Regularization in Data Mining

Liang Ma
2019/02/27
Content

• Overfitting
• Regularization and some concepts
• Ridge Regression
• Lasso Regression
• R code for Ridge Regression and Lasso Regression
Overfitting

• An example to explain Overfitting
• The definition of overfitting
• How to address overfitting
Example: What is overfitting?

Underfitting or High Bias

Overfitting or High Variance
Definition

• Overfitting is "the production of an analysis that corresponds too closely or exactly to a particular set of data, and may therefore fail to fit additional data or predict future observations reliably".

------Wikipedia

• That is, if we have too many features in our data set, the learned hypothesis may fit the training set very well, but fail to generalize to new examples.
Example: What is overfitting?
How to address overfitting

Options:

• Collect more data
• Reduce number of features
• Regularization
Regularization

• The definition of regularization
• Some concepts we need to know
• ...
Definition

• Regularization is the process of adding information in order to solve an ill-posed problem or to prevent overfitting.

• That is, regularization is to reduce the complexity of the model by adjusting the model parameters to achieve the effect of avoiding overfitting.

------Wikipedia
Assume a linear equation is as follows:
\[ \hat{h}_\theta = \theta_0 + \theta_1 x \]

And the cost function is:
\[ J(\theta) = \frac{1}{m} \sum_{i=1}^{m} (\theta_0 + \theta_1 x - y^{(i)})^2 \]

The cost function is the mean square error function (MSE), where \( m \) represents the sample size.

Generalized linear regression cost function is:
\[ J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{h}_\theta (x^{(i)}) - y^{(i)})^2 \]
Linear regression models are often fitted using the least squares approach.

Least Squares:

"Least squares" means that the overall solution minimizes the sum of the squares of the residuals made in the results of every single equation.

------Wikipedia
the sum of the squares of the residuals

\[ J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{h}_\theta(x^{(i)}) - y^{(i)})^2 \]
Ridge Regression

...
In regularization, we usually use two algorithms, **Ridge Regression** and **Lasso Regression**.

Regularization is achieved by adding different constraints to the parameter after the cost function of linear regression. (here I take linear regression as an example.)
\[ J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{h}_{\theta}(x^{(i)}) - y^{(i)})^2 \]
\[ j(\theta) = \frac{1}{2m} \sum_{i=1}^{m} \left( h_{\theta}(x^{(i)}) - y^{(i)} \right)^2 \]
\[ J(\theta) = \frac{1}{2m} \sum_{i=1}^{m} (\hat{h}_\theta(x^{(i)}) - y^{(i)})^2 \]
Linear Regression using least Squares minimizes the sum of the squares of the residuals

Regularization using Ridge Regression minimizes the sum of the squares of the residuals

\[ y = \text{slope}^2 + \lambda \text{the slope}^2 \]

\[ y = \text{slope} \times x + \text{y-axis intercept} \]
If you want to know more about the Cross Validation, I recommend:


Ridge Regression can solve complication models as well.

\[ y = y\text{-axis intercept} + \text{slope}_1 x_1 + \text{slope}_2 x_2 + \text{slope}_3 x_3 + \ldots + \text{slope}_n x_n \]

The Ridge Regression Penalty =

\[ \lambda \times (\text{slope}_1^2 + \text{slope}_2^2 + \text{slope}_3^2 + \ldots + \text{slope}_n^2) \]
Ridge Regression can also be applied to Logistic Regression.

The cost function of Logistic Regression:

$$J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))$$

Logistic Regression is solved using Maximum Likelihood.

So, regularization using Ridge Regression optimizes the sum of the likelihoods instead of the squares of the residuals

$$+ \lambda \text{the slope}^2$$
If you want to know more about the Linear Regression and Logistic Regression, I recommend:

1. Regression Methods, [https://newonlinecourses.science.psu.edu/stat501/node/250/](https://newonlinecourses.science.psu.edu/stat501/node/250/)

2. Linear Regression With R - R-statistics.co, [http://r-statistics.co/Linear-Regression.html](http://r-statistics.co/Linear-Regression.html)


4. Lecture 2.1 — Linear Regression With One Variable | Model Representation — Andrew Ng, [https://www.youtube.com/watch?v=kHwlB_j7Hkc&index=4&list=PLLssT5z_DsK-h9vYZkQkYNWcItpqhlRJLN](https://www.youtube.com/watch?v=kHwlB_j7Hkc&index=4&list=PLLssT5z_DsK-h9vYZkQkYNWcItpqhlRJLN)
More about Ridge Regression

\[ y = y\text{-axis intercept} + \text{slope}_1 x_1 + \text{slope}_2 x_2 + \text{slope}_3 x_3 + \ldots + \text{slope}_n x_n \]

The Ridge Regression Penalty =

\[ \lambda (\text{slope}_1^2 + \text{slope}_2^2 + \text{slope}_3^2 + \ldots + \text{slope}_n^2) \]

For least Squares, we need at least \( n \) points to determine what the equation is.

When \( n \) becomes bigger and bigger, we need more and more points.

Ridge Regression can find a solution with Cross Validation and Ridge Regression Penalty.
Lasso Regression
In regularization, we usually use two algorithms, **Ridge Regression** and **Lasso Regression**.

Regularization is achieved by adding different constraints to the parameter after the cost function.
In **Ridge Regression**, we minimized the sum of the squares of the residuals

\[ + \lambda \text{the } \text{slope}^2 \]  

Ridge Regression Penalty

In **Lasso Regression**, we minimized the sum of the squares of the residuals

\[ + \lambda \text{the } \vert \text{slope} \vert \]  

Lasso Regression Penalty
Lasso Regression can solve complication models as well.

\[ y = y\text{-axis intercept} + \text{slope}_1 x_1 + \text{slope}_2 x_2 + \text{slope}_3 x_3 + \ldots + \text{slope}_n x_n \]

The Lasso Regression Penalty =

\[ \lambda \left( |\text{slope}_1| + |\text{slope}_2| + |\text{slope}_3| + \ldots + |\text{slope}_n| \right) \]
Lasso Regression can also be applied to Logistic Regression.

The cost function of Logistic Regression:

\[
J(\theta) = -\frac{1}{m} \sum_{i=1}^{m} y^{(i)} \log h_\theta(x^{(i)}) + (1 - y^{(i)}) \log (1 - h_\theta(x^{(i)}))
\]

Logistic Regression is solved using Maximum Likelihood.

So, regularization using Lasso Regression optimizes the sum of the likelihoods instead of the squares of the residuals:

\[+ \lambda^* | \text{the slope} |\]
In Lasso Regression, we can increase the $\lambda$. As $\lambda$ increases, the slope of the results line gets smaller until the slope =0.

The difference between Ridge and Lasso Regression is that Ridge Regression can only make the slope asymptotically close to 0 while Lasso Regression can make the slope =0.

$$y = y\text{-axis intercept} + \text{slope}_1 \times x_1 + \text{slope}_2 \times x_2 + \text{slope}_3 \times x_3 + \ldots + \text{slope}_n \times x_n$$
Ridge Regression can do better in the data set when most variables are useful.

Lasso Regression can do better when the data set contains lots of useless variables.
Elastic-Net Regression

the sum of the squares of the residuals

\[ + \lambda_1 (|\text{slope}_1| + |\text{slope}_2| + |\text{slope}_3| + \ldots + |\text{slope}_n|) \]

\[ + \lambda_2 (\text{slope}_1^2 + \text{slope}_2^2 + \text{slope}_3^2 + \ldots + \text{slope}_n^2) \]
R code for Ridge Regression and Lasso Regression

Let’s coding!
To do Ridge, Lasso and Elastic-Net Regression in R, we will use the glmnet library.

the sum of the squares of the residuals

\[ + \lambda \times [\alpha \times (|\text{slope}_1| + |\text{slope}_2| + |\text{slope}_3| + \ldots + |\text{slope}_n|) \\
+ (1-\alpha) \times (\text{slope}_1^2 + \text{slope}_2^2 + \text{slope}_3^2 + \ldots + \text{slope}_n^2)] \]
If you want to know more about Regularization, I recommend:

- Regularization Part 1: Ridge Regression, [https://www.youtube.com/watch?v=Q81RR3yKn30&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=37](https://www.youtube.com/watch?v=Q81RR3yKn30&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=37) ----- StatQuest

- Lecture 7.1 — Regularization | The Problem Of Overfitting — [Machine Learning | Andrew Ng], [https://www.youtube.com/watch?v=u73PU6QwlI](https://www.youtube.com/watch?v=u73PU6QwlI) ----- Andrew Ng
Reference

• Regularization Part 1: Ridge Regression, https://www.youtube.com/watch?v=Q81RR3yKn30&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=37

• Regularization Part 2: Lasso Regression, https://www.youtube.com/watch?v=NGf0voTMIcs&index=9&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF

• Regularization Part 3: Elastic Net Regression, https://www.youtube.com/watch?v=1dKRdX9bfIo&index=10&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF

• Ridge, Lasso and Elastic-Net Regression in R, https://www.youtube.com/watch?v=ctmNq7FgbvI&list=PLblh5JKOoLUICTaGLRoHQDuF_7q2GfuJF&index=11
Reference

- Lecture 7.1 — Regularization | The Problem Of Overfitting — [Machine Learning | Andrew Ng], https://www.youtube.com/watch?v=u73PU6Qwl1I

- Lecture 7.2 — Regularization | Cost Function — [Machine Learning | Andrew Ng | Stanford University], https://www.youtube.com/watch?v=KvtGD37Rm5I&index=41&list=PLLssT5z_DsK-h9vYZkQkYNWcItqhlRJLN&t=0s

- Lecture 7.3 — Regularization | Regularized Linear Regression — [Machine Learning | Andrew Ng], https://www.youtube.com/watch?v=qbvRdrd0yJ8&list=PLLssT5z_DsK-h9vYZkQkYNWcItqhlRJLN&index=41

- Lecture 7.4 — Regularization | Regularized Logistic Regression — [Machine Learning | Andrew Ng], https://www.youtube.com/watch?v=IXPgm1e0IOo&index=42&list=PLLssT5z_DsK-h9vYZkQkYNWcItqhlRJLN
Reference

• Regularization (mathematics), https://en.wikipedia.org/wiki/Regularization_(mathematics)
• Linear regression, https://en.wikipedia.org/wiki/Linear_regression
• Least squares, https://en.wikipedia.org/wiki/Least_squares
• Logistic regression, https://en.wikipedia.org/wiki/Logistic_regression
Thank you very much!