Ensemble learning—

It refers to the procedures to combine output of multiple machines and treating them as a committee makers.

The principle is that the resultant decision should have better overall accuracy, on average than any individual committee member.

The members of ensemble might be predicting—

1.) real valued numbers.

2.) class labels.

3.) Rankings and clusterings etc.

4.) posterior probabilities.

An ensemble consists of a set of model and a method to combine them.

The underlying principle of ensemble learning is a recognition Thea in real world situations every machine learning model has limitations and will make errors.

Model Selection—

1- Always use cross validation.

2- The outcomes of each model are usually average out.

3- Final goal should be kept in mind.

4- Study the stability of your data sample.

Data Quantity—

—> We can split the bulky data into many parts.Train each and every part separately.

—> If the dataset is not big in size, use different bootstrap of data to train different classifiers.

(here each bootstrap sample is the random sample of the data drain with replacement and treated as if the was independently drawn from the underlying distribution.)

Divide and Conquer—

The decision boundary that separates the data may be too complex for certain problems.

Here, the classification system follows a divide and conquer approach.

Divide the dataspace in such a way the each classifier learns only of the simpler partitions.

Data fusion—

Suitable combination of data from different sources is called data fusion.

This can lead to improved accuracy.

Confidence estimation—

The decision if highly confident when majority of classifiers agree with their decisions.

It may be noted that ensemble having high confidence decision does mean that decision is correct. However, its vice-versa is correct.

ENSEMBLE LEARNING ALGORITHMS—

1— Bagging/bootstrap aggregating

—>oldest and simplest.

—> diversity is obtained by using bootstrapped replicas.

2— Boosting

—> it also creates replicas by resampling the data.

However, in boosting resembling is strategically geared to provide the most info train data for each consecutive classifiers.

Each iteration- 3 weak classifiers —

1.) classifier c1 trained with random subset.

2.) c2 trained on the most info subset (half of training data correctly classifies by c1, and the other half misclassified)

3.) c3 trained with instances on which c1 and c2 disagree.

Boosting is designed for binary class problems.

Ada Boost( adaplive boosting)

—> instances are drawn into the subset dataset from on iteratively updated sample distribution for the training data.

—> the classifiers training errors decides the voting weights to combined classifiers through weights majority voting.

Random forests—

Ensemble of decision tress trained with a bagging mechanism.

—>They naturally handle both regression and multi class classification.

—> relatively fast and depends only on one or two tuning parameter.

DECORATE(diverse ensemble creation of oppositional relabelling of artificial training examples)

Based on the specially constructed artificial e.g.

—> used to create diverse ensembles.

Combining Ensemble Members—

\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_

1-> majority voting

2-> weighted majority voting

3->borda count.