On the Size and Membership of Voting Groups

David Steiner

Abstract

This paper provides guidance on how to form groups to make optimal decisions. Specifically, it covers how group decisions differ from individual decisions, and under what conditions it is worthwhile to add a new member to a voting group. A voting group in this context is any collection of entities (people or software) contributing equally to making a decision.

1 Introduction

Many are familiar with the concept of "the wisdom of the crowd", that is, the idea that a group of average individuals can outperform a single expert on a variety of tasks. Unfortunately, the literature is quite light on the mathematical explanation for this effect, as well as on explaining the required conditions for the effect to be most potent. This paper expands on the proof by Bishop [1] and adds commentary to make it suitable for a non-mathematical audience.

2 Error of an individual vs. error of a group

Each individual in a group, hereafter indexed m, will have some model of the world. Their (their model's) opinion (prediction, in machine learning literature) on a particular point of inquiry x (train/test example, in machine learning literature) can be written as

$$y_m(x) = h(x) + \epsilon_m(x)$$

Where $y_m(x)$ is member m's learned model of the world, $h(x)$ is the world as it is (in other words, true reality), and $\epsilon_m(x)$ is some error. x is some specific point to evaluate. For example, x could be "what will the price of oil be in California tomorrow?", $y_m(x)$ is model m's prediction for the price, $h(x)$ is what the price ended up being, and $\epsilon_m(x)$ is the difference between the predicted value and actual value.

In English, this equation is saying that "a model of the world = reality + some error". This can be re-written as

$$\epsilon_m(x) = y_m(x) - h(x)$$

Which in words is saying "your error is the difference between your model of the world and reality".

We can talk about the average (squared) error on all possible topics x

$$\mathbb{E}_x[(y_m(x) - h(x))^2] = \mathbb{E}_x[\epsilon_m(x)^2]$$

In words, this is the expected square of the error the model of the world is making over all possible inquiries.

Now, imagine that instead of one model of the world, you have M models. $y_m(x) \in \{y_1(x), y_2(x), ..., y_M(x)\}$ What is the average error across all the models taken individually?

$$E_{AVG} = \frac{1}{M}\sum_{m=1}^{M} \mathbb{E}_x[\epsilon_m(x)^2]$$

If we combine all the models into one instead of evaluating the error of each model individually, we can compare the average error of the individual models with the error of the combined model. We will call the combined model a "committee". The committee's decision is the average of each member's vote

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

And we can solve for the error of the committee:

$$\begin{aligned}
E_{committee} &= \mathbb{E}_x[\{(\frac{1}{M}\sum_{m=1}^{M} y_m(x)) - h(x)\}^2] \\
&= \mathbb{E}_x[\{\frac{1}{M}(\sum_{m=1}^{M} y_m(x) - Mh(x))\}^2] \\
&= \mathbb{E}_x[\{\frac{1}{M}(\sum_{m=1}^{M} y_m(x) - \sum_{m=1}^{M} h(x))\}^2] \\
&= \mathbb{E}_x[\{\frac{1}{M}\sum_{m=1}^{M}(y_m(x) - h(x))\}^2] \\
&= \mathbb{E}_x[\{\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(x)\}^2]
\end{aligned}$$

Notice that the real-world never changes ($h(x)$ is the same for all). If the errors between each model m are uncorrelated and each has a mean of zero, something interesting happens. Under these assumptions

$$
\begin{aligned}
E_{committee} &= \mathbb{E}_x[\{\frac{1}{M}\sum_{m=1}^{M}\epsilon_m(x)\}^2] && (1)\\
&= \frac{1}{M^2}\,\mathbb{E}_x[\sum_{m=1}^{M}\epsilon_m(x)^2 + \sum_{\substack{\forall i\in(1..M),\\ \forall j\in(1..M)\\ i<j}} 2\epsilon_i(x)\epsilon_j(x)] && (2)\\
&= \frac{1}{M^2}(\mathbb{E}_x[\sum_{m=1}^{M}\epsilon_m(x)^2] + \mathbb{E}_x[\sum_{\substack{\forall i\in(1..M),\\ \forall j\in(1..M)\\ i<j}} 2\epsilon_i(x)\epsilon_j(x)]) \quad \text{(Expectation of sums is the sum of expectations)} && (3)\\
&= \frac{1}{M^2}(\sum_{m=1}^{M}\mathbb{E}_x[\epsilon_m(x)^2] + \sum_{\substack{\forall i\in(1..M),\\ \forall j\in(1..M)\\ i<j}} 2\,\mathbb{E}_x[\epsilon_i(x)\epsilon_j(x)]) \quad \text{(Expectation of sums is the sum of expectations)} && (4)\\
&= \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}_x[\epsilon_m(x)^2] \quad \text{(See below)} && (5)\\
&= \frac{1}{M}E_{AVG} && (6)
\end{aligned}
$$

Step 5 comes from the two assumptions listed above. Pull out the constant 2 since $\mathbb{E}[cX] = c\,\mathbb{E}[X]$ when c is a constant. Then by the definition of uncorrelated, $\mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$, and by the assumption of the mean of each model's errors being zero $\mathbb{E}[X] = 0$. Therefore the $\sum 2\,\mathbb{E}_x[\epsilon_i(x)\epsilon_j(x)]$ term disappears since $\sum 2\,\mathbb{E}_x[\epsilon_i(x)\epsilon_j(x)] = \sum 2\,\mathbb{E}_x[\epsilon_i(x)]\,\mathbb{E}_x[\epsilon_j(x)] = \sum 2\times 0\times 0 = 0$.

This tells us that we can endlessly reduce our error simply by adding more members to our committee *if we choose our new members carefully.* So with $M = 2$ we cut our average error in half. With $M = 100$, the average error of the committee (which simply averages each member's guess) across all possible topics x will be just 1% of the average of each of their errors taken individually.

The reason this result is interesting, and not obvious, is because it places absolutely no requirement on each model $y_m(x)$'s quality. That is to say, even if each model is very weak on its own, as long as it is uncorrelated with the other models, and makes errors equally in both directions, the average of the weak models is very strong compared to what they were individually.

3 Criteria for adding new members

Obviously, most people would reject the assumptions above as being impractical, and it is true that it is difficult to make models that are truly uncorrelated and have mean errors centered at zero.

The qualitative defense of these assumptions is that averaging different models is used widely in practice by data scientists and by societies across the world with generally good results.

The quantitative defense is that because of the smoothness of the limits of the covariance and expectation functions, we don't actually have to be completely uncorrelated, or completely mean zero, to reap the benefits. $\operatorname{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$, so $\mathbb{E}[XY] = \operatorname{Cov}[X,Y] + \mathbb{E}[X]\,\mathbb{E}[Y]$. Therefore,

$$\lim_{\operatorname{Cov}[X,Y]\to 0} \mathbb{E}[XY] = \mathbb{E}[X]\,\mathbb{E}[Y]$$

and

$$\lim_{\mathbb{E}[X]\to 0 \text{ or } \mathbb{E}[Y]\to 0} \mathbb{E}[X]\,\mathbb{E}[Y] = 0$$

This shows that every tiny step we take towards reducing the covariance of our models, and every step we take to center the model errors at zero, is helpful. In other words, the function is convex, so there will *never* be a case where reducing the correlation and getting mean error closer to zero hurts us. That being proven, we now have a precise way to judge whether adding a new member to our committee will bring a reduction in error.

What makes a new member worth adding to the committee? Obviously, we want the error after adding the member to be better than the current committee error:

$$E_{new} < E_{current}$$

Plugging in what we solved for above for E_{new} and $E_{current}$ gives us:

$$\frac{1}{(M+1)^2}(\sum_{m=1}^{M+1} \mathbb{E}_x[\epsilon_m(x)^2] + \sum_{\substack{\forall i\in(1..M+1),\\ \forall j\in(1..M+1)\\ i<j}} 2\,\mathbb{E}_x[\epsilon_i(x)\epsilon_j(x)]) < \frac{1}{M^2}(\sum_{m=1}^{M} \mathbb{E}_x[\epsilon_m(x)^2] + \sum_{\substack{\forall i\in(1..M),\\ \forall j\in(1..M)\\ i<j}} 2\,\mathbb{E}_x[\epsilon_i(x)\epsilon_j(x)])$$

Now extract the terms containing ϵ_{m+1} to make it clear exactly what changed.

$$\frac{1}{(M+1)^2}(\sum_{m=1}^{M} \mathbb{E}_x[\epsilon_m(x)^2] + \mathbb{E}_x[\epsilon_{m+1}(x)^2] + \sum_{\substack{\forall i\in(1..M),\\ \forall j\in(1..M)\\ i<j}} 2\,\mathbb{E}_x[\epsilon_i(x)\epsilon_j(x)] + \sum_{m=1}^{M} 2\,\mathbb{E}_x[\epsilon_i(x)\epsilon_{m+1}(x)])$$
$$< \frac{1}{M^2}(\sum_{m=1}^{M} \mathbb{E}_x[\epsilon_m(x)^2] + \sum_{\substack{\forall i\in(1..M),\\ \forall j\in(1..M)\\ i<j}} 2\,\mathbb{E}_x[\epsilon_i(x)\epsilon_j(x)])$$

Multiply both sides by M^2

$$\frac{M^2}{(M+1)^2}(\sum_{m=1}^{M} \mathbb{E}_x[\epsilon_m(x)^2] + \mathbb{E}_x[\epsilon_{m+1}(x)^2] + \sum_{\substack{\forall i\in(1..M),\\ \forall j\in(1..M)\\ i<j}} 2\,\mathbb{E}_x[\epsilon_i(x)\epsilon_j(x)] + \sum_{m=1}^{M} 2\,\mathbb{E}_x[\epsilon_i(x)\epsilon_{m+1}(x)])$$
$$< \sum_{m=1}^{M} \mathbb{E}_x[\epsilon_m(x)^2] + \sum_{\substack{\forall i\in(1..M),\\ \forall j\in(1..M)\\ i<j}} 2\,\mathbb{E}_x[\epsilon_i(x)\epsilon_j(x)]$$

Now let's use this inequality to examine two scenarios: $M = 1 \to 2$ and $M = \infty \to \infty + 1$. For $M = 1 \to 2$ we have

$$\frac{1}{4}(\mathbb{E}_x[\epsilon_1(x)^2] + \mathbb{E}_x[\epsilon_2(x)^2] + 2\,\mathbb{E}_x[\epsilon_1(x)\epsilon_2(x)]) < \mathbb{E}_x[\epsilon_1(x)^2]$$
$$\frac{1}{4}\,\mathbb{E}_x[\epsilon_2(x)^2] + \frac{1}{2}\,\mathbb{E}_x[\epsilon_1(x)\epsilon_2(x)] < \frac{3}{4}\,\mathbb{E}_x[\epsilon_1(x)^2]$$

This is the inequality that must hold to see if it is worth adding model 2 to our committee. This is saying that the sum $\frac{1}{4}$ of model 2's squared error plus half of the expectation of $\epsilon_1(x)\epsilon_2(x)$ must be less than $\frac{3}{4}$ of model 1's squared error by itself. The pertinent thing here is that model 2 doesn't have to have better error than model 1 to help, nor do model 1 and 2 need to be *perfectly* uncorrelated with mean zero. Imagine model 2 has the same squared error as model 1. Then we have

$$\begin{aligned}
\frac{1}{2}\,\mathbb{E}_x[\epsilon_1(x)\epsilon_2(x)] &< \frac{1}{2}\,\mathbb{E}_x[\epsilon_1(x)^2] \\
\mathbb{E}_x[\epsilon_1(x)\epsilon_2(x)] &< \mathbb{E}_x[\epsilon_1(x)^2] && \text{(Multiply by 2)} \\
\mathrm{Cov}[\epsilon_1(x),\epsilon_2(x)] + \mathbb{E}_x[\epsilon_1(x)]\,\mathbb{E}_x[\epsilon_2(x)] &< \mathbb{E}_x[\epsilon_1(x)^2]
\end{aligned}$$

The last step comes from the definition of covariance: $\mathrm{Cov}[X,Y] = \mathbb{E}[XY] - \mathbb{E}[X]\,\mathbb{E}[Y]$, so $\mathbb{E}[XY] = \mathrm{Cov}[X,Y] + \mathbb{E}[X]\,\mathbb{E}[Y]$.
For the case $M = \infty \to \infty + 1$ we have $\lim_{M\to\infty} \frac{M^2}{(M+1)^2} = 1$ so

$$\begin{aligned}
&\sum_{m=1}^{M} \mathbb{E}_x[\epsilon_m(x)^2] + \mathbb{E}_x[\epsilon_{m+1}(x)^2] + \sum_{\substack{\forall i\in(1..M),\\ \forall j\in(1..M)\\ i<j}} 2\,\mathbb{E}_x[\epsilon_i(x)\epsilon_j(x)] + \sum_{m=1}^{M} 2\,\mathbb{E}_x[\epsilon_i(x)\epsilon_{m+1}(x)] \\
&< \sum_{m=1}^{M} \mathbb{E}_x[\epsilon_m(x)^2] + \sum_{\substack{\forall i\in(1..M),\\ \forall j\in(1..M)\\ i<j}} 2\,\mathbb{E}_x[\epsilon_i(x)\epsilon_j(x)]
\end{aligned}$$

The original error terms cancel out giving us the following criterion for adding new members:

$$\mathbb{E}_x[\epsilon_{m+1}(x)^2] + \sum_{m=1}^{M} 2\,\mathbb{E}_x[\epsilon_i(x)\epsilon_{m+1}(x)] < 0$$

Since error will always be ≥ 0, this condition is impossible to meet. In practice we will never have an infinite amount of models, but it shows that as $M \to \infty$, it becomes harder and harder to find new models that will result in any improvements, and the improvements they bring will be smaller and smaller, since they are being averaged out among so many other models.

4 Conclusion

The math supports the ideal that diversity of opinion is valuable when making decisions. Liberties, such as freedom of thought, are thus especially important in democratic societies, as a means to reduce the correlation among voters.

The criteria does, however, go against the intuitive idea that preserving or improving the intelligence of the average voter is the easiest way to improve the performance of a voting group. Reducing individual error, say, through better education, may not be as valuable as reducing the correlation, and when education is deployed on a large-scale (i.e. as part of the society-wide curriculum) it should be focused on teaching things of which there is little room for doubt, or on teaching balanced, open-minded ways of thinking, in order to encourage people's errors to average out towards the center. Improper education is especially dangerous as it increases correlation among the members in a group without a concomitant decrease in error.

The results demonstrated above also help inform the choice of government (in terms of how big the electorate should be). Monarchy is hard to justify as it should not be difficult to improve decision-making by finding a co-ruler who is uncorrelated with the first, even if they are slightly less intelligent. However, there are clear diminishing returns as the size of the electorate is increased. Keeping votes uncorrelated depends not just on the amount of liberty in a society but on the depth and breadth of the problem space, as it is impractical to have many uncorrelated models in a narrow domain.

If only a few types of topics need to be voted on, a diverse group of ten to one hundred people should be sufficient. If the range of issues to be decided spans the entirety of human knowledge, up to ten thousand could perhaps be justified, but after that the benefits of enlarging the electorate are more philosophical (appealing to the fairness, justice, and ethicality of universal suffrage) than mathematical (actually resulting in better decisions).

References

[1] C. M. Bishop. Pattern recognition and machine learning, Chapter 14.2. Springer, 2006.

Appendices

A Squared error vs. error

What is the difference between $\mathbb{E}_x[\epsilon_m(x)^2]$ and $\mathbb{E}_x[\epsilon_m(x)]$? Isn't it cheating to assume $\mathbb{E}_x[\epsilon_m(x)] \approx 0$ since that looks like we aren't making any errors?

$\mathbb{E}_x[\epsilon_m(x)] \approx 0$ is not actually saying we aren't making errors. Consider trying to predict the stock price for a stock tomorrow and the next day. If tomorrow you were off by +$100 and the next day you were off by -$100, then your average error was zero, but you obviously made mistakes. In this example, $\mathbb{E}_x[\epsilon_m(x)] = \frac{(100)+(-100)}{2} = 0$. $\mathbb{E}_x[\epsilon_m(x)^2]$ better captures the fact that we made mistakes. In this case, $\mathbb{E}_x[\epsilon_m(x)^2] = \frac{(100)^2+(-100)^2}{2} = 10000$.

So when we made the assumption that $\mathbb{E}_x[\epsilon_m(x)] \approx 0$, we are not making any statement on the absolute quality of the models, but rather are saying the model is generally making mistakes of equal magnitude in both directions.

One last question is why not use $\mathbb{E}_x[|\epsilon_m(x)|]$ instead of $\mathbb{E}_x[\epsilon_m(x)^2]$, since it also deals with the issue of positive mistakes canceling out negative mistakes? $\mathbb{E}_x[|\epsilon_m(x)|]$ would also work, but $\mathbb{E}_x[\epsilon_m(x)^2]$ is more mathematically convenient since x^2 is differentiable at all points whereas $|x|$ is not. The other difference that can be useful is that $\epsilon_m(x)^2$ punishes large mistakes much more than small mistakes, whereas $|\epsilon_m(x)^2|$ punishes all mistakes equally.