Ensemble Methods

Andrés Ponce

May 6, 2020

Ensemble Methods refer to when we use an aggregate of L different models to find more accurate predictions for our data. There are several ways of implementing a selection from several methods: adaboost, decision trees, bagging, random forests, etc...

Boosting

The straightforward approach to using multiple models is to take the average value of each of the models that we use. Supposing we want to use M models, and we want to predict the value given by a certain function, we could say

$$y_{COM} = \frac{1}{M}\sum_{m=1}^{M} y_m(x)$$

where COM refers to the *decision by committee* given by the equations. We take the average value of the M models to form our final result. When calculating the error, we also take the aggregate error of all the models.

Boosting then refers to the building of a strong classifier for a problem by using several weaker classifiers. When boosting, we again can have M models whose scores are summed up.

The type of boosting called **Adaptive Boosting** will pass the input through the models sequentially, and *train* the models sequentially. Then, when training the next model, items that were misclassified will be given greater weight. This may result in an overall model which performs very well even if the models themselves perform as well as random ones.

The process for AdaBoost looks something like:

1. Given inputs $(x_1, y1), ..., (x_m, y_m)$ where $x_i \in X, y_i \in \{-1, +1\}$, initialize their distributions $D_t(i) = \frac{1}{m}, i = 1, ..., m$ [1]
2. Find the classifier h_t which minimizes the error w.r.t. D_t.[2]
3. Calculate the weight classifier $\alpha_t = \frac{1}{2} ln \frac{1-\epsilon_t}{\epsilon t}$
4. Update our distribution [3]

$$D_{t+1}(i) = \frac{D_t \exp[-\alpha_t y_i h_t(x_i)]}{Z_t}$$

Once we have the final classifier $H(x)$, comprised of the sum of the modesl with the adjusted weights and distributions, we can calculate

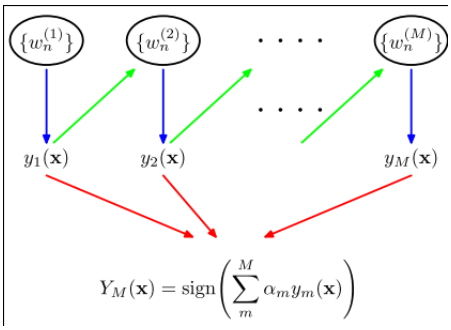

Figure 1: Each of the models contributes to how we pass the input set to the next model to be trained. At the end, we also use each model's results to calculate the final prediction.

[1] The distributions $D_t(i)$ mean the probability distribution of x_i's at time t

[2] We can basically take the min of

$$\epsilon_j = \sum_{i=1}^{m} D_t(i)[y_i \neq h_j(x_i)]$$

[3] Z_t is basically a value to normalize the distribution.

the **margin** of the classifier on a sample (x, y). The margin is given by [4]

$$yH(x)$$

The AdaBoost algorithm will try to maximize the margins of our variables by minimizing the **exponential loss** of the predictions.

$$loss_{exp}[H(x)] = E_{x,y}[e^{-yH(x)}]$$

[4] The margin tells us when we classified the input correctly, and our confidence with our classification.