Decision Tree:

In both datasets, when the value of max\_depth is lower, the models underfit. As the max\_depth increases, the models overfit. When max\_depth increases beyond 10, as the models fit exactly like training data resulting to overfitted models. It looks like a good max\_depth for both datasets will be between 5 and 10 based on initial model exploration. We will now use GridSearchCV to find the most optimal max\_depth. While we are at it, let’s also find the optimal min\_samples\_leaf.

Looking at the learning curves of both datasets, we can see that our first dataset (with small sample) suffers from high variance (backed up by the spread of light green and light green area). As the training instances increase, there isn’t much improvement in cross validation score. The high variance can be reduced using less features and by increasing the training samples. In case of second dataset, as training instances increase, the cross-validation score increases. The predictions for second dataset also have a high variance. As mentioned earlier, we can decrease the variance by removing some features (making our tree less complex) or by using performing ensemble learning like boosting.

Support Vector Machine (SVM)

During our early model exploration, we see that “linear” kernel performs better than “sigmoid” kernel in both datasets. For “sigmoid” kernel, as value of “C” increases beyond 0.1, the models start underfitting. In dataset one, the difference between training accuracy and testing accuracy indicating that the dataset suffers from high variance (also deduced in decision tree section). It looks like, for both dataset, the optimal kernel will be “linear” and the optimal value of “C” will be between 0.1 to 100. We will now use GridSearchCV to find the most optimal kernel and value of C.

Looking at the learning curves, for both datasets, the training score and cross validation score converge as the number of training instances increase. Even with less regularization (increased value of C), the models do not improve indicating underfitting. Therefore, this model will benefit from increase training samples for both datasets.

K Nearest Neighbors (KNN)

During our early model exploration, both datasets suffer from overfitting for k\_neighbors less than six indicating overfitting. However, as the value of k increases, the testing and training accuracy start to converge for both datasets. Based on the accuracy graphs, for both datasets, the optimal value of k\_neighbors will be around 10. We will now use GridSearchCV to find the most optimal value of k\_neighbors.

Looking at the learning curves, for both datasets, both training score and cross validation score increase as training instances increase indicating that the model will benefit from more data. We can also observe that the bias and variance both decrease as more training samples are introduced further solidifying the idea that the model will highly benefit from introduction of more data.