ASSIGNMENT: 8

Name: FALAK KHAN
Batch: MACHINE LEARNING AND AI BATCH A3

1) When should you use L1 regularization over L2 regularization?

Ans: L2 regularization can deal with multicollinearity (independent variables are highly correlated) problems through constricting the coefficient and by keeping all the variables. L2 regression can be used to estimate the significance of predictors and based on that it can penalize the insignificant predictors. L1 regularization is the preferred choice when having a high number of features as it provides sparse solutions. Even, we obtain the computational advantage because features with zero coefficients can be avoided.

2) Explain the difference between Ridge Regularization and Lasso Regularization.

Ans: Ridge Regularization: In Ridge Regularization, we add a penalty term that is equal to the square of the coefficient. The L2 term is equal to the square of the magnitude of the coefficients. We also add a coefficient lambda to control that penalty term. In this case, if lambda is zero then the equation is the basic, or else if lambda>0 it will add a constraint to the coefficient. As we increase the value of lambda this constraint causes the value of the coefficient to tend towards zero. This leads to both low variance and low bias.

Lasso Regularization: Lasso Regularization stands for Least Absolute Shrinkage and Selection Operator. It adds penalty terms to the cost function. This term is the absolute sum of the coefficients. As the value of coefficients increases from 0, this term penalizes, causing the model, to decrease the value of coefficients in order to reduce loss. This difference between ridge and lasso regression is that it tends to make coefficients to absolute zero as compared to Ridge which never sets the value of coefficient to absolute zero.

3) When should one use L1 and L2 regularization instead of dropout to reduce overfitting?

Ans: Dropout actually does a little bit more than just providing a form of regularization, in that it is really adding robustness to the network, allowing it to try out many different networks. This is true because the randomly deactivated neurons are essentially removed for that forward/backward pass, thereby giving the same effect as if you had used a totally different network. L1 versus L2 is easier to explain, simply by noting that L2L2 treats outliers a little more thoroughly - returning a larger error for those points.