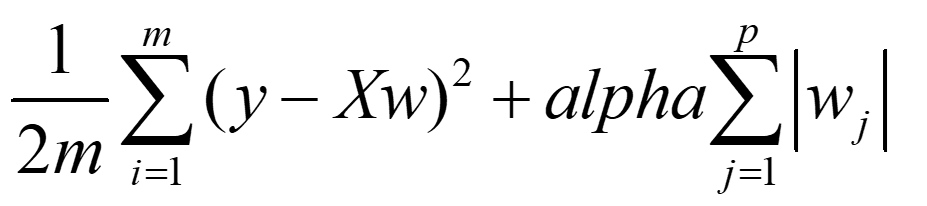
# **Machine learning**

# 10. Explain how does the adjusted R-squared penalize the presence of unnecessary predictors in the model?

The adjusted R-squared is a modified version of R-squared that has been adjusted for the number of predictors in the model. The adjusted R-squared increases only if the new term improves the model more than would be expected by chance. It decreases when a predictor improves the model by less than expected by chance. The adjusted R-squared can be negative, but it’s usually not. It is always lower than the R-squared.

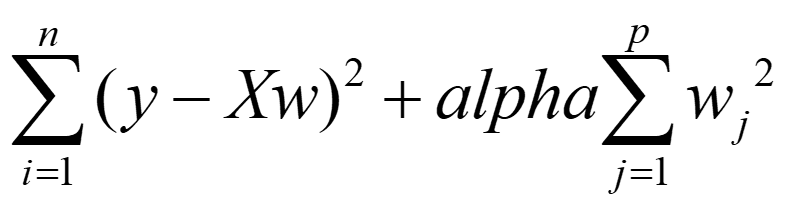
11. Differentiate between Ridge and Lasso Regression.

Lasso is a modification of linear regression, where the model is penalized for the sum of absolute values of the weights. Thus, the absolute values of weight will be (in general) reduced, and many will tend to be zeros. During training, the objective function become:



Lasso introduced a new hyperparameter, *alpha*, the coefficient to penalize weights.

Ridge penalizes the model for the sum of squared value of the weights. Thus, the weights not only tend to have smaller absolute values, but also really tend to penalize the extremes of the weights, resulting in a group of weights that are more evenly distributed. The objective function becomes:



12. What is VIF? What is the suitable value of a VIF for a feature to be included in a regression modelling?

The Variance Inflation Factor (VIF) measures the severity of multicollinearity in regression analysis. It is a statistical concept that indicates the increase in the variance of a regression coefficient as a result of collinearity. In general, a VIF above 5 indicates high correlation and is cause for concern.

13. Why do we need to scale the data before feeding it to the train the model?

Feature scaling is essential for machine learning algorithms that calculate distances between data. ... Therefore, the range of all features should be normalized so that each feature contributes approximately proportionately to the final distance

14. What are the different metrics which are used to check the goodness of fit in linear regression?

• R-square

• Adjusted R-square

• Root mean squared error (RMSE)

15. From the following confusion matrix calculate sensitivity, specificity, precision, recall and accuracy. Actual/Predicted True False

True 1000 50

False 250 1200

Sensitivity=TP/(TP+FN)=1000/(1000+250)=0.8

Specificity=TN/(TN+FP)=1200/1200+50=0.96

Recall=Precision=TP/(TP+FP)=1000/1000+50=0.9523

Accuracy=(TP+TN)/(TP+TN+FN+FP)=0.91142