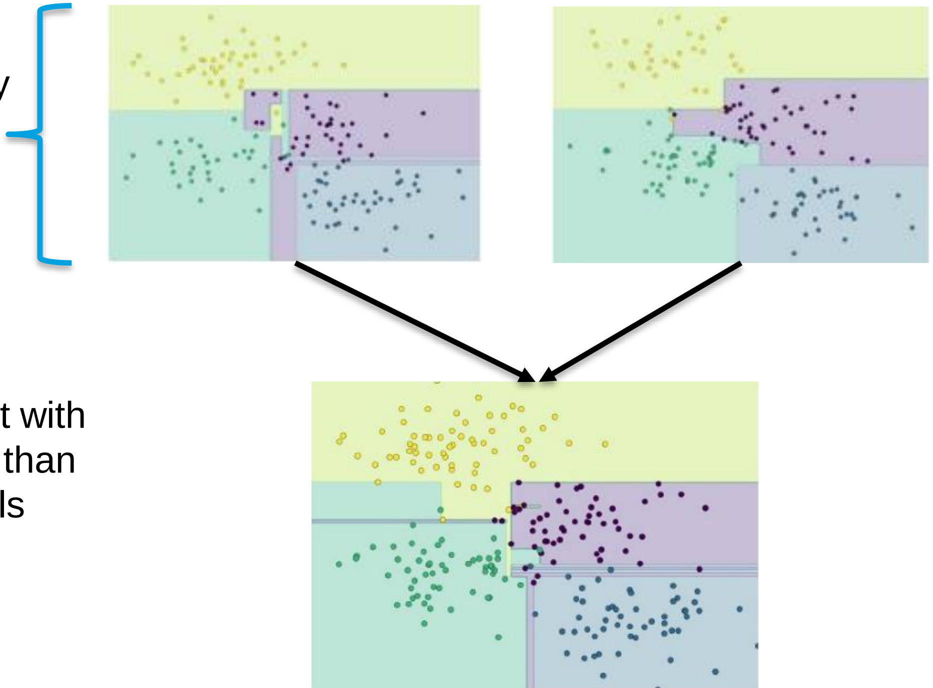
Ensemble Methods

Ensemble Methods -

1. Train multiple weak predictors on a dataset such that they get slightly different results some learn some patterns better and others learn other patterns
2. Combine their predictions to get an overall better performance
3. The combined group of learners is call meta model



Meta models Trained on slightly differ data

Combined result with better accuracy than individual models

Source: Python Data Science Handbook / Page 424

Ensemble Methods -

4. In some parts of the feature space, the different instances produce similar results for e.g. in the extreme regions

4. In regions where the data points from different classes overlap, the instances give different results

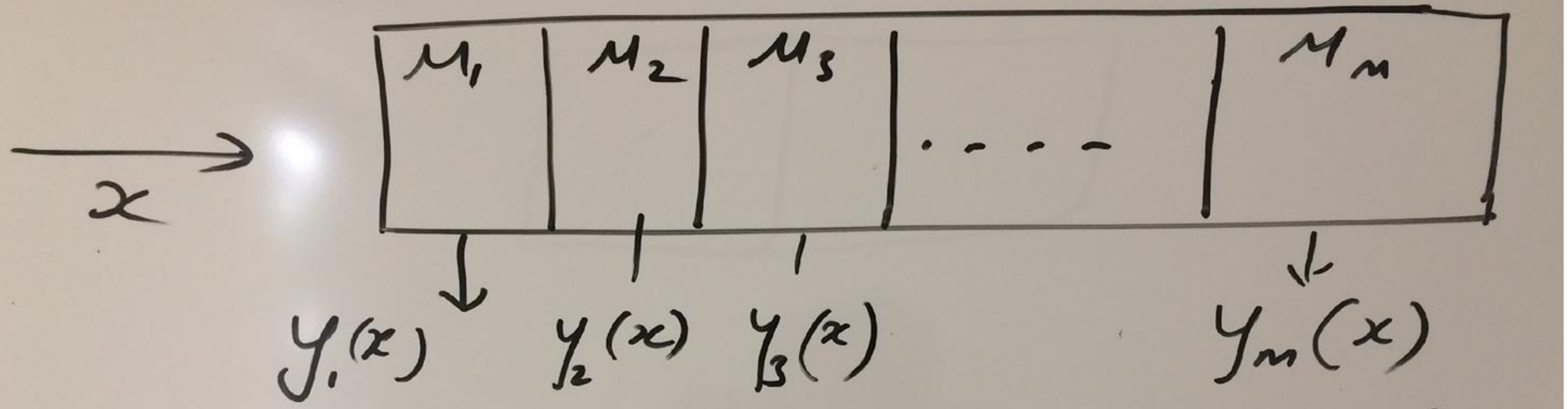
4. Thus by using information from all the instances, may give overall better result than individual instances

Ensemble Methods -

1. To ensure each learner gets to see slightly different data can be done in many ways
    a. Random sampling with replacement (Bootstrapping / sampling with replacement)
    b. By adjusting the weights assigned to each data point to force an instance to focus on certain data points more

1. Only Random Forest ensemble technique is designed only for decision Tree

1. For other ensemble methods, each instance in the meta model can be build out of different algorithm. For instance DecisionTree, Naïve Bayes, Support Vector Machine.



$$y_{com}(x) = \frac{\text{Sum}}{m}\left(\downarrow + \downarrow + \downarrow + \cdots + \downarrow \right)$$

$$\Rightarrow y_{com}(x) = \frac{\sum_{i=1}^{m} y_i(x)}{m}$$

let true y for x be $h(x)$

let error of a model be $e_m(x)$

$$y_m(x) = h(x) + e_m(x)$$

$$e_m(x)^2 = \left(h(x) - y_m(x)\right)^2 \longleftarrow SSE_m \text{ for } x$$

$$\text{Avg } SSE_m = E_x\left(e_m(x)^2\right) \longleftarrow \text{avg acron all data}$$

∴ Avg SSE of all models 1 to M —

$$E_{AV} = \frac{\sum_{i=1}^{m} SSE_i}{m} \longrightarrow \text{Avg SSE of committee}$$

Ensemble avg error is scaled down by a factor of m assuming errors of individual models are independent

Two families of ensemble methods are usually distinguished:

1. Averaging methods, the driving principle is to build several estimators independently and then to average / vote their predictions. On average, the combined estimator is usually better than any of the single base estimator because its variance is reduced.

   E.g. Bagging methods, Forests of randomized trees, ...

1. Boosting methods, base estimators are built sequentially and one tries to reduce the bias of the combined estimator. The motivation is to combine several weak models to produce a powerful ensemble. E.g. AdaBoost, Gradient Tree Boosting, ...

Ensemble Methods – Averaging method - **Bagging (B**ootstrap **Agg**regation) **:**

1. Designed to improve the stability and accuracy of classification and regression models

1. It reduces variance errors and helps to avoid overfitting

1. Can be used with any type of machine learning model, mostly used with Decision Tree

1. Uses sampling with replacement to generate multiple samples of a given size. Sample may contain repeat data points

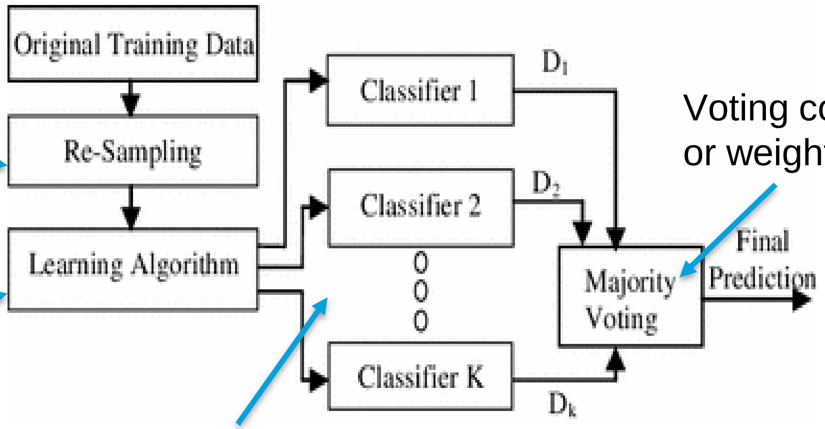
1. For large sample size, sample data is expected to have roughly 63.2% (1 – 1/e) unique data points and the rest being duplicates

1. For classification bagging is used with voting to decide the class of an input while for regression average or median values are calculate

Ensemble Methods – Averaging method - **Bagging (B**ootstrap **Agg**regation) **:**

Re-sampling done for every classifier using a random function. For large n, 63.2% unique samples likely to be selected

Algorithm to generate classifiers. Could be Decision Tree, Naïve Bayes etc



Voting could be simple or weighted

K classifiers created in parallel and independently on respective training data

Source: https://link.springer.com/article/10.1007/s13721-013-0034-x

Ensemble Learning – **Bagging**:

Lab- 6 Improve defaulter prediction of the decision tree using bagging ensemble technique

Description – Sample data is available at local file system as credit.csv

The dataset has 16 attributes described at
https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
or in the notes page of this slide

Sol: Bagging+Credit+Decision+Tree.ipynb

Ensemble Methods – Boosting Method – **AdaBoosting** :

1. Similar to bagging, but the learners are grown sequentially; except for the first, each subsequent learner is grown from previously grown learners

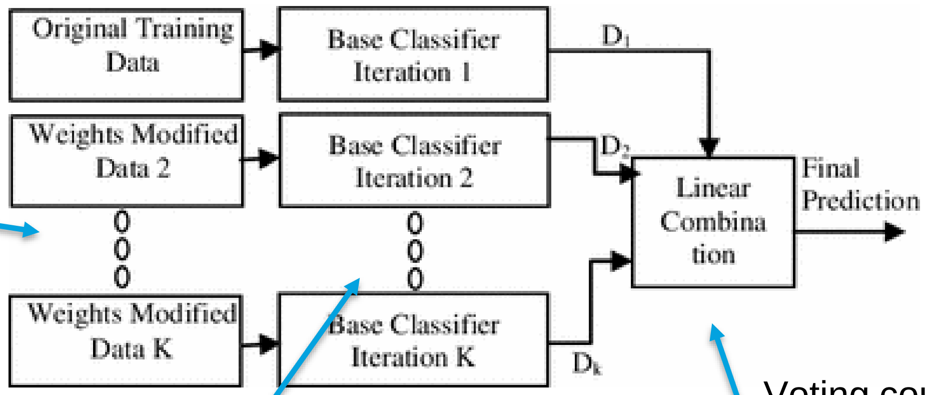
1. If the learner is a Decision Tree, each of the trees can be small, with just a few terminal nodes (determined by the parameter d supplied)

1. During voting higher weight is given to the votes of learners which perform better in respective training data unlike Bagging where all get equal weight

1. Boosting slows down learning (because it is sequential) but the model generally performs well

Ensemble Methods – Boosting method - **AdaBoosting:**

Training data from base data with focus on instances which were incorrectly classified by earlier model (if any)



K similar classifiers created in sequence with respective training data with focus on addressing the misclassified data, not the usual cost functions

Voting could be simple or weighted

It is called Adaptive Boosting as the weights are re-assigned to each instance, with higher weights to incorrectly classified instance

Source: https://link.springer.com/article/10.1007/s13721-013-0034-x

Ensemble Methods – Boosting Method – **AdaBoosting** :

7. Two prominent boosting algorithms are AdaBoost, short for Adaptive Boosting and Gradient Descent Boosting

7. In AdaBoost, the successive learners are created with a focus on the ill fitted data of the previous learner

7. Each successive learner focuses more and more on the harder to fit data i.e. their residuals in the previous tree

Ensemble Learning – **AdaBoosting**:

Lab- 7 Improve defaulter prediction of the decision tree using Adaboosting

Description – Sample data is available at local file system as credit.csv

The dataset has 16 attributes described at https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
or in the notes page of this slide

Sol: Adaboost+Credit+Decision+Tree.ipynb

Ensemble Methods – Averaging Method – **Gradient Descent Boosting** :

1. Each learner is fit on a modified version of original data (original data is replaced with the x values and **residuals** from previous learner

1. By fitting new models to the residuals, the overall learner gradually improves in areas where residuals are initially high

Ensemble Methods – Averaging Method – **Gradient Descent Boosting** :

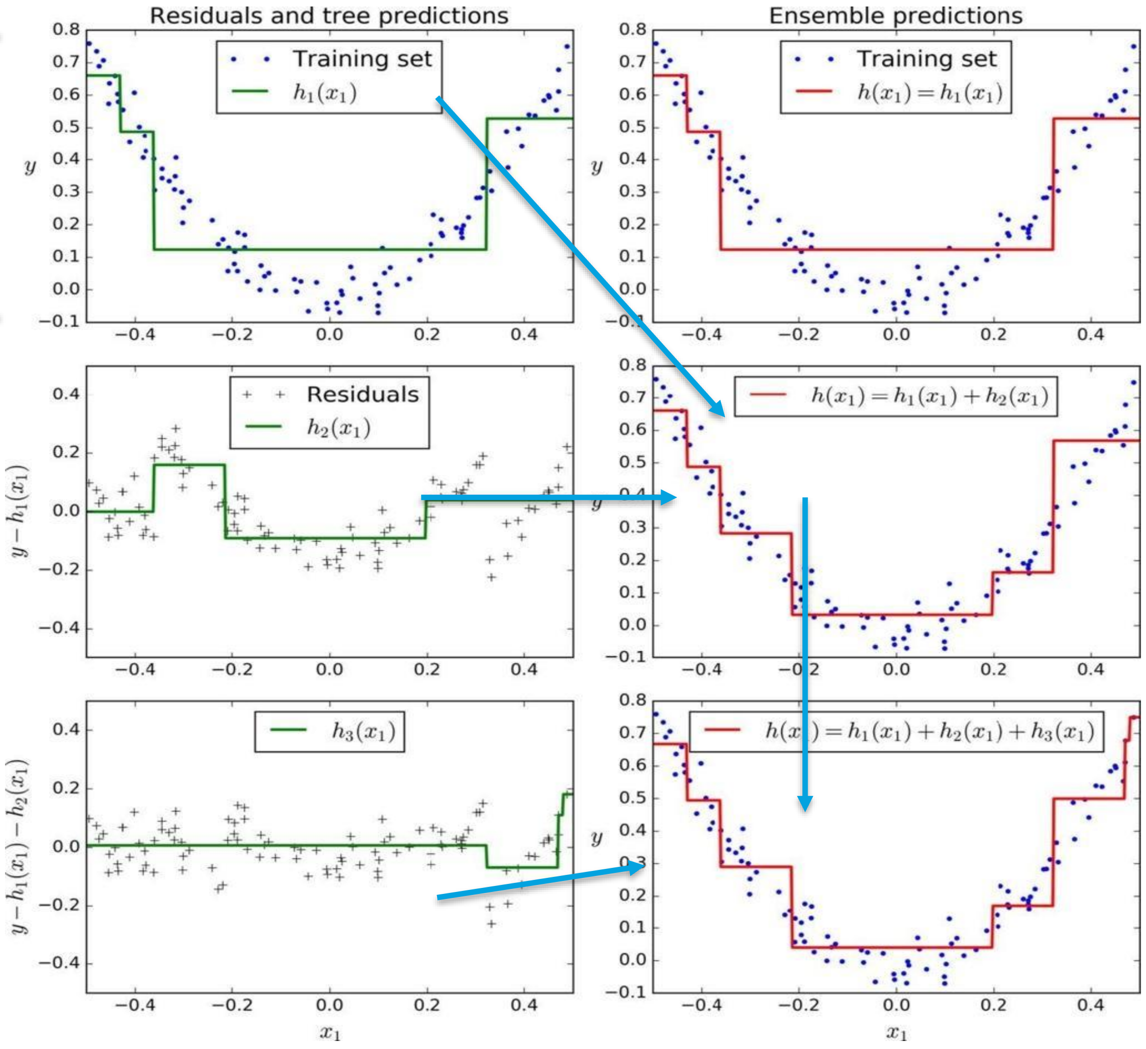
First learner results in residuals (dots that fall above and below the surface. The result (red) is same as first classifier

Next classifier focuses on the residuals of the first classifier to reclassify them as correctly as possible

The combined effect of this surface and previous classifier surface is shown in red

The third learner focusses on the residuals of the previous classifier

The combine result of the new surface with the previous surface is shown in red



Ensemble Learning – **Gradient Boosting**:

Lab- 8 Improve defaulter prediction of the decision tree using Gradient boosting

Description – Sample data is available at local file system as credit.csv

The dataset has 16 attributes described at https://archive.ics.uci.edu/ml/datasets/statlog+(german+credit+data)
or in the notes page of this slide

Sol: GRB+Credit+Decision+Tree.ipynb