Machine Learning to build Intelligent Systems

Manas Dasgupta

Model Regularization

Structure of this Module

Understanding Regularization

TOPICS
Ridge Regression
Lasso Regression
ElasticNet

What is Regularization

One of the major aspects of training your machine learning model is avoiding overfitting.

*The model will have a **low test accuracy** if it is overfitting.* This happens because your model is trying too hard to capture the noise in your training dataset.

Noise is the data points that don't really represent the true properties of your data, but random chance. Learning such data points, makes your model more flexible during training, at the risk of overfitting.

Revisit Bias and Variance.

This is a form of regression, that constrains/ regularizes or shrinks the coefficient estimates. In other words, ***this technique discourages the model from learning a more complex or flexible model, so as to avoid the risk of overfitting.***

Ridge Regression

Linear Equation model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\ldots + \beta_n x_n$

Cost Function of a Linear Equation model (MSE): $\sum_{i=1}^{n} (y - \hat{y})^2 / n$

- ***Ridge Regression*** (also called Tikhonov regularization) is a regularized version of Linear Regression: a regularization term equivalent to the Squares of the weight vector (*L2* norm) is added to the cost function.
- This forces the learning algorithm to not only fit the data but also ***keep the model weights as small as possible***. Note that the regularization term should only be ***added to the cost function during training***.
- Once the model is trained, you want to evaluate the model's performance using the unregularized performance measure.

$$\alpha \sum_{i=1}^{n} \theta_i^2$$

[θ_i is are the weights or coefficients the model is set out to determine]

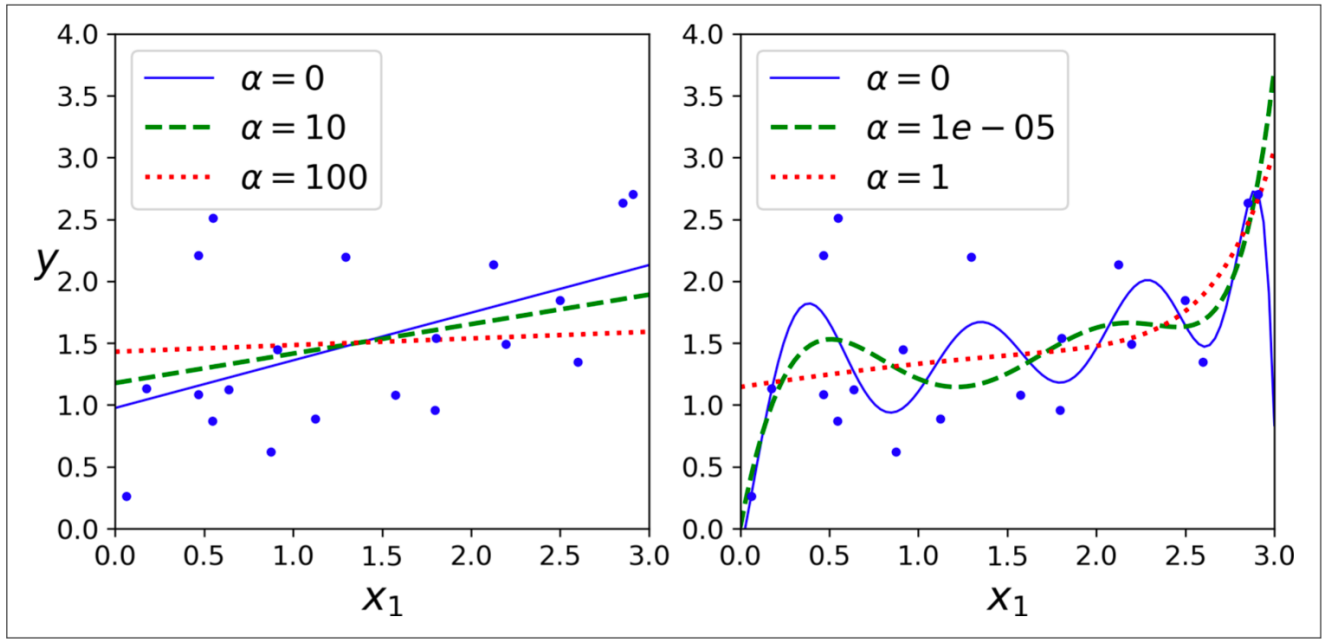
It is important to scale the data (e.g., using a StandardScaler) before performing Ridge Regression, as it is sensitive to the scale of the input features.
This is true of most regularized models.

Ridge Regression

- The hyperparameter α controls how much you want to regularize the model.
- If $\alpha = 0$ then Ridge Regression is just Linear Regression.
- If α is very large, then all weights end up very close to zero and the result is a flat line going through the data's mean.

Ridge Regression Cost Function:

$$J(\boldsymbol{\theta}) = \mathrm{MSE}(\boldsymbol{\theta}) + \alpha \frac{1}{2} \Sigma_{i=1}^{n} \theta_i^2$$



Lasso Regression

Linear Equation model: $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \ldots\ldots + \beta_n x_n$

Cost Function of a Linear Equation model (MSE): $\sum_{i=1}^{n} (y - \hat{y})^2 / n$

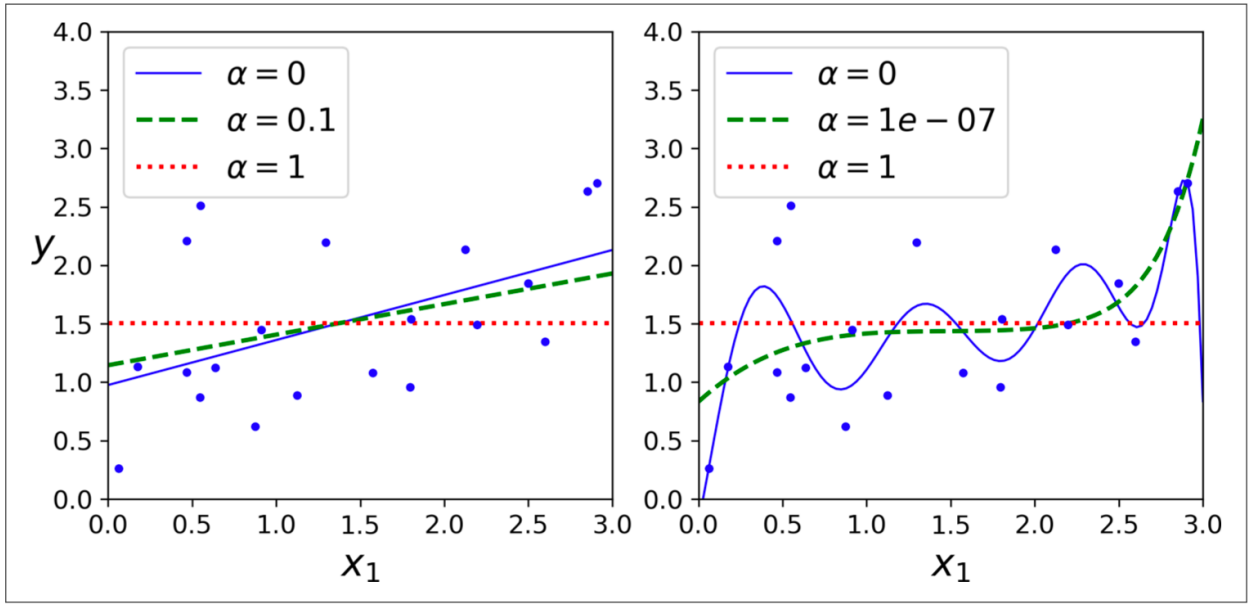
Least Absolute Shrinkage and Selection Operator Regression (simply called *Lasso Regression*) is another regularized version of Linear Regression: just like Ridge Regression, it adds a regularization term to the cost function.

Lasso differs from ridge regression in penalizing only the high coefficients. It uses the **modulus of the weight vector instead of squares**, as its penalty. In statistics, this is known as the L1 norm.

Lasso Regression

A difference in Lasso Regression from Ridge Regression is that ***it tends to completely eliminate the weights of the least important features*** (i.e., set them to zero).

For example, the dashed line in the right plot on the figure below (with $\alpha = 10^{-7}$) looks quadratic, almost linear: all the weights for the high-degree polynomial features are equal to zero. In other words, Lasso Regression automatically performs feature selection and outputs a *sparse model* (i.e., with few nonzero feature weights).



Lasso Regression Cost Function:

$$J(\boldsymbol{\theta}) = \mathrm{MSE}(\boldsymbol{\theta}) + \alpha \Sigma_{i=1}^{n} |\theta_i|$$

[θ_i is are the weights or coefficients the model is set out to determine]

Comparison between Ridge and Lasso

There is a disadvantage of ridge regression, which is model interpretability. It will shrink the coefficients for least important predictors, very close to zero. But it will never make them exactly zero. In other words, the final model will include all predictors.

However, in the case of the lasso, the L1 penalty has the effect of forcing some of the coefficient estimates to be exactly equal to zero when the tuning parameter λ is sufficiently large. **Therefore, the lasso method also performs variable selection and is said to yield sparse models.**

A standard least squares model tends to have some variance in it, i.e., this model won't generalize well for a data set different than its training data. ***Regularization, significantly reduces the variance of the model, without substantial increase in its bias***. So, the tuning parameter ∝, controls the impact on bias and variance.

As the value of ∝ is increased, it reduces the value of coefficients and thus reducing the variance. ***Till a point, this increase in ∝ is beneficial in reducing overfitting and without losing any important properties in the data.***

However, after certain value of ∝, the model starts losing important information inherent in the training data, increasing bias leading to underfitting. Therefore, the value of ∝ should be carefully selected.

ElasticNet

Elastic Net is a middle ground between Ridge Regression and Lasso Regression.

The regularization term is a simple mix of both Ridge and Lasso's regularization terms, and you can control the mix ratio *r*.

- When *r* = 0, Elastic Net is equivalent to Ridge Regression
- When *r* = 1, it is equivalent to Lasso Regression

ElasticNet Cost Function:

$$J(\boldsymbol{\theta}) = \mathrm{MSE}(\boldsymbol{\theta}) + r\alpha\Sigma_{i=1}^{n}\left|\theta_i\right| + \frac{1-r}{2}\alpha\Sigma_{i=1}^{n}\theta_i^{2}$$

When should you use plain Linear Regression (i.e., without any regularization), Ridge, Lasso, or Elastic Net?

- It is almost always preferable to have at least a little bit of regularization, so generally you should avoid plain Linear Regression.
- Ridge is a good default, but if you suspect that only a few features are actually useful, you should prefer Lasso or Elastic Net since they tend to reduce the useless features' weights down to zero as we have discussed.
- In general, Elastic Net is preferred over Lasso since Lasso may behave erratically when the number of features is greater than the number of training instances or when several features are strongly correlated.

Linear Equation

Python Demo

- **Ridge Regression**
- **Lasso Regression**
- **ElasticNet**

Hope you have liked this Video.
Please help us by providing your Ratings and Comments for this Course!

Thank You!!
Manas Dasgupta

Happy Learning!!