Ensemble Learning

Junaid Mahmud

Supervise Learning Problem

$$y = f(x)$$

Training Examples

$$(x_1, y_1), ..., (x_m, y_m)$$
$$x_1, x_2, x_3, ..., x_m$$

Hypothesis

A hypothesis is a proposed explanation for a phenomenon. For a hypothesis to be a scientific hypothesis, the scientific method requires that one can test it.

- Assumption of factors/ data
- Scientific method/ process
- Result

We conduct experiments and do tests using some predefined **steps** to get output and make trial and error to get the expected performance. We apply scientific methods to data to predict and get the result.

For example, Bangladesh’s chance of winning a cricket match is based on scored runs.

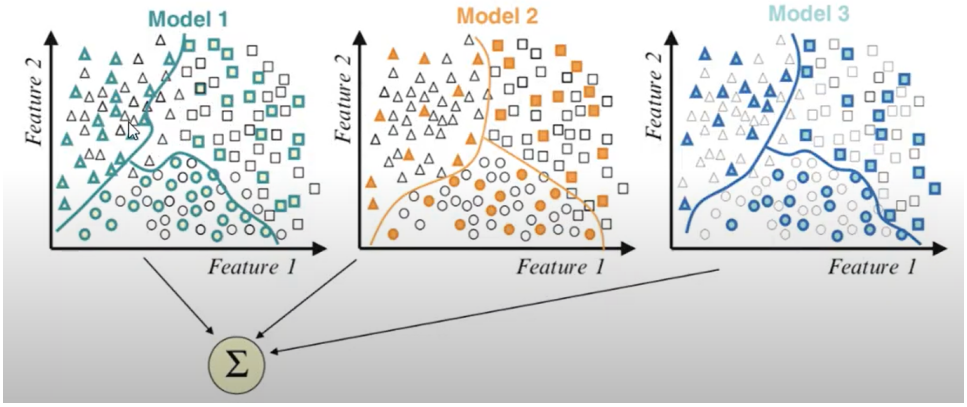
Ensemble

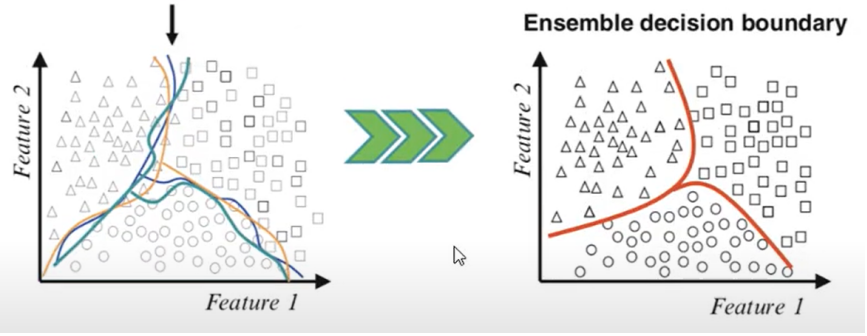
An ensemble of classifiers is a set of classifiers whose individual decisions are combined in some way to classify new examples.

The main idea is that ensembles are often much more accurate than the individual classifiers that make them up.

Given a set S of training example, a learning algorithm outputs a classifier. The classifier is a hypothesis about the true function f. Given new x values, it predicts the corresponding y values. Denote Classifier by h1, … hL.

- True function
- Representational function/hypothesis





- Here there are three models, working to classify a dataset of three classes. They give different results. By ensembling them, we get a combined result.

Ensemble Conditions

- Classifiers are:
  - **Accurate**: Error rate better than random guessing(Accuracy > 50%) on new x values.
  - **Diverse**: Making different errors on new data points.

$$h_1, h_2, h_3$$

Three classifiers are identical (i.e. not diverse)

$$h_1(x) = wrong$$
$$h_2(x) = wrong$$
$$h_3(x) = wrong$$

If they are diverse

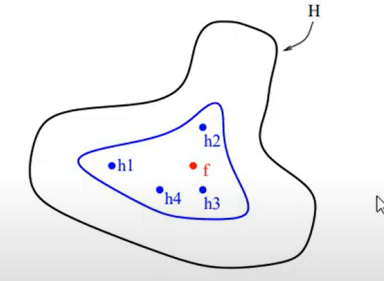
$$h_1(x) = wrong$$
$$h_2(x) = right$$
$$h_3(x) = wrong$$

Reason for making an ensemble

The three most important ways in which existing learning algorithms **fail**

- Statistical Reason
- Computational Reason
- Representational Reasons

Statistical Reason

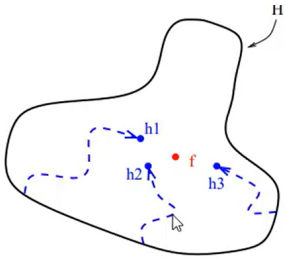


Statistical Reason

A learning algorithm can be viewed as searching a space H of hypotheses to identify the best hypothesis in the space. The statistical problem arises when the amount of training data available is too small compared to the size of the hypothesis space. Without sufficient data, the learning algorithm can find many different hypotheses in H that all give the same accuracy on the training data.

- For an overfitting scenario, the classifier will not do well in the test. For multiple classifiers, the weaknesses and strengths are combined, hence is better than a single classifier.
- If some training data is changed in KNN, the result will not be much different, but for Neural Network, the change will be significant, as NN learns from all the data.

Computational Reason



Computational Reason

Many learning algorithms work by performing some form of local search that may get stuck in local optima.

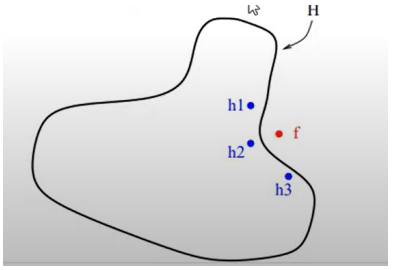
For example, neural network algorithms employ gradient descent to minimize an error function over the training data, and decision tree algorithms employ the greedy splitting rule to grow the decision tree.

There is enough training data (so that the statistical problem is absent), but it may still be very difficult computationally for the learning algorithm to find the best hypothesis.

There is enough training data (so that the statistical problem is absent), but it may still be very difficult computationally for the learning algorithm to find the best hypothesis.

An ensemble constructed by running the local search from many different starting points may provide a better approximation to the true unknown function than any of the individual classifiers.

Representational Reason



Representational Reason

- If the dataset is imbalanced. (One class sample is much bigger than another one), the classifier will not perform well.
- Different weight is assigned to the hypotheses for the true function to be in the region.

In most applications of machine learning, the true function f cannot be represented by any of the hypotheses in H. By forming weighted sums of hypotheses drawn from H, it may be possible to expand the space of representable functions.

- NN
- DT

With a finite training sample, these algorithms will explore only a finite set of hypotheses and they will stop searching when they and a hypothesis that fits the training data.

Simple Ensemble Techniques

Majority/Hard Voting

- 5 Colleagues rating a movie. The final rating is the average of their values.

Colleague 1	Colleague 1	Colleague 1	Colleague 1	Colleague 1	Final Rating
5	4	5	4	4	4

- Classifiers on a data point, make the following predictions.
    - Classifier 1 ⇒ Class 1
    - Classifier 2 ⇒ Class 1
    - Classifier 3 ⇒ Class 2

VotingClassifier (with voting = ‘hard’) would classify the samples as ‘class 1’ based on the majority class label.

⇒ If something like this happens

- Classifier 1 ⇒ Class 1
- Classifier 2 ⇒ Class 2

If multiple classes have the same vote, then labels will be sorted and the first one will be taken.

Averaging

i.e. (5 + 4 + 5 + 4 + 4) / 5 = 4.4

Colleague 1	Colleague 2	Colleague 3	Colleague 3	Colleague 4	Final Rating
5	4	5	4	4	4.4

Weighted Average Probabilities (Soft Voting)

- For regression

	Colleague 1	Colleague 2	Colleague 3	Colleague 3	Colleague 4	Final Rating
Weight	0.23	0.23	0.18	0.18	0.18	
Rating	5	4	5	4	4	4.41

- For Classification

Classifier	Class 1	Class 2	Class 3
Classifier 1	w1 * 0.2	w1 * 0.5	w1 * 0.3
Classifier 1	w2 * 0.6	w2 * 0.3	w2 * 0.1
Classifier 1	w3 * 0.3	w3 * 0.4	w3 * 0.3
Weighted Average	0.37	0.4	0.23

Here we will take the probability value for each data point by the classifier.