Regression

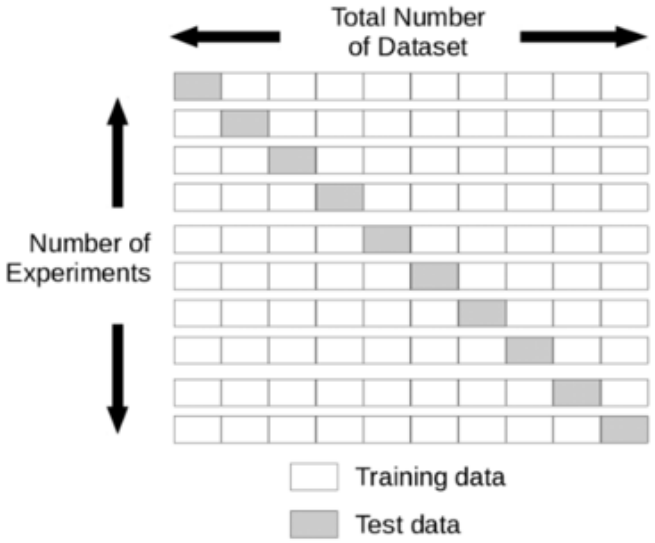
Day-3

K-Fold Cross Validation

- It is a sampling Technique used primarily for small data sets, where data is too small to partition into training and testing subsets.
- Main advantage is all of the data is used in building models, so all the patterns represented in the training data is used in building the models.
- K refers to how many subsets are used for the modeling

Procedure for K-fold cross validation

- Create k distinct data subsets through random sampling
- Assign a role to each subset. k-1 will be used for training, 1 for testing. Begin by assigning subset 1 to testing and subsets 2 through k for training.
- Rotate roles so each subset is used for testing one time and training k-1 times.



Source: https://www.researchgate.net/figure/Example-of-10-fold-Cross-validation_fig5_313851520

K-Fold Cross Validation

greatlearning

- A model is built from each of the folds – total of k models
- Any number of folds can be used.
- The hold-out subset is used to test the model, and the errors are averaged over the subsets.
- If the accuracy is similar for all folds, the modeling procedure is viewed as being stable and not overfit.

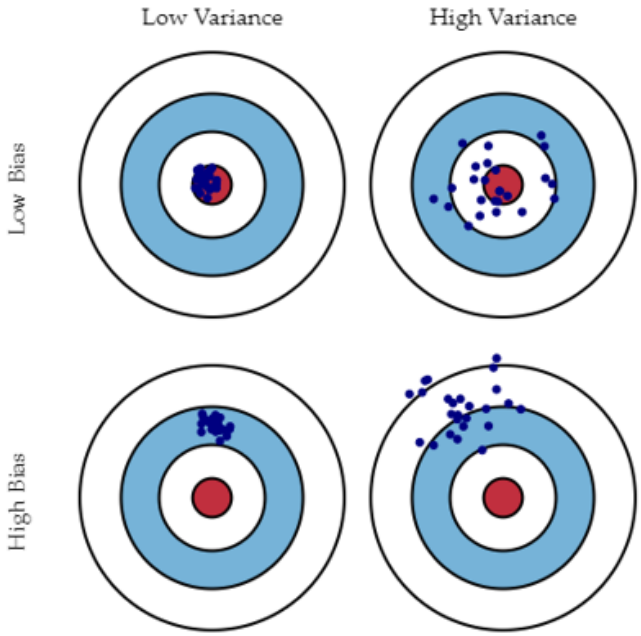
Bias-Variance Trade-off

Bias-Variance Trade-Off

- Prediction errors can be decomposed into two main subcomponents we care about: error due to "bias" and error due to "variance".
- There is a tradeoff between a model's ability to minimize bias and variance.
- **Bias refers to the model error** and **variance refers to the consistency in predictive accuracy of models applied to other data sets.**
- The best models have low bias (low error, high accuracy) and low variance (consistency of accuracy from data set to data set)

Bias-Variance Trade-Off

- Bulls Eye diagram - plots representing combinations of both high and low bias and variance
- Centre of the target is a model that perfectly predicts the correct values

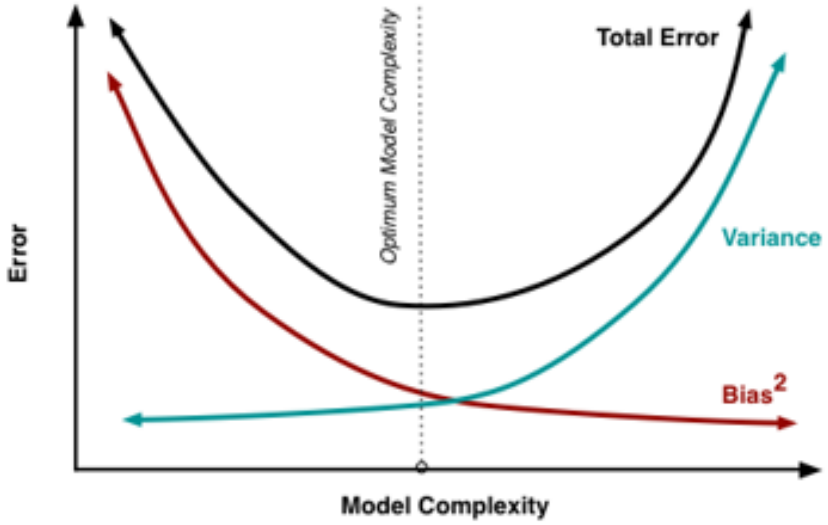


Source: http://scott.fortmann-roe.com/docs/BiasVariance.html

Bias-Variance Trade-Off

- Understanding these two types of error can help us diagnose model results and avoid the mistake of over- or under-fitting.

- Our goal is to achieve low bias and low variance.
  - High bias leads to under fitting
  - High variance leads to over fitting



Source: http://scott.fortmann-roe.com/docs/BiasVariance.html

Cost Function

Cost Function

- A cost function is a measure of how wrong the model is in terms of its ability to estimate the relationship between X and Y
- This is typically expressed as a difference or distance between the predicted value and the actual value.
- Objective of a ML model is to find parameters, weights or a structure that minimizes the cost function.

Cost Function

- The cost function helps to figure out the best possible values for β_0 and β_1 which would provide the best fit line for the data points
- Since we want the best values for β_0 and β_1, this search problem is converted into a minimization problem, to minimize the error between the predicted value and the actual value.
- We square the error difference and sum over all data points and divide that value by the total number of data points.
- Therefore, the cost function is also known as the Mean Squared Error(MSE) function.
- Using this function, β_0 and β_1 are changed to reduce the cost function (MSE)

Cost Function

greatlearning

Hypothesis:

$$h_\theta(x) = \theta_0 + \theta_1 x$$

Parameters:

$$\theta_0, \theta_1$$

Cost Function:

$$J(\theta_0, \theta_1) = \frac{1}{2m} \sum_{i=1}^{m} (h_\theta(x^{(i)}) - y^{(i)})^2$$

Goal:

$$\underset{\theta_0, \theta_1}{minimize} J(\theta_0, \theta_1)$$

Regularization / Shrinkage methods

Regularization

greatlearning

Overfitting

- Your model should ideally fit an **infinite sample** of the type of data you're interested in.
- In reality, you only have a **finite set** to train on. A good model for this subset is a good model for the infinite set, up to a point.
- Beyond that point, the model quality (measured on new data) starts to **decrease**.
- Beyond that point, the model is **over-fitting** the data.

Regularization

- One way to deal with overfitting is regularization.
- Overfitting is typically caused by inflation of the coefficients.
- To avoid overfitting, the coefficients should be regulated by penalizing potential inflation of coefficients.
- The coefficients are penalized by adding the coefficient terms to the cost function.
- If the coefficients become large, the cost increases significantly.

Regularization

Approaches for adding a penalty to the cost function.

- LASSO Regression
- Ridge Regression
- Elastic Net Regression

Ridge Regression

Ridge Regression

- **It is the summation of the squared value of coefficients / slope.**
- Ridge helps to reduce or shrink the variance and making prediction less sensitive to the unwanted variable but not removes it
- The shrinkage of the coefficients is achieved by penalizing the regression model with a penalty term called L2-norm, which is the sum of the squared coefficients.
- The amount of the penalty can be fine-tuned using a constant called lambda (λ). Selecting a good value for λ is critical.
- When λ=0, the penalty term has no effect, and ridge regression will produce the classical least square coefficients. However, as λ increases to infinite, the impact of the shrinkage penalty grows, and the ridge regression coefficients will get close zero.

Ridge Regression

- L2 norm cost function to minimize is given below,

$$\min\left(\left|\left|Y - X(\theta)\right|\right|_2^2 + \lambda||\theta||_2^2\right)$$

- λ given here is actually denoted by alpha parameter in the ridge function. So by changing the value of alpha we are controlling the penalty term

Ridge Regression

- It shrinks the parameters, therefore it is mostly used to prevent multicollinearity.
- One important advantage of the ridge regression,
  - It still performs well, compared to the ordinary least square method in a situation where you have a large multivariate data with the number of predictors (p) larger than the number of observations (n).
- One disadvantage of the ridge regression,
  - It will include all the predictors in the final model, unlike the stepwise regression methods which will generally select models that involve a reduced set of variables.
- Ridge regression shrinks the coefficients towards zero, but it will not set any of them exactly to zero. The lasso regression is an alternative that overcomes this drawback.

LASSO Regression

LASSO Regression

- It is “Least Absolute Shrinkage and Selection Operator”
- **It is the summation of the absolute value of the coefficients.**
- It shrinks the regression coefficients toward zero by penalizing the regression model with a penalty term called **L1-norm**, which is the sum of the absolute coefficients.
- LASSO is used when you have more variables and when you want to remove unwanted variables to the model, as it can bring the value to 0
- Lasso can be also seen as an alternative to the subset selection methods for performing variable selection in order to reduce the complexity of the model.

LASSO Regression

- L1 norm cost function to minimize is given below,

$$\min\left(\left|\left|Y - X\theta\right|\right|_2^2 + \lambda||\theta||_1\right)$$

- λ is the hypermeter, whose value is equal to the alpha in the Lasso function

LASSO Regression

LASSO is generally used when we have more number of features, because it automatically does feature selection.

One obvious advantage of lasso regression over ridge regression,

- It produces simpler and more interpretable models that incorporate only a reduced set of the predictors.

However, neither ridge regression nor the lasso will universally dominate the other.

LASSO and Ridge

- The lasso has a major advantage over ridge regression, in that it produces simpler and more interpretable models that involved only a subset of predictors.
- The lasso leads to qualitatively similar behavior to ridge regression, in that as λ increases, the variance decreases and the bias increases.
- The lasso can generate more accurate predictions compared to ridge regression.
- Cross-validation can be used in order to determine which approach is better on a particular data set.

Elastic Net Regression

- Elastic Net produces a regression model that is penalized with both the L1-norm and L2-norm.
- It is a combination of both L1 and L2 regularization.
- The consequence of this is to effectively shrink coefficients (like in ridge regression) and to set some coefficients to zero (as in LASSO).

Elastic Net Regression

- It uses both L1 and L2 penalty term
- Cost function to minimize is,

$$\min\left(\left|\left|Y - X\theta\right|\right|_2^2 + \lambda_1||\theta||_1 + \lambda_2||\theta||_2^2\right)$$

Conclusion

- Understand best subset selection and stepwise selection methods for reducing the number of predictor variables in regression.

- Indirectly estimate test error by adjusting training error to account for bias due to overfitting
  - Adjusted R^2
  - Estimation of the quality of model can as well be checked by AIC and BIC
  - AIC (Akaike information criterion)
  - BIC (Bayesian information criterion)
  - Difference Between AIC and BIC should be as low as possible for

Thank You