We know that the regularization is used to train the models which generalizes better on unseen data by preventing the algorithm from overfitting the training dataset. It also gives more importance to only important features while ignore the others thus reducing the model complexity. If we set the C values between 10^-7and 10^7.then model gives us the varied Conditional Log Likelihood values. Both the test and train set follow the same growth trend for CLL for given C values except dataset 3. Train and Test CLL values are increasing with the increase in model complexity and for the C values from 10^-7 to 10^-1 or 10^0. While train and test CLL becomes stable or constant for values of C from 10^0 or 10^1. There is significant increase in L2 model complexity with the increase in C values from 10^-7to 10^0 Model complexity tends to become almost constant with the higher values of C (over 10^7). For dataset 6, the model complexity reaches its maximum i.e. 86K for C value 10^7. However, Conditional log likelihood computed for test set is less negative than that of train set i.e. absolute sum of train set is greater than absolute sum of test sum.

Figures-

Favorite\_complexity\_vs\_overfit.png

For dataset 3, we can see that Conditional log likelihood does not follow the same trend in growth for train and test set with the increase in C values and model complexity. CLL for train set increases until C value 10^-1 or 10^0 and becomes constant thereafter. However, test set shows that CLL increases until C value 10^-5 and significantly drops thereafter. We can see that there is overfitting to training set for dataset 3.

Favorite\_feature\_selection.png

We know that L2 places emphasis on larger weights and forces them to get closer to zero and L1 also penalizes large weights and make few weights zero. But for dataset 4, we can see that for L2, 1 feature is not used for decision while L1 says that 49 features are not considered until value of C 10^-3. L1 and L2 weights are not seem to follow the trend with values of C. Thus we can say that L2 and L1 does not seem to prove on a same number of features to make the decision in case of feature selection.