Ensemble Prediction - Max Dama - 2009/11/08

In this note I'll explain how you can use ensembles of prediction algorithms. An ensemble of learners is a group that you make predictions by having them take a majority vote. Instead of just running one classifier, you train and run more and then treat each prediction as a single vote. The prediction with the most votes wins of course.

One fact that has surprised researchers is that a full ensemble can be more accurate than any of its constituents - just by combining them you get a free improvement without needing to come up with a more powerful non-linear model or anything complicated.

Consider this scenario, Let's start calling the prediction algorithms "experts". You're trying to predict whether to buy a stock, bond, commodity, or currency. So there are 4 possible target values and you want to pick the best one. Let's say that each one is equally likely with 25% probability, and that the best one to buy is the stock but you don't know that. Let's say that you're experts are accurate 33% of the time. This is worth something since $\frac{1}{3} > \frac{1}{4}$.

Look at what happens as you build up your ensemble of experts. The first one gives you a $\frac{1}{3} - \frac{1}{4} = \frac{1}{12}$ edge. And you know that when the expert says to buy something, you should follow that and you will be 33% accurate.

Now you have an ensemble of 2 experts. They will be in agreement and correct 33% of 33% of the time ie $\frac{1}{3} * \frac{1}{3} = \frac{1}{9} = 11\%$ of the time. This seems low. But how often are they both wrong when they're in agreement? There are three wrong buys and each has the same probability, so the total is $3 * \frac{2}{9} * \frac{2}{9} = \frac{4}{27} = 15\%$. Both these numbers are pretty small so it's hard to compare them to 33% vs 66% like we had with just 1 expert. Let's look at the odds ratio of being right, $\frac{\frac{1}{9}}{\frac{1}{9}+\frac{4}{27}} = \frac{3}{7} = 43\%$. This beats $\frac{\frac{1}{3}}{\frac{1}{3}+\frac{2}{3}} = \frac{1}{3} = 33\%$, although notice that in most cases the two experts will disagree and we won't buy anything that round (the probability they disagree is $2 * 3 * \frac{1}{3} * \frac{2}{9} + 6 * \frac{2}{9} * \frac{2}{9} = \frac{20}{27} = 74\%$ - see the table at the end - so about three-fourths of the time we don't buy anything, compared to buying something every round with just one expert.)

Now you get one more expert, so you have 3. Similarly to before, they correctly agree $\frac{1}{3} * \frac{1}{3} * \frac{1}{3} = \frac{1}{27} = 3.7\%$ and incorrectly agree $3 * \frac{2}{9} * \frac{2}{9} * \frac{2}{9} = \frac{8}{243} = 3.3\%$ of the time. Now our odds ratio is $\frac{\frac{1}{27}}{\frac{1}{27}+\frac{8}{243}} = \frac{9}{17} = 53\%$ - above 50% accurate!

The bottom line is that we've been able to greatly improve our prediction ability by simply combining a few predictors and having them vote. This is an extremely easy function to apply in practice.

Table of all possible predictions by two experts

	TT	FF			TF			FT			FF					
$\star stock\star$	12				1	1	1	2	2	2						
bond		12			2			1			1	1		2	2	
commodity			12			2			1		2		1	1		2
currency				12			2			1		2	2		1	1

More Math Section

We can see general formulas for the probability of the experts being right when they are in agreement, $p(\text{all T}) = {p_T}^n$ where n is the number of experts; and similarly for them all being wrong, $p(\text{all F}) = {p_F}^n$. So the accuracy of the unanimous ensemble is $\frac{{p_T}^n}{{p_T}^n + {p_F}^n}$. Notice that this is a specialization of the general formula, Bayes Rule, for condition probability - $P(\text{all T}|\text{all same}) = \frac{P(\text{all same}|\text{all T})P(\text{all T})}{P(\text{all same})} = \frac{P(\text{all same}|\text{all T})P(\text{all T})}{P(\text{all T})+P(\text{all F})}$. We've assumed the experts are independent throughout. If they're correlated/dependent, then the effective number of experts is scaled down. The number of unanimous signals you receive, ${p_T}^n + {p_F}^n$, falls exponentially with the number of experts n since p_T and p_F are less than 1, while the accuracy increases more and more slowly toward 100%. If you wait till the end of the world when every expert says to do the same thing, they'll probably be right, but it's not going to happen often (notice that we're not talking about TV experts, who always say the same thing are are also always wrong).

It would be interesting to look at how to carry out this kind of analysis with real valued predictions instead of discrete. You'd probably have to define a regret/loss function and show how that is expected to be lower if you only act on predictions where all experts predict sufficiently close to each other.

Fyi, some synonyms or generalizations of these ideas are called boosting, bagging, and stacking in the literature in case you want to check Google Scholar or videolectures.net for more.