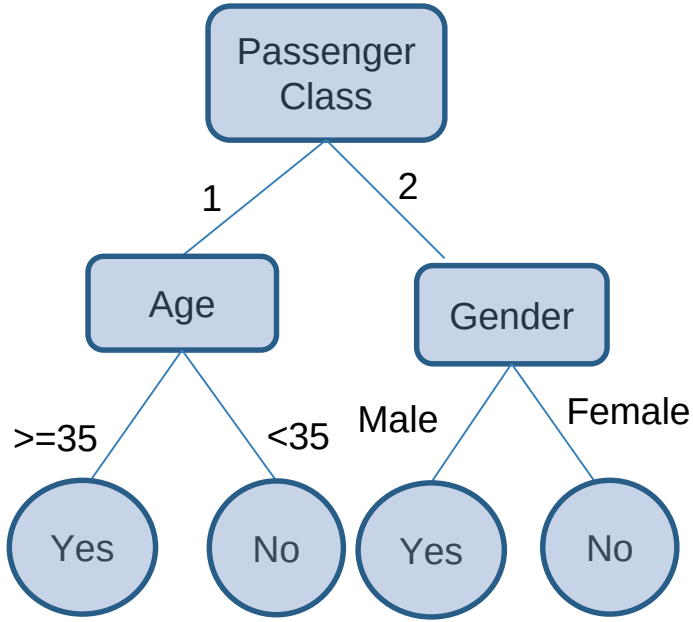
PERFORMANCE ANALYSIS OF DECISION TREES, RANDOM FOREST, XGBOOST TREES AND SVM CLASSIFIER

Subhadeep Dash – S20160010021
V. Sai Rathan – S20160010104

DECISION TREES

- Logic seems easy to visualize.
- Doesn’t require normalization of data, removal of blank values or dummy variables.
- Cost of construction is quite less.



DECISION TREES(Contd..)

- Generally built using CART or ID3 algorithms.
- Use Information Gain or Gini Index for best split.

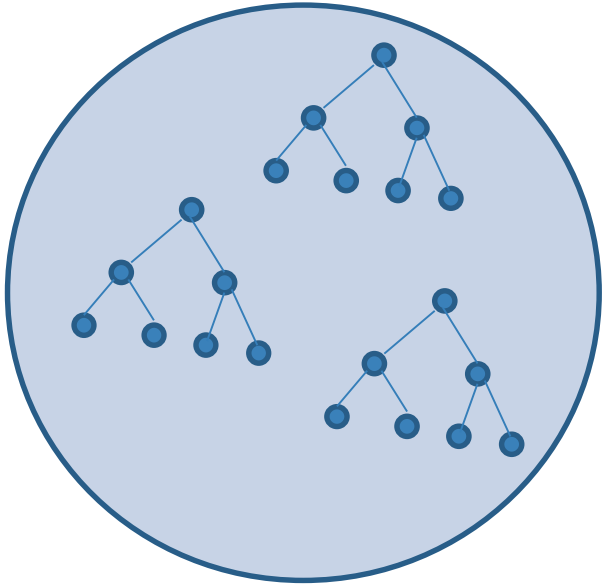
$$Information\ Gain = Entropy(Target) - Entropy(Feature)$$

$$Gini\ Index = 1 - \sum_j p_j^2$$

- Prone to overfitting.
- Pre-pruning or Post-pruning reduces overfitting.

RANDOM FOREST

- Consists of many decision trees.
- More the decision trees, more the accuracy.
- Average output of all the Decision Trees is considered the final output.
- m out of n attributes are considered for recursive split while building a tree.



RANDOM FOREST(Contd..)

- Overfitting is still an issue.
- Hyperparameters, namely no. of attributes needed to be considered and no. of decision trees to be constructed, should be provided.
- Computational cost is higher than Decision Trees.
- Takes relatively more time to build.
- Each tree is constructed parallelly and independently.

XGBOOST TREES

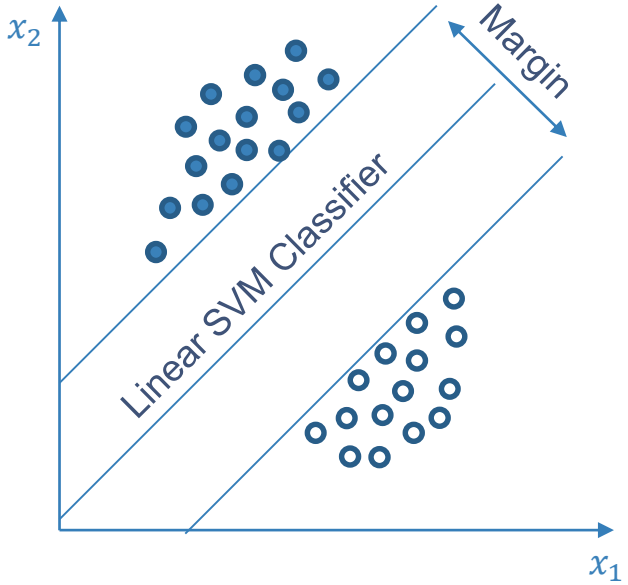
- More than 50% of award winning solutions in Machine Learning challenges hosted in Kaggle adopt XGBoost.
- Involves Gradient Boosting implementation.
- Unlike Random Forest, these are constructed sequentially.
- For every tree built, the error caused in the constructed tree is tried to be rectified in the next tree.

XGBOOST TREES(Contd..)

- An additional hyperparameter i.e. learning rate is required while training.
- Harder to tune when compared to Random Forest.
- Logic or rules can’t be easily interpreted or visualized.
- Training takes more time since trees are built one after the other.
- More sensitive to over-fitting if the data is noisy.

SVM CLASSIFIER

- Considered one of the best options for unsupervised learning.
- Generally resistant to overfitting.
- Works great even when the number of features is large.
- Powerful when an appropriate kernel is used.



SVM CLASSIFIER(Contd..)

- Selection of an appropriate kernel is a difficult task.
- Usage of inappropriate kernel leads to large errors.
- The final model is difficult to visualize or interpret.
- It’s algorithmic complexity is high which may result in slower testing phase.

RESULTS ON TITANIC DATASET

	ACCURACY ON TRAINING SET(%)	ACCURACY ON TESTING SET(%)
Decision Trees	**84.43**	**79.10**
Random Forest	**96.63**	**84.70**
XGBoost Trees	**88.44**	**85.07**
SVM Classifier	**78.97**	**77.99**