# STA243 2020 Computational Statistics
# Homework 2

**Instructor:** Krishna Balasubramanian
**Author1:** Jian Shi 917859483
**Author2:** Bohao Zou 917796070

**2020-05-06**

## Question 1.

$(i) : f(\theta) : \mathbf{R}^d \to \mathbf{R}$ is convex
$(ii) : f(\theta_2) \geq f(\theta_1) + \nabla f(\theta_1)^T(\theta_2 - \theta_1)$

$(i) \Rightarrow (ii)$ If $f$ is convex, by definition

$$f(\lambda\theta_2 + (1-\lambda)\theta_1) \leq \lambda f(\theta_2) + (1-\lambda)f(\theta_1), \forall\lambda \in [0,1], \theta_1, \theta_2 \in \text{dom}(f)$$

After rewriting, we have

$$\begin{aligned} &f(\lambda\theta_2 + (1-\lambda)\theta_1) \leq f(\theta_2) + \lambda(f(\theta_1) - f(\theta_2)) \\ \Rightarrow &f(\theta_1) - f(\theta_2) \geq \frac{f(\theta_2 + \lambda(\theta_1 - \theta_2)) - f(\theta_2)}{\lambda}, \forall\lambda \in (0,1] \end{aligned}$$

As $\lambda \to 0$, we get

$$\begin{aligned} f(\theta_1) - f(\theta_2) &\geq \nabla f^T(\theta_2)(\theta_1 - \theta_2) \\ \Rightarrow f(\theta_2) &\geq f(\theta_1) + \nabla f(\theta_1)^T(\theta_2 - \theta_1) \end{aligned} \tag{1}$$

$(ii) \Rightarrow (i)$ Suppose (1) holds $\forall\theta_1, \theta_2 \in dom(f)$. Take any $\theta_1, \theta_2 \in dom(f)$ and let

$$z = \lambda\theta_2 + (1-\lambda)\theta_1$$

We have

$$f(\theta_2) \geq f(z) + \nabla f^T(z)(\theta_2 - z) \tag{2}$$
$$f(\theta_1) \geq f(z) + \nabla f^T(z)(\theta_1 - z) \tag{3}$$

Multiplying (2) by $\lambda$, (3) by $(1-\lambda)$ and adding, we get

$$\begin{aligned} \lambda f(\theta_2) + (1-\lambda)f(\theta_1) &\geq f(z) + \nabla f^T(z)(\lambda\theta_2 + (1-\lambda)\theta_1 - z) \\ &= f(z) \\ &= f(\lambda\theta_2 + (1-\lambda)\theta_1) \end{aligned}$$

So, $f$ is convex.

## Question 2.

The origin of the dataset housingprice.csv we will use in this question is from the Coursera open course Machine Learning Foundations: A Case Study Approach by Prof. Carlos Guestrin and Prof. Emily Fox. Load the training data train.data.csv and testing data test.data.csv. We'll build our regression model on the training data and evaluate the model on the testing data.

**(a)** Build a linear model on the training data using lm() by regessing the housing price on these variables: bedrooms, bathrooms, sqft living, and sqft lot. What's the $R^2$ of the model on training data? What's the $R^2$ on testing data?

## Solution

By building the linear model on *bedrooms, bathrooms, sqft living, and sqft lot* variables, we can get the linear model (a). Than we can get the $R^2$ of training data and testing data. The $R^2$ of training data set is 0.5101139. The $R^2$ of testing data set is 0.5049945.

| | Training data | Testing data |
|---|---|---|
| $R^2$ | 0.5101139 | 0.5049945 |

Table 1: *The $R^2$ of model (a) and each data set.*

**(b)** The image below is Bill Gates' house. Load the file fancyhouse.csv to obtain the features of the house. Guess the price of his house using your linear model. Do you think the predicted price is reasonable?

## Solution

By using the linear model (a), we can guess the price of Bill Gates' house. The final result is 15436770 US dollars. In my consideration, the price is reasonable because from the picture we can know that this house is a villa. This tells us this house is not cheap.

**(c)** Let's continue to improve the linear model we have. Instead of throwing only the raw data into the statistical model, we might want to use our intuition and domain expertise to extract more meaningful features from the raw data. This step is called feature engineering. Using meaningful features in the model is often crucial for successful data analysis. Add another variable by multiplying the number of bedrooms by the number of bathrooms, which describes the combined benefit of having more bedrooms and bathrooms. Add this variable to the linear model we have in Part (a). What's the $R^2$ of the new model on the training data and testing data respectively?

## Solution

By adding another variable which is multiplied the number of bedrooms by the number of bathrooms into the linear model in (a), we can get a new model (c). The $R^2$ of of training data set is 0.5173533. The $R^2$ of testing data set is 0.5105355.

| | Training data | Testing data |
|---|---|---|
| $R^2$ | 0.5173533 | 0.5105355 |

Table 2: *The $R^2$ of model (c) and each data set.*

**(d)** Perform all the things above without using the in-built function lm() in R, but by using gradient descent algorithm on the sample-based least-squares objective function, to estimate the OLS regression parameter vector. How does your result compare to the result from previous part ? Note that you have to set the tuning parameter appropriately for this method.

## Solution

At this question, we need to use the gradient descent algorithm on the sample-based least-squares objective function to get the estimation of the coefficients of linear models in question (a) and question (c) and calculates the $R^2$ of each models.
We have tried to use the original training data set to gradient descent algorithm. The result is that if we use a big $\eta_t$ to GD, the result will be $Nan$. This phenomenon indicates that we should use a smaller $\eta_t$ to GD algorithm. However, if we used the small $\eta_t$ to GD algorithm, the result would not convergence. This indicates that we should use a bigger $\eta_t$. Those contradiction phenomenons told us we should not use the original data set to GD and we have found that the gradient of initial $\hat{\beta}$ of one element ($\hat{\beta}_5$) which the truth estimated value is the smallest has the biggest gradient which is $1.02 \times 10^{12}$. It is too huge to convergence.
For dealing those questions, we used a normalization to all $X$ variables and response variable $Y$. The formula of this normalization is :

$$X_{new} = \frac{1}{\sqrt{n-1}} \frac{X - \bar{X}}{sd(X)}$$

For ensuring convergence, we set the condition that stop the while loop is that

$$\frac{||\hat{\beta}_{Truth} - \hat{\beta}_{Current}||_2^2}{||\hat{\beta}_{Truth}||_2^2} <= \tau, \qquad where \ \tau = 0.01$$

By setting the stop condition, using normalized training data set and set the learning rate as $\eta_t = 1 \times 10^{-5}$. We get the $R^2$ of model in (a) and model in (c). The result will show in the table.

| | Model (a) | Model (c) |
|---|---|---|
| $R^2$ of training data set | 0.5100987 | 0.5173506 |
| $R^2$ of testing data set | 0.5053426 | 0.5108538 |

Table 3: *The $R^2$ of each model and each data set by using gradient descent algorithm.*

The relationship between the value of $\frac{||\hat{\beta}_{Truth} - \hat{\beta}_{Current}||_2^2}{||\hat{\beta}_{Truth}||_2^2}$ and the iteration is showing on the plot.

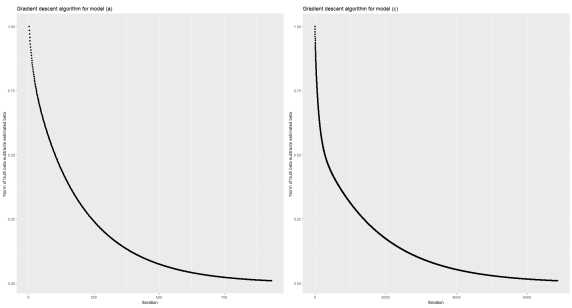


Figure 1: *The relation ship of the norm value and iterations. The left plot is for model (a), and the right plot is for model (c).*

**(e)** Perform all the things above now using stochastic gradient descent (with one sample in each iteration). How does your result compare to the result from previous parts ? Note: while running stochastic gradient descent, you can sample without replacement and when you run out of samples, just start over. Note that you have to set the tuning parameter appropriately for this method.

## Solution

At this question, we need to re-estimate the $R^2$ of model (a) and model (c) by using stochastic gradient descent with one sample in each iteration. We also used the normalized training data set to the SGD algorithm. However, the SGD algorithm is more difficult to convergence than GD algorithm. So, we released the stop condition that

$$\frac{||\hat{\beta}_{Truth} - \hat{\beta}_{Current}||_2^2}{||\hat{\beta}_{Truth}||_2^2} <= \tau, \qquad where\ \tau = 0.02$$

. The result will show in the table.

| | Model (a) | Model (c) |
|---|---|---|
| $R^2$ of training data set | 0.5095858 | 0.5166523 |
| $R^2$ of testing data set | 0.5050082 | 0.5105888 |

Table 4: *The $R^2$ of each model and each data set by using stochastic gradient descent algorithm.*

The relationship between the value of $\frac{||\hat{\beta}_{Truth}-\hat{\beta}_{Current}||_2^2}{||\hat{\beta}_{Truth}||_2^2}$ and the iteration is showing on the plot.

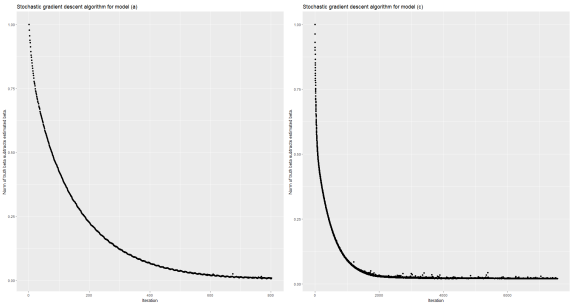


Figure 2: *The relation ship of the norm value and iterations. The left plot is for model (a), and the right plot is for model (c).*

From above analysis, we can know that the **BEST** estimation of coefficients of $\beta$ is by using the formula $\hat{\beta} = (X^T X)^{-1} X^T Y$. The **SECOND** wonderful estimation is by using gradient descent algorithm. The **THIRD** great estimation is by using stochastic gradient descent algorithm.
This result is in line with expectations. Obviously, the best estimation of $\beta$ is that uses exact formula to estimate. For the GD and SGD algorithm, because the gradient descent algorithm used all samples to estimate the gradient. It is more accurate and easier to convergence than using one sample to estimate the gradient.
However, the $R^2$ of test data set which calculates by using SGD and GD algorithm is higher than the $R^2$ of test data set which calculates by using formula. This may because that if we used formula to calculate the $\hat{\beta}$, we may have more overfitted situation on model. But, the random factor of GD and SGD releases the overfitted situation on model.

**Question 3.**

**Fact 6.1.1.** If the function is $\mu-$ strongly convex, then $(\nabla f(\boldsymbol{\theta}_1)-\nabla f(\boldsymbol{\theta}_2))^T(\boldsymbol{\theta}_1-\boldsymbol{\theta}_2) \geq \mu||\boldsymbol{\theta}_1-\boldsymbol{\theta}_2||_2^2$

**Proof:**

Assume $g(\boldsymbol{x})) = f(\boldsymbol{x}) - \frac{\mu}{2}||\boldsymbol{x}||_2^2$, where $\mu = \lambda_{min}(\nabla^2 f(\boldsymbol{x}))$, we can have

$$\nabla g(\boldsymbol{x}) = \nabla f(\boldsymbol{x}) - \frac{\mu}{2}(I + I^T)\boldsymbol{x}$$
$$\nabla^2 g(\boldsymbol{x}) = \nabla^2 f(\boldsymbol{x}) - \mu I$$

The diagonal elements of $\nabla^2 g(\boldsymbol{x})$ are non-negative, which means that $\nabla^2 g(\boldsymbol{x})$ is positive semi-definite. So, $g(\boldsymbol{x})$ is convex and we can have:

$$g(\boldsymbol{\theta}_1) \geq g(\boldsymbol{\theta}_2) + \nabla g(\boldsymbol{\theta}_2)^T(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \tag{1}$$
$$g(\boldsymbol{\theta}_2) \geq g(\boldsymbol{\theta}_1) + \nabla g(\boldsymbol{\theta}_1)^T(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1) \tag{2}$$

Add up (1) and (2) and there is :

$$\begin{aligned}
&(\nabla g(\boldsymbol{\theta}_1)^T - \nabla g(\boldsymbol{\theta}_2)^T)(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \geq 0 \\
\Rightarrow &(\nabla f(\boldsymbol{\theta}_1) - \nabla f(\boldsymbol{\theta}_2) - \mu\boldsymbol{\theta_1} + \mu\boldsymbol{\theta_2})^T(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \geq 0 \\
\Rightarrow &(\nabla f(\boldsymbol{\theta}_1) - \nabla f(\boldsymbol{\theta}_2))^T(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \geq (\mu\boldsymbol{\theta_1} - \mu\boldsymbol{\theta_2})^T(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \\
\Rightarrow &(\nabla f(\boldsymbol{\theta}_1) - \nabla f(\boldsymbol{\theta}_2))^T(\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2) \geq \mu||\boldsymbol{\theta}_1 - \boldsymbol{\theta}_2||_2^2
\end{aligned}$$

So, **Fact 6.1.1.** is proved.

Then, we start to solve the recursion to obtain the final result of Theorem 6.1 and we have had:

$$\mathrm{E}\left[\left\|\boldsymbol{\theta}^{(t+1)} - \boldsymbol{\theta}^*\right\|_2^2\right] \leq \left(1 - 2\mu\eta_t + \eta_t^2 M_g L^2\right)\mathrm{E}\left[\|\boldsymbol{\theta}^{(t)} - \boldsymbol{\theta}^*\|_2^2\right] + \eta_t^2\sigma_{\mathrm{g}}^2 \tag{3}$$

Assume $-1 + 2\mu c - c^2 M_g L^2 > 0$, and let

$$c_0 \geq \max\left(\frac{c^2\sigma_g^2}{-1 + 2\mu c - c^2 M_g L^2}, \left\|\theta^{(0)} - \theta^*\right\|_2^2\right)$$

For $t = 0, \mathrm{E}\left[\|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^*\|_2^2\right] = \|\boldsymbol{\theta}^{(0)} - \boldsymbol{\theta}^*\|_2^2 \leq c_0$,

For $t \geq 1$, assume that Theorem 6.1 holds when $t = k$ that:

$$\mathrm{E}\left[\left\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}^*\right\|_2^2\right] \leq \frac{c_0}{k+1},$$

and when $t = k + 1$,

$$
\begin{aligned}
\mathrm{E}\left[\left\|\boldsymbol{\theta}^{(k+1)}-\boldsymbol{\theta}^{*}\right\|_2^2\right] &\leq \left(1-2\mu\eta_k+\eta_k^2 M_g L^2\right)\mathrm{E}\left[\|\boldsymbol{\theta}^{(k)}-\boldsymbol{\theta}^{*}\|_2^2\right]+\eta_k^2\sigma_{\mathrm{g}}^2 \\
&\leq \left(1-2\mu\eta_k+\eta_k^2 M_g L^2\right)\frac{c_0}{k+1}+\eta_k^2\sigma_g^2 \\
&= \left(1-2\frac{\mu c}{k+1}+\frac{c^2 M_g L^2}{(k+1)^2}\right)\frac{c_0}{k+1}+\frac{\sigma_g^2 c^2}{(k+1)^2} \\
&= \frac{c_0}{k+2}\left[\left(1-2\frac{\mu c}{k+1}+\frac{c^2 M_g L^2}{(k+1)^2}\right)\frac{k+2}{k+1}+\frac{(k+2)\sigma_g^2 c^2}{c_0(k+1)^2}\right] \\
&= \frac{c_0}{k+2}\frac{k+2}{(k+1)^2}\left[(k+1)-2\mu c+\frac{c^2 M_g L^2}{(k+1)}+\frac{\sigma_g^2 c^2}{c_0}\right] \\
&\leq \frac{c_0}{k+2}\frac{k+2}{(k+1)^2}\left[(k+1)-2\mu c+\frac{c^2 M_g L^2}{(k+1)}+\frac{\sigma_g^2 c^2}{\frac{c^2\sigma_g^2}{-1+2\mu c-c^2 M_g L^2}}\right] \\
&= \frac{c_0}{k+2}\frac{k+2}{(k+1)^2}\left[(k+1)-2\mu c+\frac{c^2 M_g L^2}{(k+1)}-1+2\mu c-c^2 M_g L^2\right] \\
&= \frac{c_0}{k+2}\frac{k+2}{(k+1)^2}\left[k+\frac{(-k)c^2 M_g L^2}{(k+1)}\right] \\
&\leq \frac{c_0}{k+2}\frac{k(k+2)}{(k+1)^2} \\
&= \frac{c_0}{k+2}\frac{k^2+2k}{k^2+2k+1} \\
&\leq \frac{c_0}{k+2}
\end{aligned}
$$

So, the theorem also holds when $t = k+1$.

Therefore, Theorem 6.1 holds out under the assumptions

$$
c_0 \geq \max\left(\frac{c^2\sigma_g^2}{-1+2\mu c-c^2 M_g L^2}, \left\|\theta^{(0)}-\theta^{*}\right\|_2^2\right), -1+2\mu c-c^2 M_g L^2 > 0
$$

if $\eta_t = \frac{c}{t+1}$ for some $c > 0$, then we have

$$
\mathrm{E}\left[\left\|\boldsymbol{\theta}^{(t)}-\boldsymbol{\theta}^{*}\right\|_2^2\right] \leq \frac{c_0}{t+1}
$$

**Pledge:**
Please sign below (print full name) after checking (✓) the following. If you can not honestly check each of these responses, please email me at kbala@ucdavis.edu to explain your situation.

- We pledge that we are honest students with academic integrity and we have not cheated on this homework.
- These answers are our own work.

- We did not give any other students assistance on this homework.
- We understand that to submit work that is not our own and pretend that it is our is a violation of the UC Davis code of conduct and will be reported to Student Judicial Affairs.
- We understand that suspected misconduct on this homework will be reported to the Office of Student Support and Judicial Affairs and, if established, will result in disciplinary sanctions up through Dismissal from the University and a grade penalty up to a grade of “F" for the course.

Team Member 1: Jian Shi Team Member 2: Bohao Zou