^{k_1} m_2^{k_2}}{r^{k_3}}\exp(-k_4 r)$$

and we would like to see if k_1 and k_2 turn out to be close to 1 and k_3 close to 2.

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

Q3. In the homework folder you would have access to a data file named `Polarization.csv`. The file corresponds to some wave field measurements through a material. The matter causes a change in the polarization (and magnitude) of the field when the temperature changes. The data file contains 1100 measurements of the temperature and the corresponding field measurement. The goal is to find a simple model which reliably relates the temperature to the filed value, i.e.,

$$field = G(temp)$$

Our goal is to have G in a polynomial form, and at most of order 10, i.e.,

$$G(temp) = \beta_0 + \beta_1 temp + \beta_2 temp^2 + \ldots \beta_{10} temp^{10}$$

- Read the data file and split it into two sets. Set 1 includes the first 500 rows of the data (do not count the row associated with the feature/response names), and set 2, which includes the last 600 rows of the data. Name the first set `train` and the second set `test`.

– While we allow polynomial fits of up to degree 10, we ultimately want to develop a simple model which only contains few powers of *temp* (i.e., some of the β_i's are zero). Since we do not know which powers are the most key components to describe the model, we ultimately would like to apply a feature selection scheme.

(a) Use the training data to develop a model of the form

$$field = \beta_0 + \beta_1 temp + \beta_2 temp^2 + \dots \beta_{10} temp^{10}$$

report the fitted parameters, the 95% confidence interval for each estimated parameter, the p-values and the R^2 statistic. Are all the p-values in the safe range? Explain what the R^2 statistic tells you.

(b) Use the fitted model and apply it to the test data to generate the corresponding response $\boldsymbol{f}_{test}^{pred}$ (a vector of length 600). Now compare this quantity with the original responses in your test file $\boldsymbol{f}_{test}$ by finding the Euclidean distance between them:

$$E_1 = \|\boldsymbol{f}_{test}^{pred} - \boldsymbol{f}_{test}\| = \sqrt{\sum_{i=1}^{600} \left(f_{test\ i}^{pred} - f_{test i} \right)^2}.$$

(c) An earlier scientific study shows that the best model to pick takes the following form

$$field = \beta_0 + \beta_3 temp^3 + \beta_6 temp^6.$$

Use the training data to estimate β_0, β_3 and β_6 in the proposed model, report the fitted parameters, the 95% confidence interval for each estimated parameter, the p-values and the R^2 statistic. Are all the p-values in the safe range? Compare the R^2 statistic of this model with the full model in part (a). Also similar to part (b) apply your model to the test data and calculate

$$E_2 = \|\boldsymbol{f}_{test}^{pred} - \boldsymbol{f}_{test}\|$$

for this model.

(d) Your textbook suggests a scheme for feature selection. It suggests picking a full model (a model containing all ten powers of *temp*) and removing the variable with the highest p-value (or the second largest if the largest p-value was for the intercept), then redo the modeling and remove the next variable with the highest p-value, and continue removing until all p-values are small (what is called backward selection and described on page 79, also in the slides). Implement this approach for the problem above and stop when all p-values are below 0.01? Which features are selected as the most important ones when your code stops? Is this model close to the model suggested in part (c)?

(e) Apply the model developed in part (d) to the test data and calculate

$$E_3 = \|\boldsymbol{f}_{test}^{pred} - \boldsymbol{f}_{test}\|$$

for the corresponding model. How do E_1, E_2 and E_3 compare? Which model do you suggest to pick?

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