Logistic Regression -

It uses sigmoid function as the decision function and loss function in logistic regression is cross entropy loss.

SVM –

Support vector machines attempt to pass a linearly separable hyperplane through a dataset in order to classify the data into two groups. This hyperplane is a linear separator for any dimension; it could be a line (2D), plane (3D), and hyperplane (4D+).

Data description –

The data contains 11 different categories of 256\*256\*3 pixel images. The class imbalance is not very high.

EDA and Feature selection –

The image dataset has very high number of features, hence it is very difficult to train on basic statistical machine learning algorithms. In order to reduce the size of features, applied PCA on the entire dataset and observed that the 1000 principal components are preserving around 98 percent of variance. Hence choose the first 1000 principal components as features.

Modelling Choices –

Initially tried with a basic logistic regression with OVR method and with out any regularization and found that the model is overfitting the image data and hence tried to insert l2 penalty term. By using grid search and cross validation got the best l2 penalty value. But still with the given amount of data and added regularization the model is not able to fit the data properly, under training it got 40% of accuracy on train set and 28% of accuracy on test set.

Empirical results and comparisons –

The results for models with different parameters and their respective scores are in the jupyter notebook.