REPORT ON LINEAR CLASSIFIERS

BRIEF DESCRIPTION:

- Dataset: It is obtained from Kaggle – Heart disease dataset.
- The description of features(13) are mentioned:

Features	Description
Age	Age
Sex	Gender
cp	chest pain type
trestbps	resting blood pressure (in mm Hg on admission to the hospital
chol	serum cholestoral in mg/dl
fbs	(fasting blood sugar &gt; 120 mg/dl) (1 = true; 0 = false)
restecg	resting electrocardiographic results
thalach	maximum heart rate achieved
exang	exercise induced angina (1 = yes; 0 = no)
oldpeak	ST depression induced by exercise relative to rest
slope	the slope of the peak exercise ST segment
ca	number of major vessels (0-3) colored by flourosopy
thal	3 = normal; 6 = fixed defect; 7 = reversable defect

- The label is target (1 means person is having heart disease and 0 just opposite).

a) The analysis of linear sepability is done using two features- ***trestbps and thalach***

b) Linear separability is tested using Perceptron algorithm which gives accuracy score of **60.39%** using above two features.

Working process for different classifiers:

1. Split the dataset using train-test split from sklearn in 80:20 ratio.

2. Scale the features by importing Standard Scaler from sklearn.preprocessing.

3. Now import the required classifier from sklearn and fit the X and Y (train) component of split in the model and predict the values on test set.

4. Calculate the accuracy score, classification matrix and confusion matrix by importing from sklearn.metrics. Also plot AUC(Area under curve) to test the fit quality of each classifier.

Analyze the results obtained by using the different classifiers:

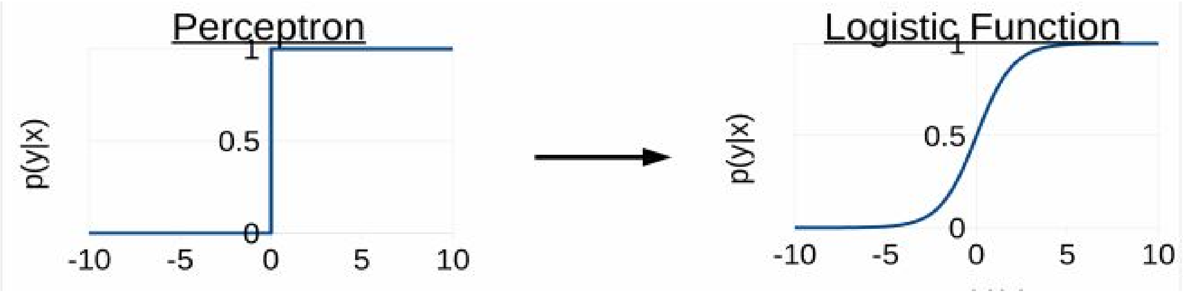
Split Chosen 80:20(80% training, 20% testing)	
Type of Classifier	**Accuracy Score**
Half Space	0.8524
Logistic Regression (using inbuilt function)	0.8524
SVM classifier (using a linear kernel)	0.8688
SVM classifier (Polynomial kernel)	0.9016
SVM classifier(Gaussian kernel)	0.8688
Logistic Regression using the SGD procedure	0.7868

Comparison:

- ## Logistic Regression v/s Half space:

a) Logistic Regression gives almost same accuracy as that of Half space. Sigmoid function used in logistic is differentiable so it gives values wide variety of values between 0 to 1. Whereas Half space gives either 0 or 1.

b) The dataset is balanced and training examples less(<500) here so the accuracy is same in both cases. Logistic Regression can also account for uncertainty.



Source: http://www.cs.umd.edu/class/fall2017/cmsc723/slides/slides_04.pdf

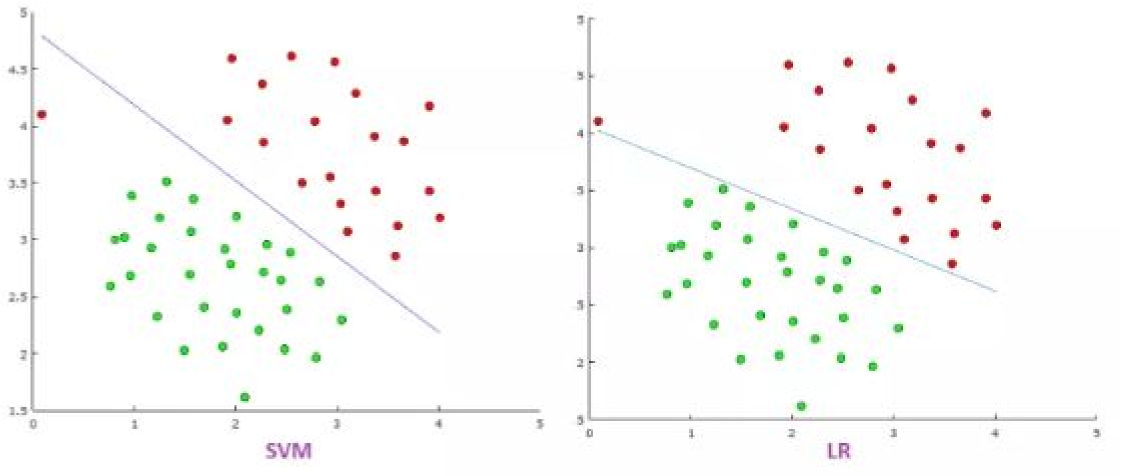
- ## Logistic Regression v/s SVM Classifier:

a) SVM classifiers give better accuracy than Logistic Regression as SVM kernels map the features in high dimension space and and classify using high polynomial degree fitting and higher dimension curves.

b) SVM finds the best margin which can separate classes and error is reduced Whereas in Logistic Regression has different decisions boundaries with different weights which are close to optimal points.

c) SVM have dual form which gives sparse solutions(using kernel trick) and has better scalability.

d) SVM with linear kernel is gives almost same result as of Logistic Regression as both are classifying using linear line but SVM selects the best margin so its accuracy is more.



Source:https://gdcoder.com/support-vector-machine-vs-logistic-regression/

- **Logistic Regression(LR) v/s LR with SGD:**

a) As we know that LR uses gradient descent which converges easily without having much noise but LR with SGD selects some mini batches and it takes so many iterations to converge also it has so much noise.

b) Due to this LR gives better results than LR with SGD(It is not able to Converge due to noise)

c) We can make LR with SGD to perform same as LR by using SGD with using SGD with momentum technique, by which unnecessary noise can be removed and it can converge easily.

Soft SVM Formulation and Support Vectors:

- Since the data is not linearly separable so our soft svm is going to give high error and the no of support vectors will also be more.

Type of Split	No of Support vectors for X_train, y_train
70:30	149
80:20	172
90:10	194

- As we can see the no of support vectors increases with increase in the training dataset %.

Effect of Regularization parameter on Performance:

a) So we found out that with increase in Value of C (Regularization parameter) on linear kernel in SVM, support vector decreases because SVM choses smaller margin hyperplane if the hyperplane classify training points correctly.

b) Similarly small value of C will increase the support vectors as the optimizer finds a large margin separating hyperplane even if the model misclassifies the points

c) The dataset is not linearly separable so the support vectors are too much effect of C is also less visible.