Datasets Description

We used in this project two dataset:

1) Digits dataset offered by Scikit-learn library
2) REALDISP Activity Recognition Dataset Data Set

2.3) Analyze the classifiers behavior:

Classifier Name	With Standardization		Without Standardization	
	Running Time (Sec)	Accuracy	Running Time	Accuracy
Perceptron	0.090	0.951	0.104	0.927
Liner SVM	0.059	0.979	0.0648	0.981
Non-Liner SVM	0.14	0.97	0.45	0.409
Decision tree	0.024	0.853	0.026	0.859
K-nearest neighbor	0.086	0.975	0.105	0.99
Logistic Regression	0.1765	0.959	0.173	0.96

Table 1: Applying Different classifiers on Digits dataset offered by Scikit-learn library

Classifier Name	With Standardization		Without Standardization	
	Running Time (Sec)	Accuracy	Running Time	Accuracy
Perceptron	70.207	0.97	70.207	0.7394
Liner SVM	234.4	0.998	2714	0.69
Non-Liner SVM	675.31	0.994	It takes a long time	
Decision tree	39.24	0.987	41.83	0.987
K-nearest neighbor	316.6	0.997	13.67	0.96574
Logistic Regression	529.57	0.99	234.2	0.742

Table 2: Applying Different classifiers on REALDISP Activity Recognition Dataset Data Set

In the table 1, linear SVM and KNN classifier are the best for the first dataset. Also, we can see that the classifiers did not be affected so much by using with /without standardization which means that the original feature scaling of dataset is good. Expect non-liner SVM do very well with features scaling. But, it fails to create non-linear combinations of features (without standardization) to uplift the samples onto a higher-dimensional feature space where you it uses a linear decision boundary to separate your classes.

In the table 2, decision tree classifier and KNN classifier are the best for the second dataset. Also, we can see all the classifiers are affected by standardization. Therefore, the accuracy is increased, Except the accuracy of decision tree classifier does not affect by using feature scaling. Moreover, we can see that the accuracy both of liner or non-liner SVM (with standardization) is very good among other classifiers. But, there is a big difference Running time between linear SVM and non-linear SVM because non-liner SVM has heavy mathematical computation. In addition, both liner SVM and non-liner SVM are affected by changing the value of variable C which helps to increase or decrease the width of the margin and Gamma “defines how far the influence of a single training example reaches, with low values meaning ‘far’ and high values meaning ‘close’”. In addition, KNN has affected by changing the value of k.in this project, I use the K=1. Increasing the value of K will increase the running time but it will decrease the accuracy. Moreover, the behavior of KNN Classifier Compute a distance value between the item to be classified and every item in the training data-set. Pick the k (in this project, k value is 1) closest data points (the items with the k lowest distances. Conduct a “majority vote” among those data points the dominating classification in that pool is decided as the final classification. The idea of Logistic Regression is to find a relationship between features and probability of outcome. In addition, Logistic Regression, which again is a classifier but uses a more complex activation function derived from taking the inverse of logit (link) function which gives you the output in the range between 0 to 1. Moreover, perceptron classifier is affected by using standardization. Moreover, perceptron classifier calculates the sum of multiply the input and the weight for each instance and it check if the sum is greater than zero or not. If it greater the prediction output will 1 otherwise will be -1. Then, it checks the prediction output is equal to actual prediction. If it is equal, it will not update the weight otherwise it will update. All of classifiers have several parameters which can affect the accuracy of classifier such max depth parameter in decision tree classifiers.

2.4

- Pre-pruning strategy is that stop if the current node does not improve impurity (min_impurity_decrease OR min_impurity_split).

  Pre-pruning strategy is max_depth

- https://github.com/scikit-learn/scikit-learn/blob/bac89c2/sklearn/tree/tree.py#L518
  Lines of code :278-293 (first strategy)
  Line of code: 171-172 (second strategy)