Machine Learning

Regularization

Linear Regression

- Linear regression attempts to find a relationship between a dependent variable and one or more explanatory (or independent) variables
- Linear regression can be used for various tasks
- For example,
  - a given dataset has data about locations of houses in a state/province, their prices, their architecture, their neighborhood, etc. This dataset can be used to estimate the prices for houses (which may not have been listed yet) in that particular state. This is useful for house owners, potential buyers, and real estate agencies.
  - Stock price prediction
  - Weather forecasting
  - Predictive analysis from survey data
  - Market research studies
  - Future sales prediction and much more

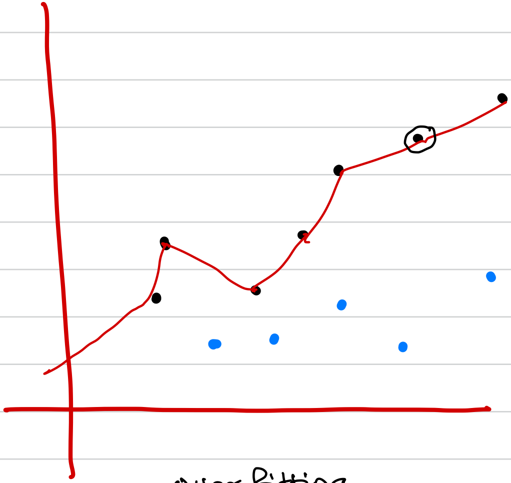
R-Squared

- R-squared is a statistical measure of how close the data are to the fitted regression line
- It evaluates the scatter of the data points around the fitted regression line
- It is also known as the coefficient of determination, or the coefficient of multiple determination for multiple regression
- It is the percentage of the response variable variation that is explained by a linear model
- R-squared = Explained variation / Total variation
- R-squared is always between 0 and 100%
  - 0% indicates that the model explains none of the variability of the response data around its mean
  - 100% indicates that the model explains all the variability of the response data around its mean
- In general, the higher the R-squared, the better the model fits your data

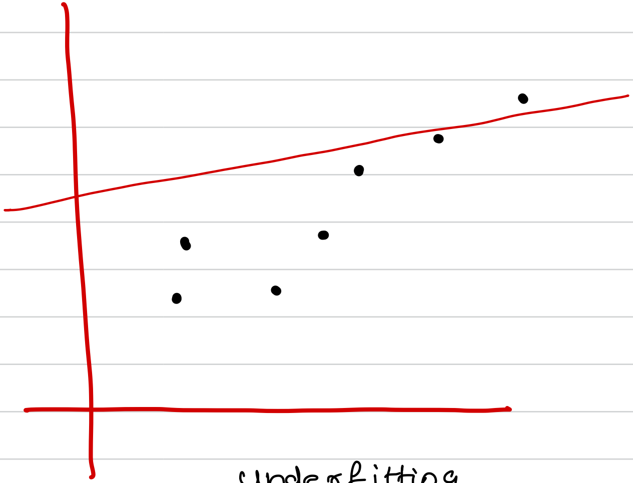
Adjusted R Squared

- R-Squared suffers from a major flaw: Its value never decreases no matter the number of variables we add to our regression model
- That is, even if we are adding redundant variables to the data, the value of R-squared does not decrease. It either remains the same or increases with the addition of new independent variables
- This clearly does not make sense because some of the independent variables might not be useful in determining the target variable
- Adjusted R-squared deals with this issue
- The Adjusted R-squared takes into account the number of independent variables used for predicting the target variable
- In doing so, we can determine whether adding new variables to the model actually increases the model fit

$$Adjusted\ R^2 = \{1 - [\frac{(1 - R^2)(n - 1)}{(n - k - 1)}]\}$$



over fitting



underfitting

accuracy with training dataset = 100%.

accuracy with testing dataset = poor

Regularization

- Regularization is a kind of regression that shrinks the coefficient estimates towards zero
- This technique discourages formation of a complex model, so as to avoid risk of overfitting

Underfitting and Overfitting

- Underfitting occurs when a model is not able to capture the underlying trend of the data
- Overfitting occurs when a model follows the trend of training data very closely but is not able to replicate the same performance on testing data
- A good fit model generalizes well and neither underfits nor overfits

Ridge Regression

- Ridge regression or **L2 regularization** brings values of coefficients near zero to enforce regularization
- Penalty is described by λ parameter
- The more the value of λ, the lesser the flexibility
- For low values of λ, the coefficients are very similar to that of a multiple linear regression model
- As λ increases, the differences between the results of Ridge model and linear regression model increase

Lasso Regression

- Lasso regression or **L1 regularization** not only brings values of coefficients near zero but to exact zero in case of weak regressors
- So, it not only shrinks coefficient estimates towards zero but also helps in feature selection
- Penalty is described by λ parameter.
- The more the value of λ, the lesser the flexibility
- For low values of λ, the coefficients are very similar to that of a multiple linear regression model
- As λ increases, the differences between the results of Lasso model and linear regression model increase

$$y = \xcancel{\beta_1 x_1} + \beta_2 x_2 + \beta_3 x_3$$

$$\beta_1 = 0$$