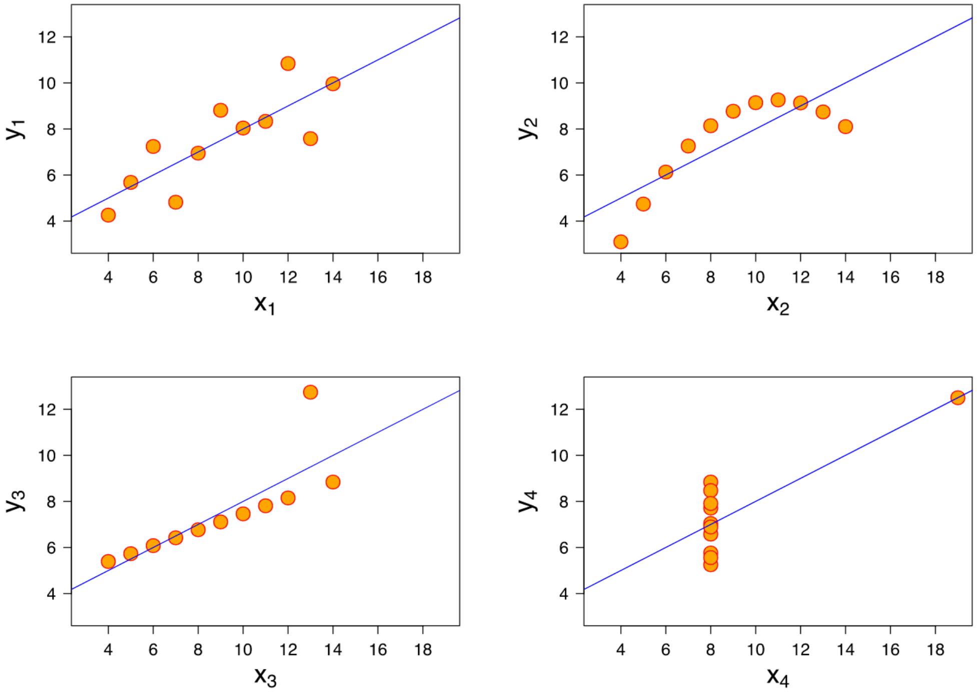
# Ensemble

January 30, 2022

## 1 Ensemble

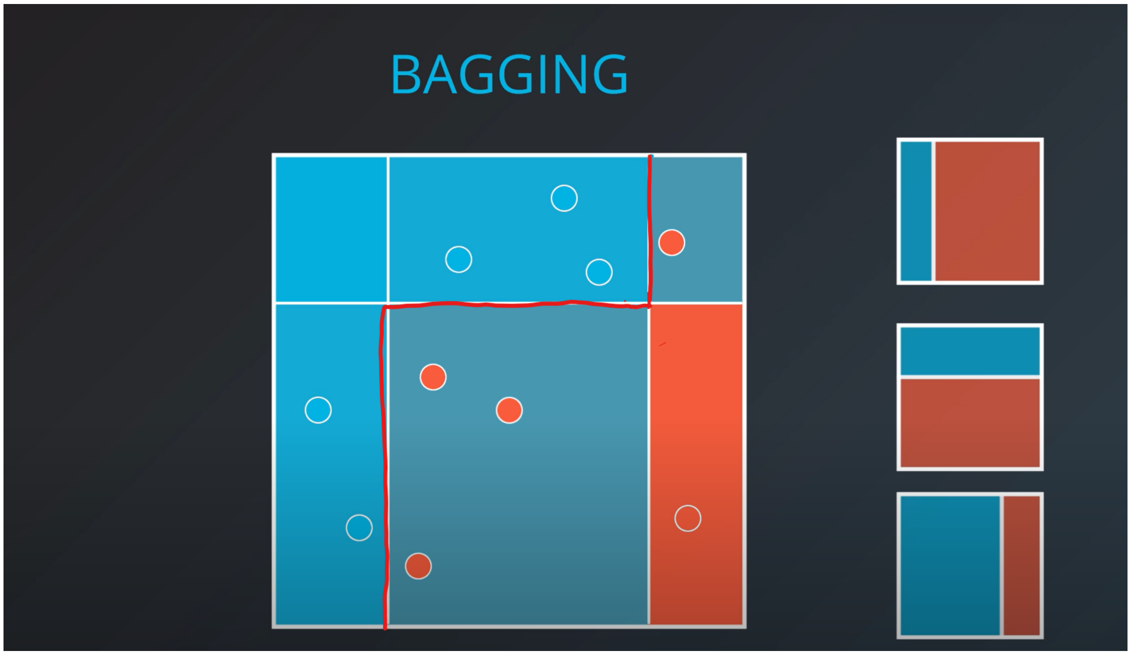
Ensemble is simply a method to combine algorithms for supervised learning together to produce strong and better learning models. The process is to gather weak learners which are basically individual model that perform best at what they do and combine them to produce a strong learning model. Decision trees are used commonly as weak learners and it combined with other methods to make the model stronger. There are two commonly known methods to do this, one is Bagging and second is Boosting.

Bias and Variance are two variables generally used to define if the model is a good learning model. An example of High Bias model is Linear Regression because since the boundary line is linear, for any dataset the model will try to fit linear line. On the contrary Decision Trees have high Variance, which means the fittin if the data is such that if no stopping parameters are specified the model will classify every single last data point. Thus making large number of branches and leaves. Thus Linear Regression model is considered to have High Bias and Low Variance while Decision trees model is considered to have Low Bias and High Variance. However, out aim is to find a best fit of Bias and Variance in a model.



**Random Forests** are a perfect example of Esemble methods which combines many decision trees models to predict the outcome. This method is used when we have a huge amount of data and we have a high Variance. Since decision trees are low bias models they tend to overfit the data and that is not a good model to predict the outcomes.

In **Bagging**, when we have to deal with huge data, the given data is divided into sample subsets and those are trained with our model, say Decision Tree model. Each sample data trained is said to be a weak learner and when we combine them we can get a strong learner thus giving better predictions results.



**AdaBoost** is one of the boosting techniques used to train model on small sample sets and then combine them to make a strong learning model. Weighting of the data is done in each of the sample sets based on the correctly and incorrectly classified points. Next, based on the calculated weights we combine the sample sets into on model and add the weights. Weight will be negative for the region which is oppositely classified. As we can see in above image there are three sample sets that classifying the points.

Formaula of weight is given by $W = ln(\frac{x}{1-x})$, where x is sum of correctly classified points and 1-x is sum of incorrectly classified points for that sample set.

### 1.0.1 Spam Classifier Model with Ensemble:

The spam classifier model was earlier developed using Naive Bayes algorithm as a part of coursework on Udacity. The extension of this is made to use Ensemble of methods like BaggingClassifier, Random Forest Classifier and AdaBoost Classifier. The resultant accuracies of these methods are compared in the notebook. You can find the notebook on my github repository at: https://github.com/hseju/Udacity-Intro-To-ML-TensorFlow/tree/main/Ensemble

```
[ ]:
```