Regularization in Machine Learning

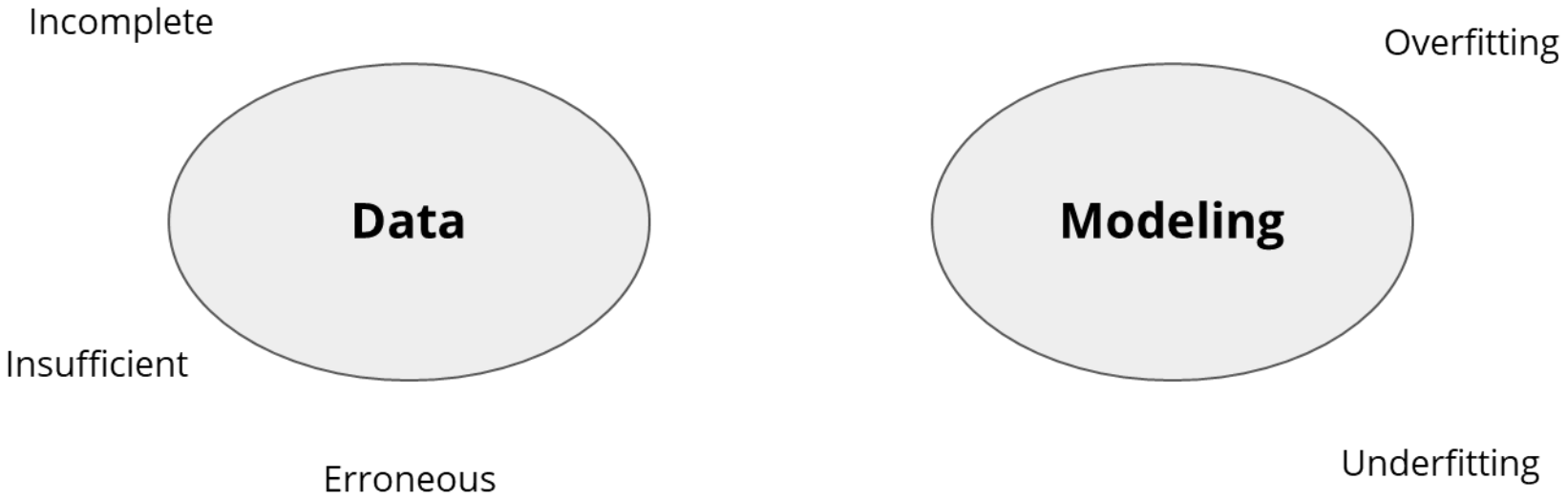
Learning Objectives

- Explain regularization techniques
- Discuss loss functions
- Describe L1 and L2 norms
- Explain Ridge regression
- Describe Lasso regression
- Discuss Elastic Net regression

Regularization Techniques

Challenges of Statistical Models

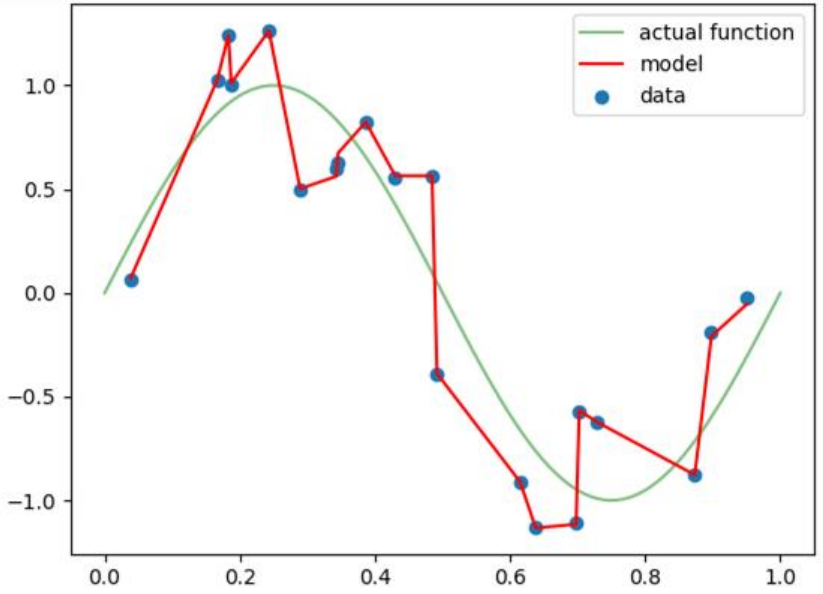
In applied fields like data science and many other disciplines, the goal isn't to create a model that looks good on paper but to create a model that's useful in a practical context.



Overfitting Example

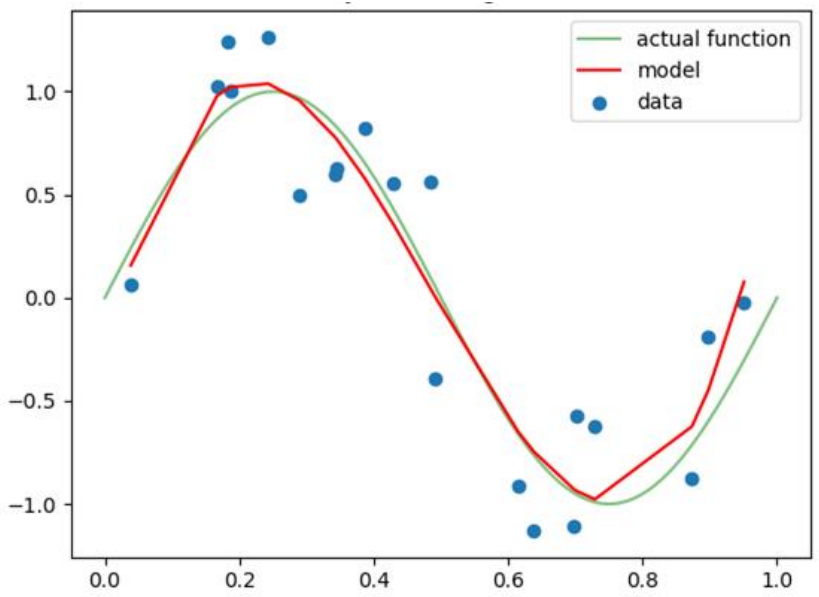
knowledgehut
upGrad

- Goal: Prevent overfitting
- Intuition: Reduce the model accuracy on the training data in favor of better overall performance on new data.



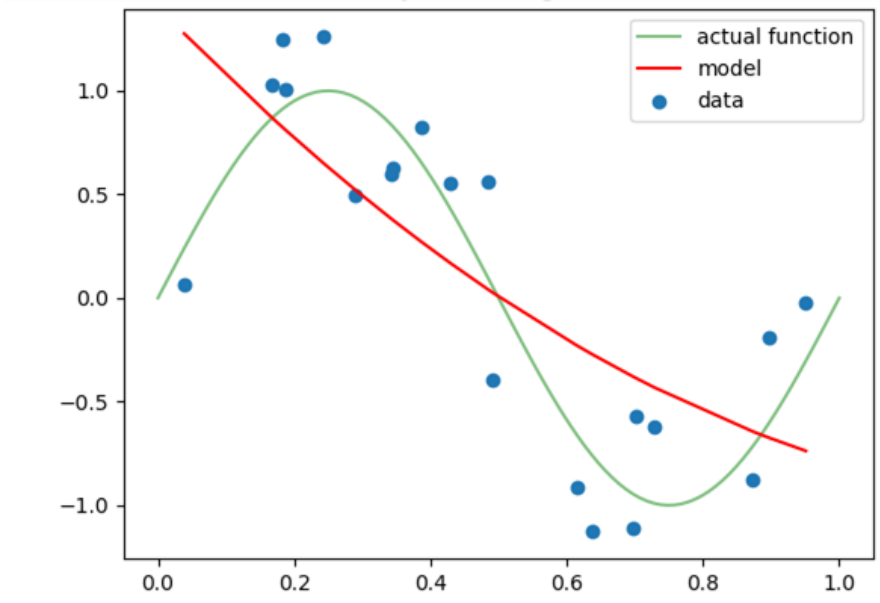
Right Fit Example

Technically, some biases are added to the model to avoid overfitting (high variance).



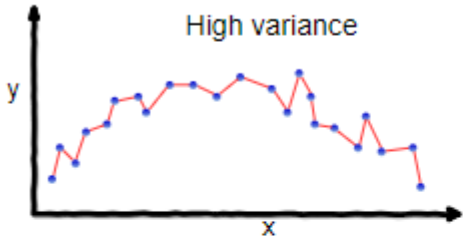
Underfitting Example

There is too much variance between the actual data and the predicted data.

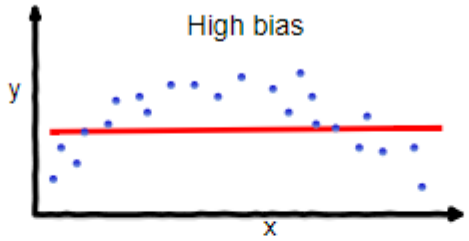


Bias-variance Tradeoff

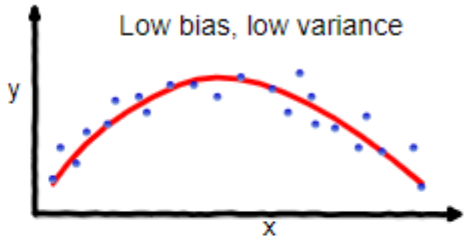
Balancing this process is called the bias-variance.



overfitting



underfitting



Good balance

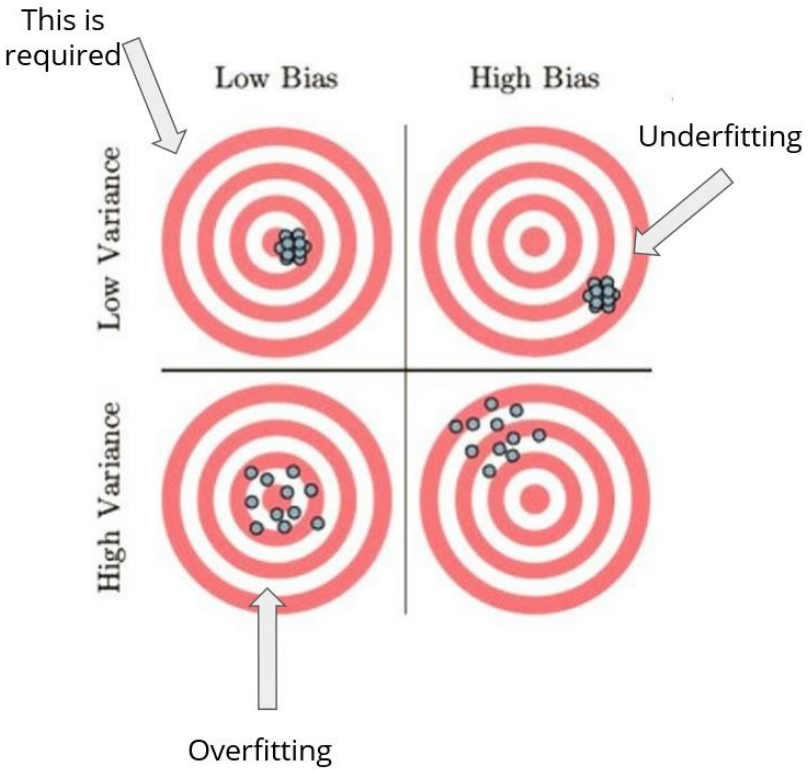
https://towardsdatascience.com/understanding-the-bias-variance-tradeoff-165e6942b229

Bias-variance Tradeoff

Balancing this process is called the bias-variance.

To get it right, it's mandatory to have the following:

- Accurate data that allows enough variation for new data points.
- Multiple iterations using fresh data.



Regularization Techniques

Techniques that can be used to prevent overfitting and help reduce the amount of noise in the data.

Penalizing model/Shrinkage methods

Early stopping

Feature selection

Data augmentation

Dimensionality reduction

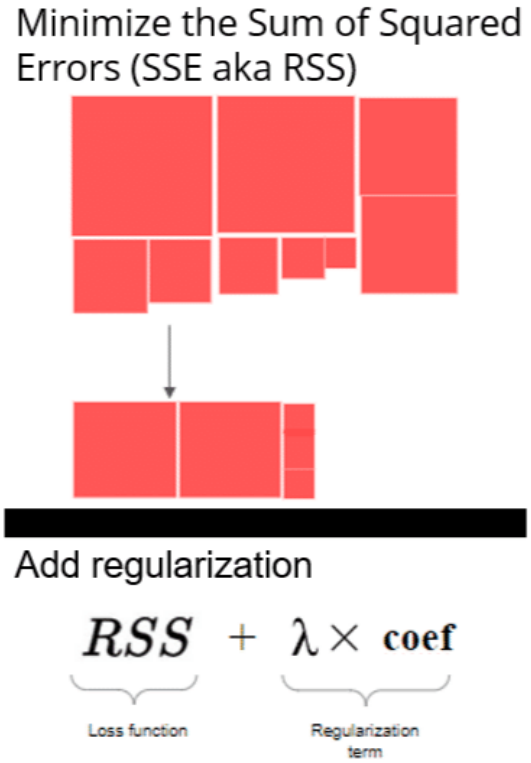
Dropout

Loss Functions

Loss Function in OLS

The Loss Function for the OLS regression is to minimize the sum of squared errors or the residual sum of squares.

To add regularization, take the RSS and add a regularization term.



Regularization Techniques

- There are two techniques – L1 regularization and L2 norm regularization.
- The regularization term is typically constructed by defining a hyperparameter called **Lambda**, multiplied by another term.

$$\underbrace{RSS}_{\text{Loss function}} + \underbrace{\lambda \times \sum_{j=1}^{p} |b_j|}_{\text{L1-Regularization / Linear}}$$

$$\underbrace{RSS}_{\text{Loss function}} + \underbrace{\lambda \times \left(\sum_{j=1}^{p} b_j \right)^2}_{\text{L2-Regularization / Quadratic}}$$

Hyperparameter Lambda

knowledgehut upGrad

- Lambda is a hyperparameter that can be greater or equal to 0.
- Regularization reduces the variance of the estimates.
- Larger values specify stronger regularization.
- Do not regularize the model when it is still underfitting in training, as it still has large errors.
- Lambda is also called alpha when directly linked to the specific regression function.
- These regularization methods produce multiple sets of coefficient estimates for each value of Lambda.

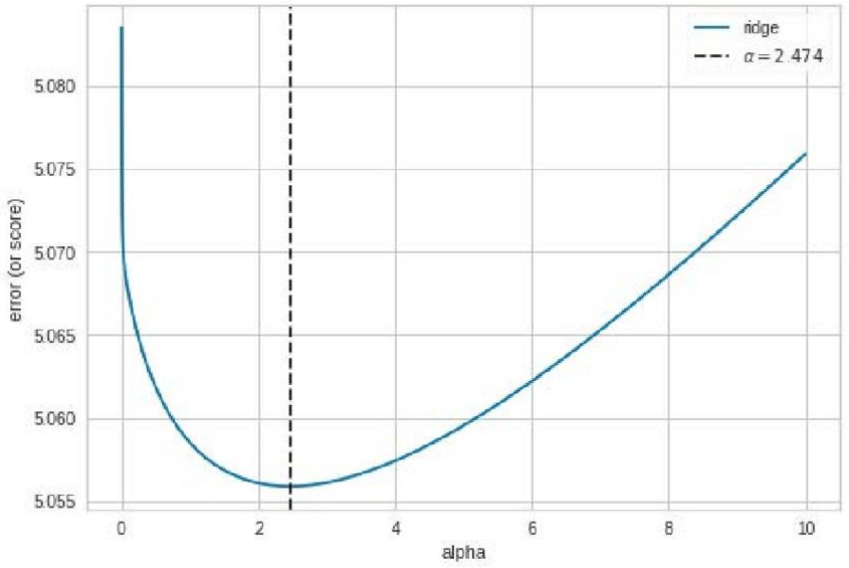
Hyperparameter Lambda

•Split the dataset into three subsets: a training set, a validation set, and a test set.

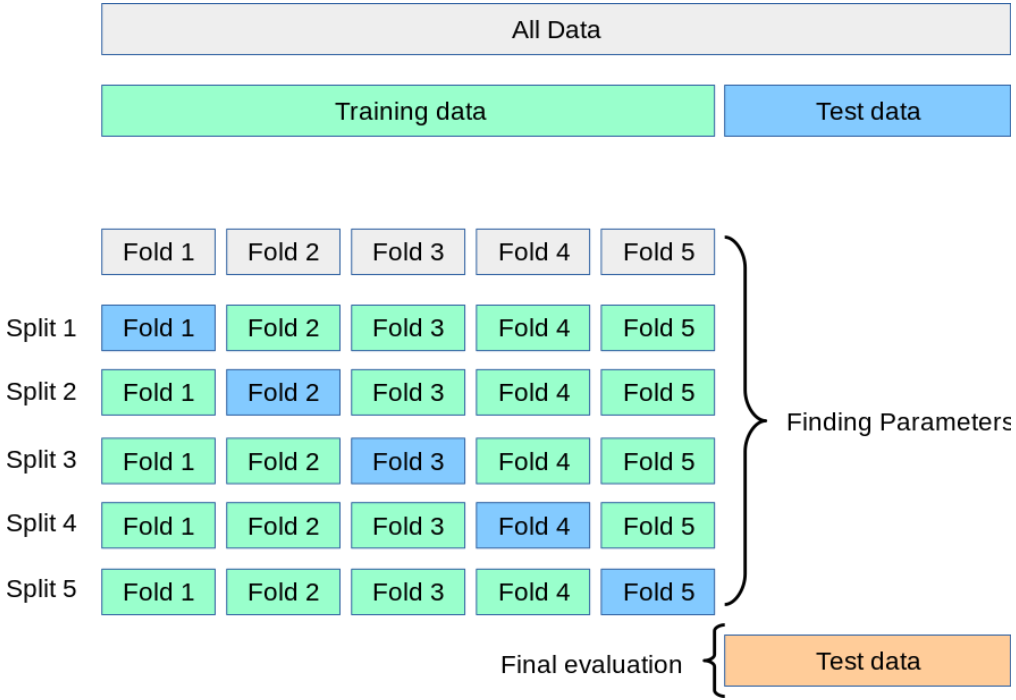
•The training set is used to train the model.

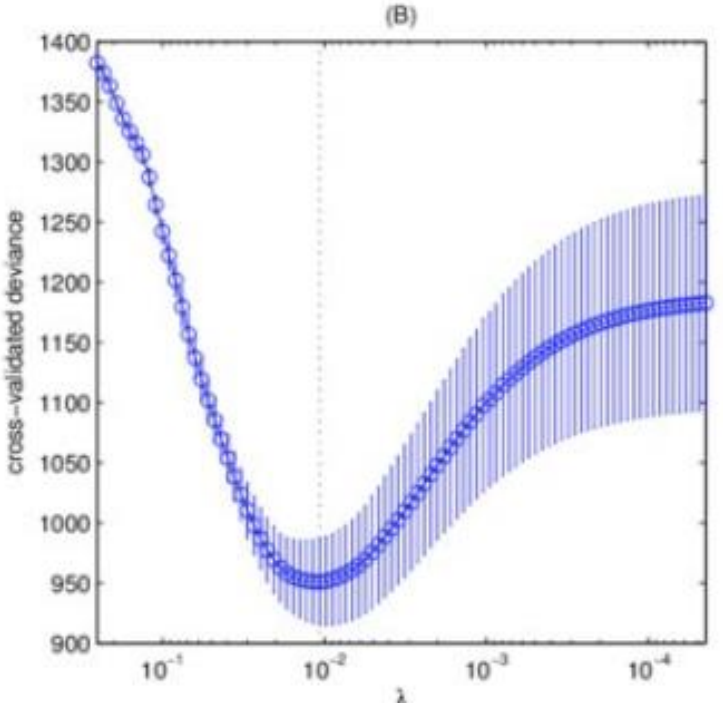
•The validation set is used to tune the hyperparameters and evaluate the model's performance during training.

•The test set is used to evaluate the final performance of the tuned model.



Hyperparameter Lambda





https://www.researchgate.net/publication/307181887_penalized_A_MATLAB_Toolbox_for_Fitting_Generalized_Linear_Models_with_Penalties/figures?lo=1

L1 and L2 Norms

Quadratic Regularization (L2 Norm)

- For the L2 norm, multiply Lambda by the squared values of the model's coefficients (or weights).
- A large coefficient means a one-unit increase in the variable greatly affects the outcome variable. This means that the slope is steep.
- On the other hand, if the coefficient is close to 0, the variable has a smaller effect on the outcome variable.
- The application of L2 regularization in regression is called Ridge regression.

$$\underbrace{RSS}_{\text{Loss function}} + \underbrace{\lambda \times \left(\sum_{j=1}^{p} b_j\right)^2}_{\textbf{L2-Regularization / Quadratic}}$$

Linear Regularization (L1 Norm)

- The L1 norm works quite similar to the L2 norm. The only difference is you do not add the squared value of the coefficient but the absolute value.
- Coefficients can become 0 and thus can be removed from the model.
- The application of L1 regularization in regression is called Lasso regression.

$$\underbrace{RSS}_{\text{Loss function}} + \underbrace{\lambda \times \sum_{j=1}^{p} |b_j|}_{\textbf{L1-Regularization / Linear}}$$

Ridge Regression

Ridge Regression

Properties

- Regularizes coefficients to enhance prediction and avoid overfitting.
- Shrinks all coefficients by a uniform factor to become as small as possible (close to 0).

Achieves balance between:

- Each predictor should contribute to the result as least as possible (less complex model).
- Good overall prediction.

It works best when all input variables have a similar range, and you want to keep all coefficients. **Remember standardization / scaling ?**

Minimize:

$$RSS + \lambda \times \left(\sum_{j=1}^{p} b_j\right)^2$$

Pros and Cons of Ridge Regression

knowledgehut upGrad

Pros:

- Reduces variance.
- Helps the model to learn more complex patterns without overfitting easily.
- Reduces the influence of several correlated variables.

Cons:

- Can't perform feature selection (coefficients don't become 0).
- Difficult to estimate λ. (but possible)

Minimize:

$$RSS + \lambda \times \left(\sum_{j=1}^{p} b_j \right)^2$$

Lasso Regression

Lasso Regression

Properties

- Regularizes coefficients to enhance prediction and avoid overfitting.
- Shrinks some of them to exactly 0.

Achieves balance between:

- Reduces the number of predictors to important ones (less complex model).
- Good overall prediction.

It works best when one wants to drop unimportant features from the model and retain only some important features. **(Feature Selection)**

Minimize:

$$RSS + \lambda \times \sum_{j=1}^{p} |b_j|$$

Pros and Cons of Lasso Regression

knowledgehut upGrad

Pro:

- Reduces variance.
- Avoids overfitting.
- Can perform variable selection.
- Eliminates several correlated variables.
- More robust to outliers.

Minimize:

$$RSS + \lambda \times \sum_{j=1}^{p} |b_j|$$

Cons:

- Not useful if the number of predictors is low and almost all are important.
- Difficult to learn complex patterns from the input data.
- Difficult to estimate λ. (but possible)

Elastic Net Regression

Elastic Net Regression

knowledgehut upGrad

Properties

- Combines L1 and L2 regularizations into one formula.
- Lambda gets multiplied with both the absolute and the squared coefficients.
- Another parameter lets you adjust the weight of each regularization method ,“L1_weight”.

Minimize:

$$\frac{\sum_{i=1}^{n}(y_i - x_i^J \hat{\beta})^2}{2n} + \lambda \left(\underbrace{\frac{1-\alpha}{2} \sum_{j=1}^{m} \hat{\beta}_j^2}_{\text{L2-Regularization}} + \underbrace{\alpha \sum_{j=1}^{m} |\hat{\beta}_j|}_{\text{L1-Regularization}} \right)$$

Pros and Cons of Elastic Net Regression

Pros:

- Allows balancing both penalties, potentially resulting in better model performance.

Cons:

- Another hyperparameter that is difficult to estimate.

Minimize:

$$\frac{\sum_{i=1}^{n}(y_i - x_i^J \hat{\beta})^2}{2n} + \lambda \left(\underbrace{\frac{1-\alpha}{2} \sum_{j=1}^{m} \hat{\beta}_j^2}_{\text{L2-Regularization}} + \underbrace{\alpha \sum_{j=1}^{m} |\hat{\beta}_j|}_{\text{L1-Regularization}} \right)$$

Thank you

Copyright © Knowledgehut., All rights reserved
Confidential Information, intended for approved distribution list of Knowledgehut