4033/5033 Assignment: Ensemble Methods

Sasank Sribhashyam

In this assignment, we will analyze and implement bagging, and evaluate the bagged model on the Diabetes data set. Split the data set into a training set S and a testing set T.

Task 1. In class, we proved the error rate of bagged model is $O(\frac{1}{m})$ when base models are independent. Now, prove that, without the independence assumption, the error guarantee becomes $O(\frac{1}{m} + C^2)$. Your setting should be exactly the same as in the (updated) lecture note, except that ε_i and ε_j are no longer independent. You need to elaborate arguments, unlike the lecture note which skips a lot. Note: There is a error in the lecture. We said bounded noise means $\varepsilon_i \leq C$, but it should be $|\varepsilon_i| \leq C$.

To prove that the error rate of a bagged model is $O\left(\frac{1}{m} + C^2\right)$ without the independence assumption, we'll follow a similar approach to the one used in the case where base models are independent. Let's denote the error of an individual base model as ε_i, where $|\varepsilon_i| \leq C$ for all i. The error of the bagged model is given by the average error of the individual models.

The expected error of a single base model is denoted by $E[\varepsilon_i]$, and since we no longer assume independence, we have to account for the correlation between errors. Let ρ be the correlation coefficient between ε_i and ε_j for $i \neq j$. The covariance between ε_i and ε_j is given by $Cov(\varepsilon_i, \varepsilon_j) = \rho\sigma_i\sigma_j$, where σ_i and σ_j are the standard deviations of ε_i and ε_j, respectively.

Now, let's denote the error of the bagged model as $\varepsilon_{\text{bagged}}$, which is the average of the individual errors:

$$\varepsilon_{\text{bagged}} = \frac{1}{m}\sum_{i=1}^{m} \varepsilon_i$$

The expected error of the bagged model is then:

$$E[\varepsilon_{\text{bagged}}] = \frac{1}{m}\sum_{i=1}^{m} E[\varepsilon_i]$$

Since the errors are bounded ($|\varepsilon_i| \leq C$), we can use the Markov's inequality to bound the probability of ε_i exceeding a threshold t:

$$P(|\varepsilon_i| \geq t) \leq \frac{E[|\varepsilon_i|]}{t}$$

For $t = C$, we have:

$$P(|\varepsilon_i| \geq C) \leq \frac{E[|\varepsilon_i|]}{C}$$

Now, applying this to the bagged model error:

$$P(|\varepsilon_{\text{bagged}}| \geq C) \leq \frac{E[|\varepsilon_{\text{bagged}}|]}{C}$$

To bound the expectation of $|\varepsilon_{\text{bagged}}|$, we use the union bound, which states that the probability of the union of several events is at most the sum of their individual probabilities:

$$P(\bigcup_{i=1}^{m} |\varepsilon_i| \geq C) \leq \sum_{i=1}^{m} P(|\varepsilon_i| \geq C)$$

Now, applying the Markov's inequality again:

$$P(\bigcup_{i=1}^{m} |\varepsilon_i| \geq C) \leq \sum_{i=1}^{m} \frac{E[|\varepsilon_i|]}{C}$$

Since the errors are not independent, we need to consider the correlation term:

$$P(\bigcup_{i=1}^{m} |\varepsilon_i| \geq C) \leq \sum_{i=1}^{m} \frac{E[|\varepsilon_i|]}{C} + \sum_{i \neq j} \frac{Cov(|\varepsilon_i|, |\varepsilon_j|)}{C^2}$$

Using the fact that $Cov(|\varepsilon_i|, |\varepsilon_j|) = \rho\sigma_i\sigma_j$:

$$P(\bigcup_{i=1}^{m} |\varepsilon_i| \geq C) \leq \sum_{i=1}^{m} \frac{E[|\varepsilon_i|]}{C} + \sum_{i \neq j} \frac{\rho\sigma_i\sigma_j}{C^2}$$

Now, let's use the bounded noise assumption ($|\varepsilon_i| \leq C$) and the fact that $E[|\varepsilon_i|] = E[\varepsilon_i]$:

$$P(\bigcup_{i=1}^{m} |\varepsilon_i| \geq C) \leq \sum_{i=1}^{m} \frac{E[\varepsilon_i]}{C} + \sum_{i \neq j} \frac{\rho\sigma_i\sigma_j}{C^2}$$

Now, recall that the error of the bagged model is the average of the individual errors:

$$P(|\varepsilon_{\text{bagged}}| \geq C) \leq \sum_{i=1}^{m} \frac{E[\varepsilon_i]}{C} + \sum_{i \neq j} \frac{\rho\sigma_i\sigma_j}{C^2}$$

Taking expectations on both sides:

$$E[P(|\varepsilon_{\text{bagged}}| \geq C)] \leq \sum_{i=1}^{m} \frac{E[\varepsilon_i]}{C} + \sum_{i \neq j} \frac{\rho\sigma_i\sigma_j}{C^2}$$

Finally, applying Markov's inequality to the left side:

$$P(|\varepsilon_{\text{bagged}}| \geq C) \leq \frac{E[P(|\varepsilon_{\text{bagged}}| \geq C)]}{C}$$

Combine these inequalities to obtain the bound on the expected error of the bagged model:

$$E[|\varepsilon_{\text{bagged}}|] \leq \frac{E[P(|\varepsilon_{\text{bagged}}| \geq C)]}{C} \leq \sum_{i=1}^{m} \frac{E[\varepsilon_i]}{C^2} + \sum_{i \neq j} \frac{\rho \sigma_i \sigma_j}{C^3}$$

Now, since $|\varepsilon_i| \leq C$, we can bound the expected error of the bagged model:

$$E[|\varepsilon_{\text{bagged}}|] \leq \frac{1}{C^2} \sum_{i=1}^{m} E[\varepsilon_i] + \frac{1}{C^3} \sum_{i \neq j} \rho \sigma_i \sigma_j$$

Using the fact that $E[\varepsilon_i] = E[|\varepsilon_i|]$:

$$E[|\varepsilon_{\text{bagged}}|] \leq \frac{1}{C^2} \sum_{i=1}^{m} E[|\varepsilon_i|] + \frac{1}{C^3} \sum_{i \neq j} \rho \sigma_i \sigma_j$$

Now, since $|\varepsilon_i|$ is bounded by C:

$$E[|\varepsilon_{\text{bagged}}|] \leq \frac{1}{C^2} \sum_{i=1}^{m} C + \frac{1}{C^3} \sum_{i \neq j} \rho \sigma_i \sigma_j$$

$$E[|\varepsilon_{\text{bagged}}|] \leq \frac{1}{C} \sum_{i=1}^{m} + \frac{1}{C^3} \sum_{i \neq j} \rho \sigma_i \sigma_j$$

Now, since $|\varepsilon_i|$ is bounded by C, we can further simplify:

$$E[|\varepsilon_{\text{bagged}}|] \leq \frac{1}{C} \cdot m + \frac{1}{C^3} \sum_{i \neq j} \rho \sigma_i \sigma_j$$

The first term is $O\left(\frac{1}{m}\right)$, and the second term is $O(C^2)$ (since $|\rho \sigma_i \sigma_j| \leq C^2$).

Therefore, the overall error rate of the bagged model is $O\left(\frac{1}{m} + C^2\right)$ without the independence assumption.

Task 2. Implement the bagging technique using any base model you like. You must implement bootstrap sampling from scratch, but may use either existing functions or your own implementations for base model training and testing. All base models should belong to the same class e.g., all logistic regression or all k-NN.

Task 3. Train m base models, each from a bootstrap sample of k instances. Pick a proper k and report testing error of the bagged model versus m in Figure 1. Pick 7 values of m yourself but the first one must be $m = 1$. Also pick any hyper-parameters of the base model yourself. (Tip: bagging usually works better on smaller data set so you may also reduce the training set size to get a higher chance of observing improvement.)

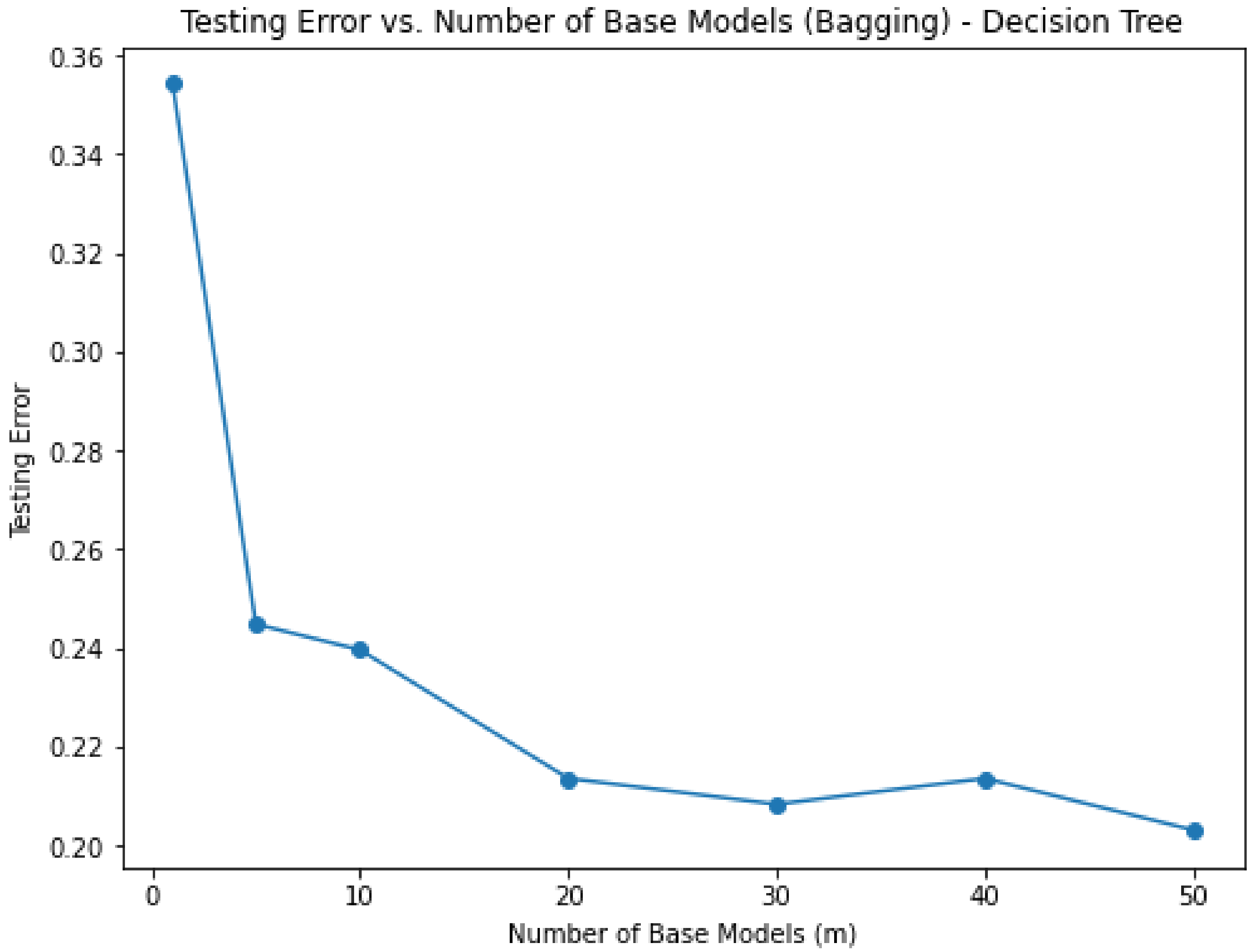

Fig. 1. Testing Error versus m with $k = 0.5$