Machine Learning Task

Mostafa Ahmed Abdelsattar

Question 1 Explain the difference between correlation and multicollinearity. Also, describe the different statistical tests used to detect multicollinearity.

Correlation determines the strength of linear relationship between two variables, while multicollinearity means that two or more independent variables are highly correlated which is problematic when trying to identify the individual effect on the dependent variable.

We can detect multicollinearity using Correlation matrix, Variance inflation factor and Eigenvalues

Question 2 What are the different types of Gradient Descent algorithms and regularizations? Explain when each type should be used, and the advantages and limitations of each.

Batch Gradient Descent, we use it when the data set is small, it's slow compared to others, but converges to the global minimum

Stochastic Gradient Descent, we use it when the data set is large, it's fast compared to others, but it doesn't converge to exactly the global minimum, it keeps oscillating, we should decrease learning rate when using it, to minimize the range of oscillation around the global minimum

Batch Gradient Descent, it comporomises the speed of SGS and the accuracy of BGS, but still requires choosing an appropriate batch size which can impact performance, it's popular in deep learning

Regularization is often used to solve the problem of overfitting in machine learning by reducing the magnitude of coefficients

L2 Regularization, it adds a penalty term based on the sum of squared of coefficients, so it reduces large coefficients but does not shrink them to zero, it's useful when all features are important.

L1 Regularization, it adds a penalty term based on the sum of absolute values of coefficients, it can shrink some features to zero effectively performing future selection, it helps when we doubt that some features are not important