**Microsoft:** DAT203x Data Science and Machine Learning Essentials

## KEY POINTS

- When operating on the data in the training and test datasets, the difference between the known ***y*** value and the value calculated by ***f(x)*** must be minimized. In other words, for a single entity, ***y - f(x)*** should be close to zero. This is controlled by the values of any baseline variables (***b***) used in the function.
- For various reasons, computers work better with smooth functions than with absolute values, so the values are squared. In other words, $(y - f(x))^2$ should be close to zero.
- To apply this rule to all rows in the training or test data, the squared values are summed over the whole dataset, so $\Sigma(y_i - f(x_i))^2$ should be close to zero. This quantity is known as the sum of squares errors (or SSE). The regression algorithm minimizes this by adjusting the values of baseline variables (***b***) used in the function.
- Minimizing the SSE for simple linear regression models (where there is only a single ***x*** value) generally works well, but when there are multiple ***x*** values, it can lead to over-fitting. To avoid this, the you can use ridge regression, in which the regression algorithm adds a regularization term to the SSE. This helps achieve a balance of accuracy and simplicity in the function.
- Support vector machine regression is an alternative to ridge regression that uses a similar technique to minimize the difference between the predicted ***f(x)*** values and the known ***y*** values.
- To determine the best algorithm for a specific set of data, you can use cross-validation, in which the data is divided into folds, with each fold in turn being used to test algorithms that are trained on the other folds.
- To determine the optimal value for a parameter (***k***) for an algorithm, you can use nested cross-validation, in which one of the folds in the training set is used to validate all possible ***k*** values. This process is repeated so that each fold in the training set is used as a validation fold. The best result is then tested with the test fold, and the whole process is repeated again until every fold has been used as the test fold.


Download on the
App Store


ANDROID APP ON
Google play