SNS COLLEGE OF TECHNOLOGY
(AN AUTONOMOUS INSTITUTION)
COIMBATORE – 35
DEPARTMENT OF COMPUTER SIENCE AND ENGINEERING

UNIT II SUPERVISED LEARNING

Introduction - Linear Models for Regression – Linear Regression Models and Least Squares
Subset Selection
Shrinkage Methods – Derived Input Directions
Linear Models for Classification- Discriminant Analysis
Logistic Regression
Separating Hyperplanes

Shrinkage Methods in Machine Learning

The best known shrinking methods are **Ridge Regression and Lasso Regression** which are often used in place of Linear Regression.

Why shrink or subset and what does this mean?

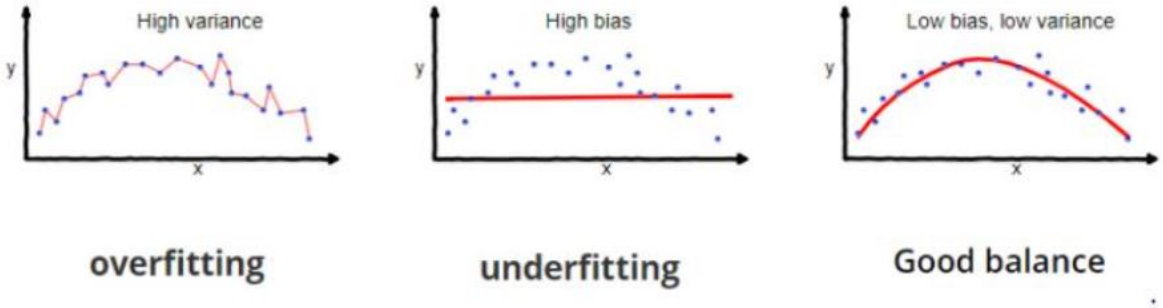
In the **linear regression** context, subsetting means choosing a subset from available variables to include in the model, thus **reducing its dimensionality.**

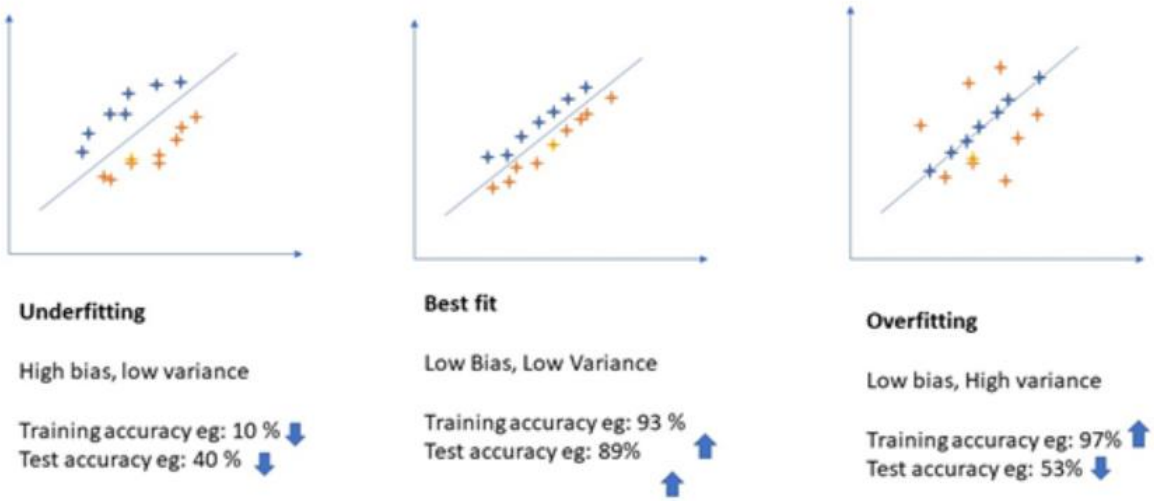
Shrinkage, on the other hand, means **reducing the size of the coefficient estimates** (shrinking them towards zero). Note that if a coefficient gets shrunk to exactly zero, the corresponding variable drops out of the model. Consequently, such a case can also be seen as a kind of subsetting.

Shrinkage and selection aim at improving upon the simple linear regression.

There are two main reasons why it could need improvement:

- **Prediction accuracy**: Linear regression estimates tend to have low bias and high variance. Reducing model complexity (the number of parameters that need to be estimated) results in reducing the variance at the cost of introducing more bias. If we could find the sweet spot where the total error, so the error resulting from bias plus the one from variance, is minimized, we can improve the model's predictions.
- **Model's interpretability**: With too many predictors it is hard for a human to grasp all the relations between the variables. In some cases we would be willing to determine a small subset of variables with the strongest impact, thus sacrificing some details in order to get the big picture.





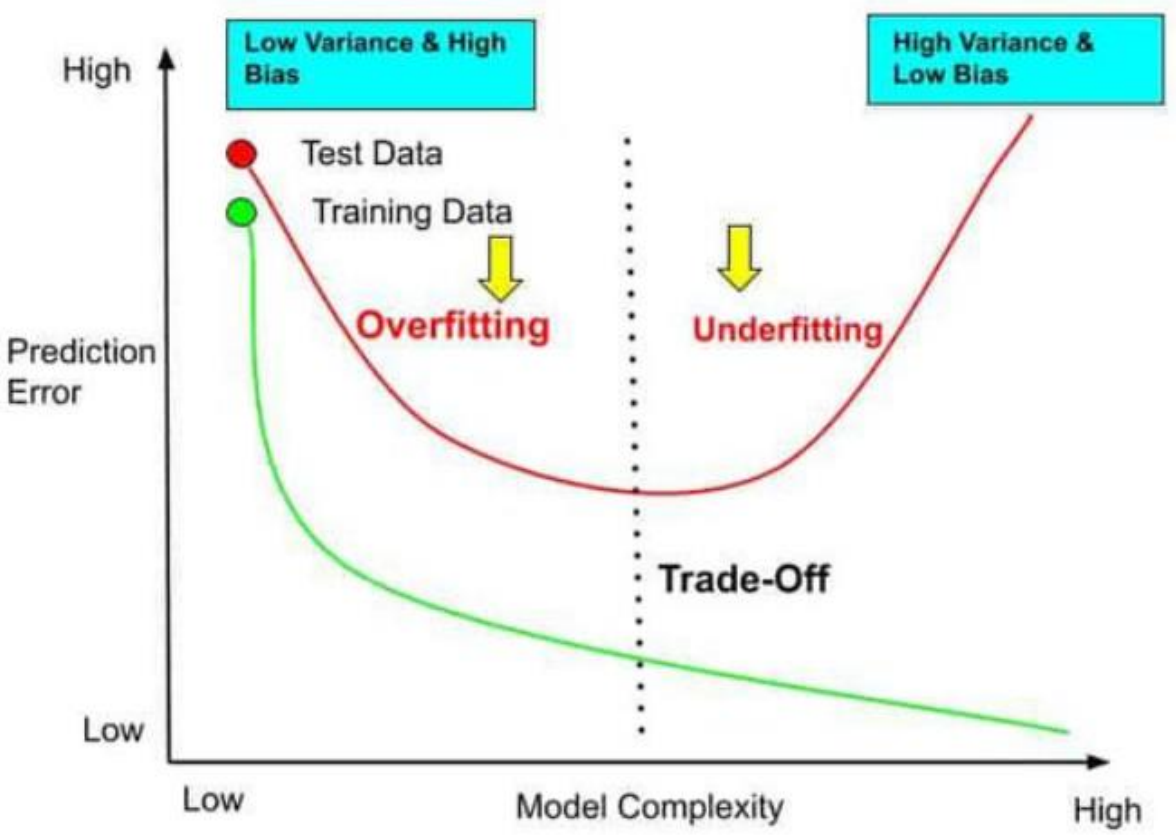
Bias — error of train data
Variance — error of test data

Underfitting:
Training and test errors both are high. Both models gives less score . But test score is more than train score (High bias, lower variance)

Overfitting:
Training error is less. Test error is more. Model gives good score to training data, and less score to test data. (Low bias, high variance)

Best fit:
Both train and test gives less errors. Model gives best score to both train and test data. (Low bias and low variance)

Bias-Variance Trade off



Bias Variance Trade off

shrinkage methods are also known as regularization

Regularization in Machine Learning

What is Regularization?

Regularization is one of the most important concepts of machine learning. It is a technique to **prevent the model** from **overfitting by adding extra information to it.**

Sometimes the machine learning **model performs well with the training data but does not perform well with the test data**. It means the model is not able to predict the output when deals with unseen data by introducing noise in the output, and hence the **model is called overfitted.**

This problem can be deal with the help of a **regularization technique.**

This technique can be used in such a way that it will allow to maintain all variables or features in the model by reducing the magnitude of the variables. Hence, it maintains accuracy as well as a generalization of the model.

It mainly regularizes or reduces the coefficient of features toward zero. In simple words, "*In regularization technique, we* ***reduce the magnitude of the features*** *by keeping the same number of features."*

How does Regularization Work?

Regularization works by adding a penalty or complexity term to the complex model. Let's consider the simple linear regression equation:

$$y= \beta0+\beta1x1+\beta2x2+\beta3x3+\cdots+\beta nxn +b$$

In the above equation,

Y represents the value to be **predicted**

X1, X2, …Xn are the **features** for Y.

β0,β1,…..βn are the **weights or magnitude** attached to the features, respectively. Here represents the bias of the model, and **b represents the intercept**.

Linear regression models try to **optimize the β0 and b to minimize the cost function.** The equation for the cost function for the linear model is given below:

$$\sum_{i=1}^{M}\left(y_i - y'_i\right)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{n}\beta_j * Xij\right)^2$$

Now, we will add a **loss function and optimize parameter** to make the model that can predict the accurate value of Y. The loss function for the linear regression is called as **RSS or Residual sum of squares.**

Techniques of Regularization

There are mainly two types of regularization techniques, which are given below:

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

Ridge Regression

- Ridge regression is one of the types of linear regression in which **a small amount of bias is introduced** so that we can get **better long-term predictions**.

- Ridge regression is a regularization technique, which is used to **reduce the complexity** of the model. It is also called as **L2 regularization**.
- In this technique, the cost function is altered by **adding the penalty term** to it. The amount of bias added to the model is called **Ridge Regression penalty**. We can calculate it by multiplying with the lambda to the squared weight of each individual feature.
- The equation for the cost function in **ridge regression** will be:

$$\sum_{i=1}^{M}(y_i - y'_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{n}\beta_j * x_{ij}\right)^2 + \lambda\sum_{j=0}^{n}\beta_j{}^2$$

- In the above equation, the penalty term regularizes the coefficients of the model, and hence ridge regression reduces the amplitudes of the coefficients that decreases the complexity of the model.
- As we can see from the above equation, if the values of **λ tend to zero, the equation becomes the cost function of the linear regression model.** Hence, for the minimum value of λ, the model will resemble the linear regression model.
- A general linear or polynomial regression will fail if there is high collinearity between the independent variables, so to solve such problems, Ridge regression can be used.
- It helps to solve the problems if we have more parameters than samples.

Lasso Regression:

- Lasso regression is another regularization technique to **reduce the complexity** of the model. It stands for **Least Absolute and Selection Operator.**
- It is similar to the Ridge Regression except that the penalty term contains only the **absolute weights** instead of a square of weights.
- Since it takes absolute values, hence, it can shrink the slope to 0, whereas Ridge Regression can only shrink it near to 0.
- It is also called as **L1 regularization.**
- The equation for the cost function of **Lasso regression** will be:

$$\sum_{i=1}^{M}(y_i - y'_i)^2 = \sum_{i=1}^{M}\left(y_i - \sum_{j=0}^{n}\beta_j * x_{ij}\right)^2 + \lambda\sum_{j=0}^{n}|\beta_j|$$

- Some of the features in this technique are completely neglected for model evaluation.
- Hence, the Lasso regression can help us to **reduce the overfitting** in the model as well as the feature selection.

Key Difference between Ridge Regression and Lasso Regression

- **Ridge regression** is mostly used to **reduce the overfitting** in the model, and it **includes all the features present in the model**. It reduces the complexity of the model by **shrinking the coefficients**.

- **Lasso regression** helps to **reduce the overfitting** in the model as well as **feature selection.**

Derived Input Directions

- **Principal component regression**
- **Partial Least Squares**

This topic presents **regression methods based on dimension reduction techniques,** which can be very useful when you have a **large data set with multiple correlated predictor variables.**

Generally, all dimension reduction methods work by first summarizing the original predictors into few new variables called principal components (PCs), which are then used as predictors to fit the linear regression model. These methods avoid multicollinearity between predictors, which a big issue in regression setting

When using the dimension reduction methods, it's generally recommended to standardize each predictor to make them comparable. Standardization consists of dividing the predictor by its standard deviation.

Here, we described two well known **regression methods based on dimension reduction: Principal Component Regression** (**PCR**) and **Partial Least Squares** (**PLS**) regression.

Principal component regression

The Principal Component Regression (PCR) first applies **Principal Component Analysis** on the **data set to summarize the original predictor variables** into few new variables also known as principal components (PCs), which are a linear combination of the original data.

These PCs are then used to build the linear regression model. The number of principal components, to incorporate in the model, is chosen by cross-validation (cv). Note that, **PCR is suitable when the data set contains highly correlated predictors.**

Partial least squares regression

A possible drawback of PCR is that we have no guarantee that the selected principal components are associated with the outcome. Here, the selection of the principal components to incorporate in the model is not supervised by the outcome variable.

An alternative to PCR is the **Partial Least Squares** (PLS) regression, which identifies new principal components that not only summarizes the original predictors, but also that are **related to the outcome**. These components are then used to fit the regression model. So, compared to PCR, **PLS uses a dimension reduction strategy that is supervised by the outcome.**

Like PCR, PLS is convenient for data with highly-correlated predictors. The number of PCs used in PLS is generally chosen by cross-validation. Predictors and the outcome variables should be generally standardized, to make the variables comparable.

Reference Links:

http://www.sthda.com/english/articles/37-model-selection-essentials-in-r/152-principal-component-and-partial-least-squares-regression-essentials/