ESTIMATING THE ACCURACY OF A BAGGED ENSEMBLE

Eugene Pinsky and Siddhant Shah

Department of Computer Science, Metropolitan College, Boston University, Boston, MA

ABSTRACT

In ensemble machine learning, we combine the decisions of weak learners to derive a decision that is, hopefully, better than the individual ones. The combination of these learners can be aggregated by a majority vote or simple averaging, or it can be more complicated and involve multiple steps such as in boosting. In this paper, we consider the question of predicting the accuracy of an ensemble created with bagging for a given number of weak learners. We achieve a low relative error on our predictions and can make this prediction in a shorter time, as compared to training the ensemble over various sizes.

KEYWORDS

Ensemble, Bagging, Weak Learners, Poisson Distribution, Normal Distribution

1. INTRODUCTION

The combination of multiple models in an ensemble often results in better predictive performance than relying on individual models (for a survey, see [1]). Ensembles use models with high variance and low bias, and use using this diversity to reduce the effective variance and minimize bias, and reduce overfitting.

Despite their established advantages, using an ensemble has a higher cost due to the increased number of learners being used. Furthermore, it is difficult to predict the accuracy of an ensemble for a given size, which leads to trial and error over increasing sizes till the desired accuracy is attained. Thus, it is essential to comprehend and measure when an ensemble is “large enough” to create effective and scalable predictive algorithms.

In 2019, Lopes [2] used a bootstrap-based method to estimate the variance of randomized ensembles, such as bagging and random forests. The key contribution is a mechanism to approximate the error distribution of finite ensembles and evaluate how closely it matches the accuracy of an infinite ensemble. Using efficient extrapolation techniques, they provide a practical guide to determine when ensemble fluctuations have become negligible, thus ensuring algorithmic convergence.

Slivinski and Snyder [3] focus on estimating the size of the ensemble within particle filters, a method widely used for data assimilation in high-dimensional dynamical systems. Their work addresses the “curse of dimensionality,” in which ensemble size requirements grow exponentially with system dimension, making particle filters computationally prohibitive in many applications. They derive asymptotic relations that link ensemble size to system properties, such as effective dimensionality and observational error covariance. By using simpler data assimilation techniques to estimate key parameters, they provide a practical framework to predict when a particle filter ensemble will achieve satisfactory performance without exhaustive experimentation.

DOI:10.5121/mlaij.2025.12106

In 2024, Christiansen [4] proposed simple formulas to predict the ensemble mean squared error, based on small initial sub-samples. This work derives analytical relations to estimate the trade-offs between the size of the ensemble and its accuracy in climate modeling. The approach emphasizes practicality, enabling users to assess the expected benefits of increasing ensemble size without the need to generate prohibitively large ensembles. Such tools are particularly valuable in computationally intensive fields where the cost of training and running large ensembles is high.

Although these studies advance our understanding of ensemble behavior in specific use cases, this paper proposes a simpler approach to predicting the ensemble accuracy and thereby, the optimal ensemble size, where the ensemble uses simple majority voting. Unlike prior models, our framework focuses on probabilistic estimation of weak learners and of the ensemble. This gives an interpretation of the ensemble as a probability distribution, allowing us to better understand its behavior.

We start by modelling the ensemble as an aggregation of weak learners making decisions using majority voting, where each weak learner is makes an independent decision with a given probability of correctness and analyze its behavior by modeling its accuracy as a probability distribution. Next, we relax the assumption of equal probabilities to generalize our analysis and better approximate real-world scenarios. Finally, we evaluatethe effectiveness of our predictions against actual ensemble performance using classifiers trained on a sample dataset.

2. ENSEMBLE WITH MAJORITY VOTING

Our general setting is as follows: We assume that we have weak learners, each learner making the right decision with some probability, and we make our final classification based on the majority decision of these weak classifiers.

Formally, we describe the model as follows. We have $N = 2n + 1$ weak learners. For majority voting, we need an odd number of such simple classifiers to ensure that there are no ties in the vote. For each weak learner i, let $p_i < 1$ denote the probability that this classifier makes the correct decision. We can then model this using a simple Bernoulli distribution [5].
Let the random variable X_i denote the decision of the i^{th}weak learner as follows:

$$\mathrm{X_i} = \begin{cases} 1, & if\ learner\ i\ is\ correct \\ 0, & otherwise \end{cases}$$

The probability distribution $P(\cdot)$ for each weak learner is given by a simple rule:

$$\mathrm{X_i} = \begin{cases} p_i, & if\ x\ =\ 1 \\ q_i\ =\ 1 -\ p_i, & otherwise \end{cases} \ldots (1)$$

Define the random variable $X = X_1 + X_2 + \cdots + X_N$. This random variable takes integer values from 0 to N. In particular, $X\ >\ n\ \equiv\ X\ \geq\ n\ +\ 1$ denotes the event when the ensemble makes the right decision.

We would like to investigate when an ensemble gives us a higher probability of success than any of the individual learners (even the best one with the highest p_i). In other words, we would like to investigate when

$$P(X \geq n + 1) > \max_i(p_i)$$

To that end, we proceed as follows:We investigate the how the accuracy of the ensemble varies with the individual probabilities of the learners. We start with a model where all weak learners are independent and make decisions with the same probability. We will then extend this to a more general case where these probabilities are not equal.

3. Independent Weak Learners with Same Accuracy

In this case, the ensemble represented by the random variable Xfollows a binomial distribution [5]. We have $N = 2n + 1$ independent learners. The probability that exactly $k$learners make the correct decision is a binomial distribution with

$$P(X = k) = \binom{N}{k} p^k q^{N-k}$$

where $p$is the probability that each weak learner makes the correct decision, $P(X_i = 0) = p$for all i.

We will find it convenient to introduce the following notation

$$F^*(k, N, p) = P(X > k) = \sum_{i=k+1}^{N} \binom{N}{i} p^i (1-p)^{N-i}$$

With this definition, the probability of a correct decision is $F^*(n, 2n + 1, p)$. In other words, it is the probability that more than $n$learners made the correct decision.

Example 1. We consider an ensemble$N = 3$ learners. The individual probabilities and hence, the probabilities of the ensemble havingat least $k + 1$ learners make the correct decision can be written explicitly and are summarized in the table below:

Table 1: Individual probabilities for the number of correct decisions in an ensemble with 3 learners

k	$P(X = k)$	$F^*(k, 3, p)$
0	q^3	$1 - q^3$
1	$3pq^2$	$1 - (\mathrm{q}^3 + 3p\mathrm{q}^2)$
2	$3p^2q$	$1 - (\mathrm{q}^3 + 3pq^2 + 3p^2q)$
3	q^3	0

For example, the probability that exactly one learner makes the correct decision is

$$P(X = 1) = 3p(1-p)^2$$

With majority voting, the probability of correct prediction is the probability that no more than one learner makes a wrong decision. Therefore, the probability of the right decision by the ensemble is:

$$F^*(1{,}3, p) = P(X = 2) + P(X = 3) = 3p^2(1-p) + p^3$$

It is obvious to think that adding more learners would make the ensemble better, but this may not necessarily be the case. Thus, the following question arises: *How does the accuracy of the ensemble vary with that of the individual weak learners? Specifically, are there values of* p *for which it is better to not use ensemble voting?*

First, we investigate graphically using Figure 1.We see that if the probability of success $p < 0.5$, then using an ensemble gives you a worse result than using an individual learner.

We can also show it algebraically. If we were to use a learner, then the probability of the right decision by such a learner is p. Using ensemble voting, the probability of acorrect decision from the ensemble is $F^*(1,3, p)$. Therefore, we are looking for p for which,

$$p^3 + 3p^2(1 - p) > p$$

Or, equivalently,

$$p^2 + 3p(1 - p) > 1$$

After some elementary algebra, we obtain $p > 0.5$. This result means that for the case of 3 learners, if the probability of success is less than 0.5 then using ensemble voting would result in a higher error rate. This result generalizes to the general case of any N.

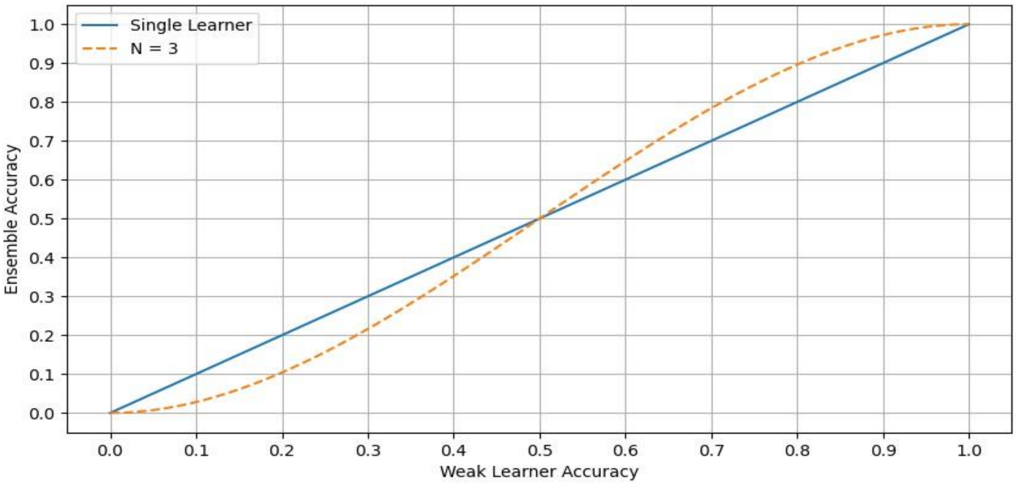

Fig.1: Variation in the accuracy of a 3-learner ensemble and a 1-learner ensemble with change to the individual weak learner accuracy

Example 2. Let us see how accuracy increases with the number of learners. In Figure 2, we plot the accuracy of the ensembles of sizes $N = 3$, 5, and 7 against the probability of success of each individual learner.

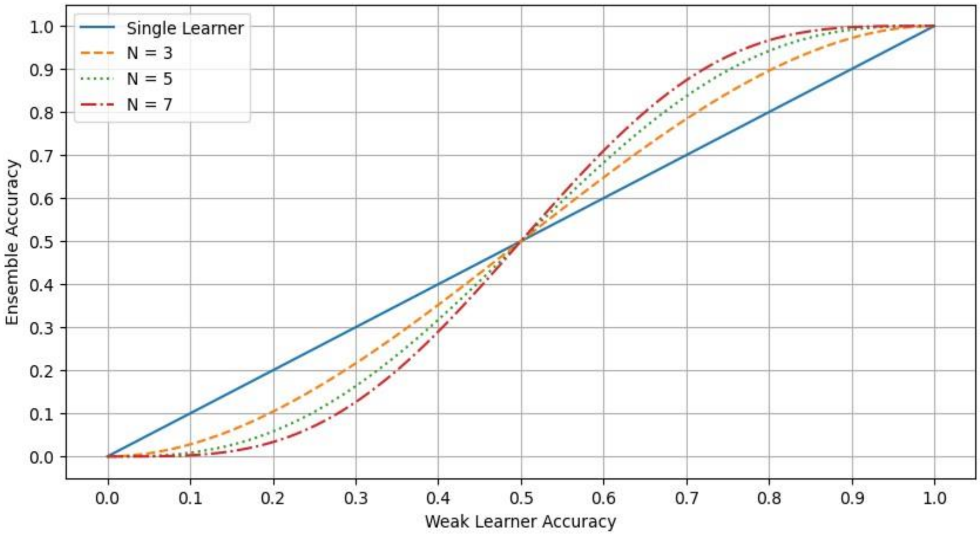

Fig.2: Variation in the accuracies of ensembles with 1, 3, 5 and 7 learners with change to the individual weak learner accuracy

Similarly to the previous case, voting by ensemble gives a higher error than using an individual learner as long as the probability of success p >0.5. In this range, we can see that for any value of the probability of success, the accuracy of the ensemble increases as we increase the number of learners. Furthermore, for the same number of learners, an ensemble where each learner has a higher probability of success, has a higher probability of success. In fact, when each individual learner is almost always correct, the ensemble's accuracy approaches 1.

Our observations from the examples have been formally stated in [6], as Condorcet's Jury Theorem, which states (translated to English):

- If $p$is greater than $1/2$ (each voter is more likely to vote correctly), then adding more voters increases the probability that the majority decision is correct. In the limit, the probability that the majority votes correctly tend to1 as the number of voters increases.
- On the other hand, if $p$is less than $1/2$ (each voter is more likely to vote incorrectly), then adding more voters makes things worse: the optimal jury consists of a single voter.

3.1. Computing Ensemble Accuracy

Now that we have some understanding of the behavior of the ensemble's accuracy, we need to formally derive it.
For the ensemble with $N = 2n + 1$ independent weak learners (each with a probability of a correct decision p), the probability of a correct decision is $P(X > n)$.
Using the Binomial Distribution, we get

$$F(N) = P\left(X > \left\lfloor\frac{N}{2}\right\rfloor\right) = 1 - P(X \leq n) = 1 - \sum_{i=0}^{n} \binom{N}{i} p^i (1-p)^{N-i} \dots (2)$$

Computing the accuracy of an ensemble directly takes time proportional to $O(N^2)$, because of the factorial. So, we try to find an approximation that takes asymptotically less time to compute.
To that end, we consider two standard approximations to the binomial distribution:

1. Approximation by a Poisson Distribution, when $N \cdot \min(p, 1-p)$ is small, usually less than 15
2. Approximation by a Normal Distribution, when $N \cdot \min(p, 1-p)$ is large, usually greater than 15

Approximation by Poisson Distribution

The binomial distribution can be approximated by a Poisson distribution with $\lambda = Np$.

$$\lim_{N\to\infty} P\,(X = k) = \lim_{N\to\infty} \binom{N}{i} p^i (1-p)^{N-i} = \frac{\lambda^k}{k!} e^{-\lambda}$$

This can be used when $N \cdot \min(p, 1-p) \leq 15$. For larger values, we need to use the Normal Approximation (See Section 3.1).

Theorem 1. The cumulative distribution function of the Poisson distribution (See [7]) is

$$F_\lambda(k) = \frac{\Gamma(\lfloor k+1 \rfloor, \lambda)}{\lfloor k \rfloor!} = \frac{\Gamma(k+1, Np)}{k!}$$

where $\Gamma(s, x)$ is the upper incomplete gamma function

$$\Gamma(s, x) = \int_x^\infty t^{s-1}\, e^{-t}\, dt$$

From Theorem 1 and Equation 2, we have that the probability of an incorrect decision in an ensemble with $N = (2n+1)$ weak learners are

$$\mathrm{F(N)} \approx 1 - \mathrm{F}_\lambda\left(\left\lfloor \frac{N}{2} \right\rfloor\right) = 1 - F_\lambda(n) = 1 - \frac{\Gamma(\mathrm{n} + 1, (2\mathrm{n}+1)\mathrm{p})}{\mathrm{n}!}$$

For large N, we have

$$F(N) \approx 1 - \frac{\Gamma(n+1, 2p(n+1) - p)}{n!} \approx 1 - \frac{\Gamma\big(n+1, 2p(n+1)\big)}{n!} = f(N) \qquad \dots(3)$$

While this approximation is asymptotically better than N^2, we can do better by expressing the upper incomplete gamma function as a probability distribution.

Theorem 2. We can estimate $f(N)$ and thus $F(N)$ as the following

$$F(N) \approx f(N) \approx f_1(N) = \Phi\left((2p-1)\sqrt{\frac{N+1}{2}}\right)$$

where Φ is the cumulative distribution function of the standard normal.

Proof. We have that

$$f(N) = 1 - \frac{\Gamma\big(n+1, 2p(n+1)\big)}{n!} = \frac{\Gamma(n+1) - \Gamma\big(n+1, 2p(n+1)\big)}{\Gamma(n+1)} = \frac{\gamma\big(n+1, 2p(n+1)\big)}{\Gamma(n+1)}$$

where $\gamma(s,x)$ is the lower incomplete gamma function

$$\gamma(s,x) = \int_0^x t^{s-1}\, e^{-t}\, dt$$

Notice that this is a special case of the regularized gamma function, which is the cumulative distribution function of a gamma random variable [8] with unit scale and shape parameter $n+1$. Thus,

$$f(N) = F_\gamma(2p(n+1); n+1,1) \qquad \ldots (4)$$

where $F_\gamma(x;k,\theta)$ is the CDF of a gamma random variable with shape parameter $k$and scale parameter θ.

The mean and variance of gamma distribution are given by

$$\mu = k\theta \qquad \sigma^2 = k\theta^2$$

In our case, we have that $k = n+1$ and $\theta = 1$. We get

$$\mu = n+1 \qquad \sigma^2 = n+1$$

Let $Y \sim Gamma(n+1,1)$. This gives us the following value for $f_1(N)$

$$f(N) = \frac{\gamma\big(n+1, 2p(n+1)\big)}{\Gamma(n+1)} = P\big(Y \le 2p(n+1)\big)$$

As the value of $n$increases (or equivalently, the value of Nincreases), a gamma distribution can be approximated by a normal distribution with the same mean and variance, using the Central Limit Theorem, i.e.,

$$\lim_{n\to\infty} Gamma(k,\theta) \mapsto Normal\big(k\theta, k\theta^2\big)$$
$$\lim_{n\to\infty} Gamma\,(n+1,1) \mapsto Normal(n+1, n+1)$$

where $Normal(\mu,\sigma^2)$ is a normal distribution with mean μ and variance σ^2.

Using this approximation, $Y \sim Normal(n+1, n+1)$. Thus,

$$f(N) \approx P\big(Y \le 2p(n+1)\big) = \Phi\left(\frac{2p(n+1)-\mu}{\sigma}\right) = \Phi\left((2p-1)\sqrt{(n+1)}\right)$$

where $\Phi$is the cumulative distribution function of the standard normal. Finally, we know that $N = 2n + 1$, this gives us that

$$f(N) \approx \Phi\left((2p-1)\sqrt{\frac{N+1}{2}}\right)$$

This concludes the proof.

This approximation gives us an estimated accuracy in constant time as we only need to look-up the value of the standard normal distribution.

Approximation by Normal Distribution

This approximation can be used when $N \cdot \min(p, 1-p) > 15$. For smaller values, we need to use the Poisson Approximation (SeeSection 3.1)

Theorem 3. We can estimate $F(N)$ as

$$F(N) \approx \Phi\left(\frac{N(2p-1)+1}{2\sqrt{Np(1-p)}}\right) \dots (5)$$

where Φ is the cumulative distribution function of the standard normal.

Proof. Because $X_i$are independent, the binomial distribution can be approximated by a normal distribution with the same mean (μ) and variance (σ^2), using the Central Limit Theorem.

$$\mu = \mathrm{E}[X] = \sum_{i=1}^{N} E[X_i] = Np$$

$$\sigma^2 = Var[X] = \sum_{i=1}^{N} Var[X_i] = Np(1-p)$$

Thus, from Equation 2 can be approximated as

$$F(N) = 1 - P(X \le n) \approx 1 - F_N(n; \mu, \sigma^2) = 1 - \Phi\left(\frac{n-\mu}{\sigma}\right) = 1 - \Phi\left(\frac{N(1-2p)-1}{2\sqrt{Np(1-p)}}\right)$$

where $F_N(x; \mu, \sigma^2)$ is the cumulative distribution function of a normally distributed random variable with mean $\mu$and variance σ^2.

For the standard normal, we know that $F_N(x; \mu, \sigma^2) = 1 - F_N(-x; \mu, \sigma^2)$. From this, we can rewrite the previous result as

$$F(N) \approx 1 - \Phi\left(\frac{N(1-2p)-1}{2\sqrt{Np(1-p)}}\right) = \Phi\left(\frac{N(2p-1)+1}{2\sqrt{Np(1-p)}}\right)$$

This concludes the proof.

Evaluation of the Approximations

Example 3. We now proceed to evaluate the performance of the Poisson and Normal approximation.

During the evaluation, we noticed that these approximations have errors that move in opposite directions. Thus, we also evaluate the performance of the average of these approximations with the hypothesis that the errors induced by either approximation will be balanced by the other, giving us a consistent result.

The performance of these three approximations can be observed in Figure 3 and the actual values can be seen in the Appendix.

We see that our approximations are close to the actual predicted values from the binomial distribution.Over the entire dataset, we get that the Average approximation gives us the best results, as seen in Table 2.

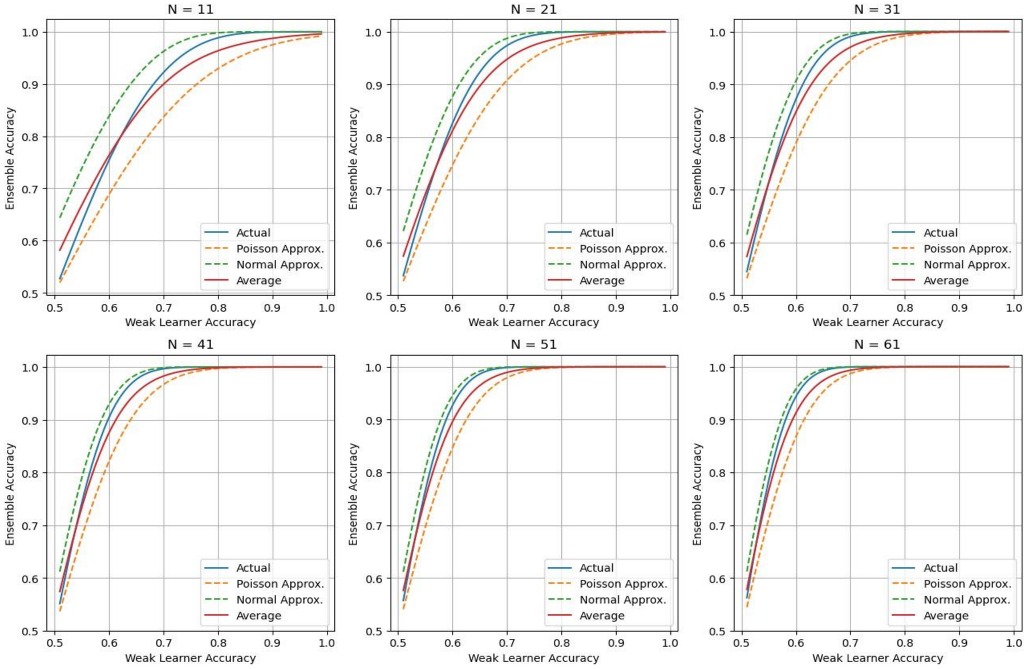

Fig.3: Comparison of Actual Ensemble Accuracy (Binomial) versus the Predicted Ensemble Accuracy (Poison, Normal and Average)for various ensemble sizes, for individual weak learner accuracy varying from 0.5 to 1.

Table 2: Mean and Standard Deviation of relative errors and relative absolute errors using the Poisson, Normal and Average approximations.

	Poisson		Normal		Average	
	RE (%)	RAE (%)	RE (%)	RAE (%)	RE (%)	RAE (%)
Mean	-5.01	5.01	4.06	4.06	-0.47	2.22
SD	3.37	3.37	5.04	5.04	2.85	1.83

As predicted, the error of each approximation outside of its domainsisnegated by the other approximation, giving us a better result. Thus, while the Poisson and Normal Approximations are individually better in their specific regions, if we were to use one approximation over the entire space, we would use the average of the two.

Therefore, we can approximate the accuracy of an ensemble with $N = 2n + 1$ independent weak learners (each with a probability of a correct decision p), as follows:

$$F(N) \approx f_1(N) = \frac{1}{2}\left(\Phi\left(\left(2\mu_p - 1\right)\sqrt{\frac{N+1}{2}}\right) + \Phi\left(\frac{N\left(2\mu_p - 1\right) + 1}{2\sqrt{N\mu_p\left(1 - \mu_p\right)}}\right)\right)$$

4. INDEPENDENT WEAK LEARNERS WITH DIFFERENT ACCURACY

After investigating the simple ensemble with identical weak learners, we extend our analysis to account for weak learners with varying accuracy levels. We derive expressions for estimating ensemble accuracy in this more general setting and discuss when traditional Normal or Poisson approximations remain valid.

We now revisit the model as defined in Section 2. The sum of Bernoulli random variables $X = X_1 + X_2 + \cdots X_N$with different probabilities is not a simple distribution unless all $p_i$are equal. If $p_i$vary, the resulting distribution is called *Poisson's Binomial Distribution*[9].
The probability mass function (PMF) for Poisson's Binomial Distribution is:

$$P(X = k) = \sum_{A \subseteq \{1,2,\ldots,n\}, |A| = k} \left(\prod_{i \in A} p_i \prod_{j \notin A} (1 - p_j) \right) \quad \ldots (6)$$

Since $X$is the sum of Nindependent Bernoulli distributed variables, its mean and variance will simply be sums of the mean and variance of the Bernoulli distributions:

$$\mu_X = \sum_{i=1}^{N} p_i \, \sigma_X^2 = \sum_{i=1}^{N} (1 - p_i) p_i$$

In practice, we often do not know the accuracy of each X_ia priori, as in the case of the Random Forrest Ensemble, as these classifiers are generated randomly. Thus, we assume that p_i are generated by a distribution with mean $\mu_p$and variance σ_p^2.

$$E[\mu_X] = \sum_{i=1}^{N} E\,[p_i] = \sum_{i=1}^{N} E\left[\mu_p\right] = \mu_p N \qquad \ldots (7)$$

$$E[\sigma_X^2] = \sum_{i=1}^{N} E[(1 - p_i) p_i] = \sum_{i=1}^{N} E[p_i] - E\left[p_i^2\right] = \left(\mu_p - \sigma_p^2 - \mu_p^2\right) N \quad \ldots (8)$$

4.1. Computing Ensemble Accuracy

From Equation 6, we get that the CDF for the Poisson's Binomial Distribution is

$$P(X \le x) = \sum_{k=0}^{x} \sum_{A \subseteq \{1,2,\ldots,n\}, |A| = k} \left(\prod_{i \in A} p_i \prod_{j \notin A} (1 - p_j) \right)$$

Thus, we get that the probability that the model makes the correct decision is

$$F(N) = P\left(X > \frac{N}{2}\right) = \sum_{k=n+1}^{N} \sum_{A \subseteq \{1,2,\ldots,n\}, |A| = k} \left(\prod_{i \in A} p_i \prod_{j \notin A} (1 - p_j) \right) \ldots (9)$$

This probability distribution is even more complex to calculate than that of the Binomial Distribution. However, under the assumption of independence of the weak learner, we get some useful properties for this distribution that help us approximate this computationally-taxing distribution.

To estimate $F(N)$, we consider three approximations:

1. Approximation by a Binomial Distribution, when the classifiers have almost equal probabilities of success, i.e. $\sigma_X^2 \approx 0$.
2. Approximation by a Poisson Distribution, when the probability of each classifier is close to 0, i.e. $\mu_p \approx 0, \sigma_X \approx 0$.
3. Approximation by a Normal Distribution, in all other cases.

Approximation by Binomial Distribution

The first case we investigate is when does Poisson's Binomial Distribution behave like the classical Binomial Distribution that we discussed in Section 3.

A Poisson binomial distribution PBcan be approximated by a binomial distribution $B$where $\mu_p$is the probability of success for each trial of$B$.
Ehm [10] determined bounds for the error introduced when approximating $PB$with B. Let $\mu = \mu_p$and $\nu = 1 - \mu$and $d(PB, B)$ be the total variation distance of $PB$with B.
Then

$$C \cdot \min\left(1, \frac{1}{N\mu\nu}\right) \sum_{i=1}^{N} (p_i - \mu)^2 \leq d(PB, B) \leq \frac{1 - \mu^{N+1} - \nu^{N+1}}{(N+1)\mu\nu} \sum_{i=1}^{N} (p_i - \mu)^2$$

where$C \geq 1 / 124$.

The closer $p_i$are to μ_p, i.e. $\sigma_p^2 \to 0$, the more $d(PB, B)$ tends to 0. We can also show this using Equations 7 and 8.

$$\lim_{\sigma_p^2 \to 0} E[\mu_X] = \lim_{\sigma_p^2 \to 0} \mu_p N = \mu_p N$$

$$\lim_{\sigma_p^2 \to 0} E[\sigma_X^2] = \lim_{\sigma_p^2 \to 0} \left(\mu_p - \sigma_p^2 - \mu_p^2\right) N = N\mu_p\left(1 - \mu_p\right)$$

which is the mean and variance of a Binomial Distribution with N trials and probability of success μ_p.

Now that we have a Binomial Distribution, we can approximate it as in Section 3.

$$F(N) \approx f_1(N) = \frac{1}{2}\left(\Phi\left(\left(2\mu_p - 1\right)\sqrt{\frac{N+1}{2}}\right) + \Phi\left(\frac{N\left(2\mu_p - 1\right) + 1}{2\sqrt{N\mu_p\left(1 - \mu_p\right)}}\right)\right)$$

Approximation by Poisson Distribution

A Poisson binomial distribution PBcan also be approximated by a Poisson distribution $Po$with mean $\lambda = \mu_X$.

Barbour and Hall [11] have shown that

$$\frac{1}{32}\min\left(\frac{1}{\lambda},1\right)\sum_{i=1}^{N} p_i^2 \leq d(PB,Po) \leq \frac{1-e^{-\lambda}}{\lambda}\sum_{i=1}^{N} p_i^2$$

where $d(PB,B)$ is the total variation distance of $PB$and Po. Clearly, the smaller the $p_i$are, i.e. $\mu_p, \sigma_p^2 \to 0$, the better Poapproximates PB. We can also show this using Equations 7 and 8.

$$\lim_{\mu_p,\sigma_p^2\to 0} E[\mu_X] = \lim_{\sigma_p^2\to 0} \mu_p\, N \ = \mu_p N$$
$$\lim_{\mu_p,\sigma_p^2\to 0} E[\sigma_X^2] = \lim_{\sigma_p^2\to 0} \left(\mu_p - \sigma_p^2 - \mu_p^2\right)N = \mu_p N$$

which is the mean and variance of a Poisson Distribution with $\lambda = \mu_X = \mu_p N$.

Thus, we can approximate $F(N)$ (as defined in Equation 9) using the CDF of the Poisson Distribution as

$$F(N) = P\left(X > \left\lfloor\frac{N}{2}\right\rfloor\right) \approx 1 - \frac{\Gamma(n+1,\lambda)}{n!}$$

From Theorem 2, we get the following.

$$F(N) \approx f_2(N) = \Phi\left(\left(2\mu_p - 1\right)\sqrt{\frac{N+1}{2}}\right)$$

Approximation by Normal Distribution

For large values of N, the Central Limit Theorem allows us to approximate Poisson's Binomial Distribution by a normal distribution with mean $\mu_X$and variance σ_X^2. Thus, we get

$$F(N) \approx f_3(N) = E\left[1 \ - \ \Phi\left(\frac{\frac{N-1}{2} - \mu_X}{\sqrt{\sigma_X^2}}\right)\right] = \ \Phi\left(\frac{2E[\mu_X] \ - \ N + 1}{2\sqrt{E[\sigma_X^2]}}\right)$$

This, with Equations 7 and 8, yields

$$F(N) \approx f_3(N) \approx \Phi\left(\frac{N\left(2\mu_p - 1\right) + 1}{2\sqrt{N\left(\mu_p - \sigma_p^2 - \mu_p^2\right)}}\right)$$

These approximations allow us to efficiently estimate the ensemble accuracy without exhaustive simulations. In fact, all these approximations use the standard normal distribution and hence, all the approximations take constant time to compute.

5. A GENERAL RECIPE

Based on our findings, this section outlines a structured approach to estimating ensemble accuracy efficiently. This step-by-step methodology starts from dataset preparation and weak learner evaluation and ends with selecting an appropriate approximation technique based on the variance of weak learner accuracies.

1. Dataset Preparation

   a. Select an appropriate dataset for the study.
   b. If the dataset has multiple classes but the analysis focuses on binary classification, exclude one or more labels to simplify the data into two classes.
   c. Split the dataset into training and validation sets to ensure balanced evaluation and avoid overfitting.

2. Setup Initial Classifiers

   a. Define the weak learners to be used in the ensemble (e.g., decision trees, Random Forest with shallow depth, etc.).
   b. Assume the probabilities of success for each classifier are drawn from a common normal distribution.
   c. Create a fixed number of weak learners (e.g., 100 classifiers) with similar configurations (e.g. random Forrest with trees of a fixed depth).
   d. Record the accuracy of each classifier on the validationset.
   e. Using the accuracies of the classifiers as their respective probabilities of success, estimate the mean (μ_p) and variance (σ_p^2) of the success probabilities from the recorded accuracies.

3. Theoretical Approximation

   Depending on the mean and variance of the success probabilities, choose an approximation to be used:

   a. Low Variance($\sigma_p^2 \approx 0$): Use the Binomial approximation.
   b. Small Success Probabilities($\mu_p, \sigma_p^2 \approx 0$): Use the Poisson approximation.
   c. General Case(Large N): Use the Normal approximation.

6. A CASE STUDY: THE IRIS DATA SET

Following the steps outlined in the previous section, we construct ensembles of varying sizes using Random Forest classifiers with limited tree depth and compare the predicted accuracy against actual ensemble performance.

The Python code for this project can be found in the following GitHub repository: github.com/SidShah2953/Estimating-Number-of-Weak-Learners.

The Iris data set [12] is a well-known data set in machine learning, commonly used for multi-class classification tasks.In this study, we focus on a binary classification problem by excluding one of the three labels. This simplified the data set to facilitate analysis with binary classification techniques. The data set was then divided equally into training and testing sets to ensure a balanced evaluation.

We begin the analysis with Random Forest classifiers restricted to a maximum depth of 1.

1. Before equation 7, we assume that the underlying classifiers have their probabilitiesof success randomly generated by a common normal distribution. The goal was to estimate the distribution of accuracies generated by these classifiers. To achieve this,

   a. A hundred Random Forest classifiers were created, each of size 1.
   b. The accuracy of each classifier was recorded to estimate. The mean and variance of the sample were used to approximate μ_p and $\sigma_p{}^2$.

   $$\mu_p \approx 0.90 \qquad \sigma_p \approx 4.44 \times 10^{-16}$$

2. The results showed a high mean accuracy and an almost negligible variance. Basedon this, we should use the binomial approximation (see Section 4.1) to predict the performance of ensembles of varying sizes.

   For the sake of demonstration, we use all three approximations.

3. Ensembles of sizes ranging from 11to 51 were analyzed using two binomial approximations. The following metrics were recorded for each ensemble size, in Table 3:

   - Actual accuracy was measured by creating a random forest of that size.
   - Predicted accuracies for all three approximations.
   - Relative errors between predicted and actual accuracies for all approximations.

This brings us to an important assumption that we have made, which might not always be true: We assume that increasing the number of learners always increases the accuracy of the ensemble, which is not always the case.

Table 3: Performance of approximations on the Iris Dataset, predicting the accuracy of Random Forests with maximum depth 1.

N	Actual	Binomial		Poisson		Normal	
		Est.	% RE	Est.	% RE	Est.	% RE
11	0.880	0.987	12.215	0.975	10.793	1.000	13.636
21	0.940	0.998	6.171	0.996	5.959	1.000	6.383
31	0.940	1.000	6.346	0.999	6.310	1.000	6.383
41	0.940	1.000	6.376	1.000	6.370	1.000	6.383
51	0.940	1.000	6.382	1.000	6.381	1.000	6.383
Mean			7.498		7.162		7.834

7. CONCLUSION

In this paper, we present a simple, probabilistic framework for estimating the accuracy of a bagged ensemble that does not necessitate costly calculations. We investigated three key approximations using majority voting: the *Binomial approximation*, the *Poisson approximation*, and the *Normal approximation*.

The advantage of these methods stems from the constant-time lookup of standard normal distribution values. Despite the need to train the weak learners individually to estimate probabilistic parameters, our system is more efficient because the learners are assumed to be independent, allowing us to parallelize this training.

Furthermore, empirical validation with the Iris dataset revealed that our approximations are close to the actual ensemble performance.

One fundamental assumption that we made is that increasing the number of learners always improves ensemble accuracy. This is not always the case. The accuracy of an ensemble is determined by several factors, including the underlying data distribution and the type of weak learners used. By representing each learner as a Bernoulli trial, we abstracted these complexities, allowing us to focus solely on probabilistic estimation.

While the assumption of independence gives us numerous advantages, it also limits the type of ensembles we can investigate. By relaxing this assumption, we can consider various boosting algorithms and other complex algorithms that dynamically change based on the performance of the learners on the dataset.

This work contributes to the development of more efficient and interpretable ensemble learning techniques by offering a structured approach to ensemble accuracy estimation. Future studies can refine these approximations and extend their applicability to a broader range of machine learning models by relaxing one or more of the approximations we have taken into consideration.

REFERENCES

[1] R. Clemen, “Combining forecasts: A review and annotated bibliography,” International Journal of Forecasting 5, pp. 559–581, 1989.

[2] M. E. Lopes, Estimating the algorithmic variance of randomized ensembles via the bootstrap, 2019. arXiv: 1907.08742[math.ST]. [Online]. Available: https://arxiv. org/abs/1907.08742.

[3] L. Slivinski and C. Snyder, “Exploring practical estimates of the ensemble size necessary for particle filters,” Monthly Weather Review, vol. 144, Nov. 2015. DOI:10.1175/MWR-D-14-00303.1.

[4] B. Christiansen, “Estimating the gain of increasing the ensemble size from analytical considerations,” Quarterly Journal of the Royal Meteorological Society, vol. 150, no. 764, pp. 4270–4284, 2024. DOI: 10.1002/qj.4815.

[5] W. Feller, Introduction to Probability Theory and Its Applications. J. Wiley and Sons, 1950.

[6] N. d. Condorcet, “Essai sur l’application de l’analyse `a la probabilit´e des d´ecisions rendues `a la pluralit´e des voix,” in Essai sur l’application de l’analyse `a la probabilit´e des d´ecisions rendues `a la pluralit´e des voix (Cambridge Library Collection Mathematics), Cambridge Library Collection - Mathematics. Cambridge University Press, 2014, pp. 1–2.

[7] F. A. Haight, Handbook of the Poisson distribution [by] Frank A. Haight. (Operations Research Society of America. Publications in operations research, no. 11), Eng. New York: Wiley, 1967, p. 2.

[8] Papoulis and S. Unnikrishna Pillai, Probability, Random Variables and Stochastic Processes, en, 4th ed. McGraw Hill Higher Education, 2001, p. 87.

[9] Y. H. Wang, “On the number of successes in independent trials,” Statistica Sinica, pp. 298–312, 1993.

[10] W. Ehm, “Binomial approximation to the Poisson binomial distribution,” Statistics & Probability Letters, vol. 11, no. 1, pp. 7–16, 1991, ISSN: 0167-7152. DOI: 10.1016/0167-7152(91)90170-V.

[11] D. Barbour and P. Hall, “On the rate of Poisson convergence,” Mathematical Proceedings of the Cambridge Philosophical Society, vol. 95, no. 3, pp. 473–480,1984. DOI: 10.1017/S0305004100061806.

[12] R. A. Fisher, “The use of multiple measurements in taxonomic problems,” Annals of Eugenics, vol. 7, no. 2, pp. 179–188, 1936. DOI: 10.1111/j.1469-1809.1936. tb02137.x.

Appendix: Predicted Vs Actual Accuracy for Identical Learners

p = 0.51							
	Actual	Poisson		Normal		Average	
N	Value	Est.	% Rel. Err.	Est.	% Rel. Err.	Est.	% Rel. Err.
11	0.527	0.520	-1.43	0.644	22.10	0.582	10.34
21	0.537	0.526	-1.96	0.622	15.78	0.574	6.91
31	0.545	0.532	-2.35	0.614	12.81	0.573	5.23
41	0.551	0.537	-2.68	0.612	11.00	0.574	4.16
51	0.557	0.541	-2.95	0.611	9.75	0.576	3.40
61	0.562	0.544	-3.20	0.612	8.82	0.578	2.81
71	0.567	0.548	-3.42	0.613	8.09	0.580	2.34
81	0.572	0.551	-3.62	0.615	7.50	0.583	1.94
91	0.576	0.554	-3.81	0.616	7.01	0.585	1.60
p = 0.55							
	Actual	Poisson		Normal		Average	
N	Value	Est.	% Rel. Err.	Est.	% Rel. Err.	Est.	% Rel. Err.
11	0.633	0.597	-5.74	0.738	16.52	0.667	5.39
21	0.679	0.630	-7.23	0.752	10.71	0.691	1.74
31	0.713	0.655	-8.10	0.770	8.02	0.713	-0.04
41	0.741	0.677	-8.67	0.788	6.41	0.732	-1.13
51	0.764	0.695	-9.05	0.805	5.32	0.750	-1.87
61	0.784	0.711	-9.31	0.820	4.52	0.765	-2.39
71	0.802	0.726	-9.47	0.833	3.91	0.779	-2.78
81	0.817	0.739	-9.58	0.845	3.42	0.792	-3.08
91	0.831	0.751	-9.63	0.856	3.03	0.804	-3.30

p = 0.60							
	Actual	Poisson		Normal		Average	
N	Value	Est.	% Rel. Err.	Est.	% Rel. Err.	Est.	% Rel. Err.
11	0.753	0.688	-8.71	0.838	11.16	0.763	1.23
21	0.826	0.746	-9.59	0.877	6.17	0.812	-1.71
31	0.872	0.788	-9.58	0.907	4.01	0.847	-2.78
41	0.903	0.820	-9.21	0.929	2.79	0.875	-3.21
51	0.926	0.846	-8.68	0.945	2.03	0.896	-3.32
61	0.943	0.867	-8.08	0.958	1.51	0.912	-3.28
71	0.956	0.885	-7.46	0.967	1.15	0.926	-3.15
81	0.966	0.900	-6.84	0.974	0.88	0.937	-2.98
91	0.973	0.913	-6.25	0.980	0.68	0.946	-2.78
p = 0.65							
	Actual	Poisson		Normal		Average	
N	Value	Est.	% Rel. Err.	Est.	% Rel. Err.	Est.	% Rel. Err.
11	0.851	0.769	-9.69	0.913	7.24	0.841	-1.23
21	0.923	0.840	-8.96	0.953	3.22	0.896	-2.87
31	0.958	0.885	-7.59	0.974	1.68	0.929	-2.95
41	0.976	0.915	-6.22	0.985	0.94	0.950	-2.64

51	0.986	0.937	-5.00	0.992	0.55	0.964	-2.22
61	0.992	0.953	-3.97	0.995	0.32	0.974	-1.82
71	0.995	0.964	-3.14	0.997	0.19	0.981	-1.47
81	0.997	0.973	-2.47	0.998	0.12	0.986	-1.17
91	0.998	0.979	-1.93	0.999	0.07	0.989	-0.93
p = 0.70							
	Actual	Poisson		Normal		Average	
N	Value	Est.	% Rel. Err.	Est.	% Rel. Err.	Est.	% Rel. Err.
11	0.922	0.836	-9.26	0.962	4.38	0.899	-2.44
21	0.974	0.908	-6.77	0.987	1.42	0.948	-2.68
31	0.990	0.945	-4.57	0.996	0.53	0.970	-2.02
41	0.996	0.967	-2.99	0.998	0.21	0.983	-1.39
51	0.999	0.979	-1.94	0.999	0.08	0.989	-0.93
61	0.999	0.987	-1.24	1.000	0.03	0.993	-0.61
71	1.000	0.992	-0.80	1.000	0.01	0.996	-0.39
81	1.000	0.995	-0.51	1.000	0.01	0.997	-0.25
91	1.000	0.997	-0.33	1.000	0.00	0.998	-0.16
p = 0.75							
	Actual	Poisson		Normal		Average	
N	Value	Est.	% Rel. Err.	Est.	% Rel. Err.	Est.	% Rel. Err.
11	0.966	0.890	-7.87	0.988	2.33	0.939	-2.77
21	0.994	0.951	-4.25	0.998	0.46	0.975	-1.90
31	0.999	0.977	-2.15	1.000	0.10	0.988	-1.02
41	1.000	0.989	-1.07	1.000	0.02	0.994	-0.52
51	1.000	0.995	-0.53	1.000	0.01	0.997	-0.26
61	1.000	0.997	-0.27	1.000	0.00	0.999	-0.13
71	1.000	0.999	-0.13	1.000	0.00	0.999	-0.07
81	1.000	0.999	-0.07	1.000	0.00	1.000	-0.03
91	1.000	1.000	-0.03	1.000	0.00	1.000	-0.02