INF554 Assessment: Trees, Bagging and Boosting

Ziad Oumzil

October 2023

1 Question 1

1. Interpretation of the three functional properties of impurity measures in the context of measuring node impurity :

   — **Property 1 :** This property implies that when the node has a diversification of its elements, the impurity measure is higher. The node is most impure when it has an equal distribution of all classes.
   — **Property 2 :** This property shows that the impurity is the least when the one unique classes is represented in the node. Hence, the node is perfectly pure, if its elements are from one class.
   — **Property 3 :** The impurity measure remains consistent regardless of the order or labeling of the classes. This implies that impurity is a function of the distribution of instances across classes and is independent of how the classes are labeled or ordered.

2. Analytical proof for binary classification that Gini index satisfies these properties : For binary classification, the gini index is given by this formula, for $p = [p_1, p_2] = [q, 1-q]$:

   $$G(p) = 1 - q^2 - (1-q)^2 = 1 + 2q - 2q^2$$

   it's differential respect to q is :

   $$G'(p) = 2 - 4q(1)$$

   — **Property 1 :** The gini index is a C^2 function respect to q, and a strictly concave function. Hence it has a unique maximum q^* solution of $G'(p) = 0$, we get $(p_1^*, p_2^*) = (q^*, 1-q^*) = (1/2, 1/2)$.
   — **Property 2 :** By (1), we can affirm that $G(q)$ is increasing on $[0, 1/2]$, and decreasing on $[1/2, 1]$, with $G(p) = 0$ when $q = 1$ or $q = 0$. Hence the function achieves it minimum at the extremum.

   $$p_{min} = [1, 0] \ or \ [0, 1]$$

   — **Property 3 :** The gini index remains the same when we swap the variables p_1 and p_2,

   $$G([1-q, q]) = 1 - (1-q)^2 - q^2 = G([q, 1-q])$$

   Hence, the Gini index satisfies all three properties for binary classification.

2 Question 2 :

Compute the probability that a given observation in a data set of size N is part of a bootstrap sample, and study the limit of this probability as N tends to infinity.

— Each observation has the same likelihood of being drawn at each step (which means selecting one observation randomly with replacement). Hence the probability of an observation not being selected.

$$P(not\ selected\ in\ one\ draw) = 1 - P(selected\ in\ one\ draw) = 1 - \frac{1}{N}$$

— An observation is not selected in a bootstrap sample, which means we performed N draws, and the observation is not selected on any of the draws.

$$P(\textit{not selected in the bootstrap}) = P(\textit{not selected in one draw})^N = (1 - \frac{1}{N})^N$$

— Hence the probability to be selected in one bootstrap is :

$$P(\textit{selected}) = 1 - P(\textit{not selected in the bootstrap}) = 1 - (1 - \frac{1}{N})^N$$

— **Analyzing the limit :** It's well known that the sequence $(1 - \frac{1}{N})^N$ converges to e^{-1}.
proof :

$$(1 - \frac{1}{N})^N = exp\left(Nln(1 - \frac{1}{N})\right) = exp\left(N \times (-\frac{1}{N} + o(\frac{1}{N}))\right)$$

Hence

$$(1 - \frac{1}{N})^N = exp\left(-1 + o(1)\right) \longrightarrow e^{-1} \quad \text{as } N \longrightarrow \infty$$

We conclude that

$$\lim P(\textit{selected}) = 1 - e^{-1} \approx 0.63212055882$$

3 Question 3 :

— *Which is the most important predictor according to MDA ?*

According to MDA, the most important predictor is Glucose. The second import one is IBM, then comes Age.

— *Is this information consistent with the fitted logistic regression parameters obtained from Task 5 ?*

To obtain information from the regression parameters, I performed logistic regression on standardized data. The coefficients corresponding to the 8 variables are as follows :

$$\begin{bmatrix} 0.18 & 1.98 & -0.05 & 0.71 & 0.62 & 1.39 & 1.05 & -0.60 \end{bmatrix}$$

For example, we observe that the second variable (that corresponds to Glucose) has the largest logistic regression parameter. The magnitude of each coefficient is an indicator of its importance. A larger magnitude implies a more significant impact of the predictor, especially when considering Glucose. The sign of a logistic regression coefficient reflects the direction of its relationship with the log-odds of the response variable. In the case of a positive coefficient, as the predictor value increases, the log-odds of the response variable also increase, thereby increasing the likelihood of the positive class and making it more probable.

4 Question 4 :

— We can observe, from the bar plot, that some few observations have very large average weights.
— This observation suggests that some of the data samples are more difficult to be classified, and then receive more emphasis during the AdaBoost iterations. For example, outliers are hard to classify and then get significant weight over the process, which leads the process to focus on these outliers. Consequently, potential issues such as overfitting and performance degradation may arise.

5 Question 5 :

In order to prevent some weights from getting larger, we propose a new way to update the weights such that the new weights will never be higher than a certain w_{max}

$$w_i^{(k+1)} = w_i^{(k)} exp(\alpha_b I[y_i \neq C_b(x_i)])$$

We then normalize the weights and apply the threshold.

$$w_i = minimum(w_i, w_max)$$

other methods could be either introducing a penalty when weights become large or deleting observations that have a large weight.

6 Question 6 :

— **XGBoost :** XGBoost is an optimized gradient boosting algorithm, known for its impressive predictive accuracy. XGBoost includes regularization techniques to prevent overfitting and the handling of missing values. It excels in both regression and classification tasks, employing a gradient boosting framework. It supports parallel and distributed computing, which significantly speeds up the learning process, especially on larger datasets. item **LightGBM :**LightGBM is another gradient boosting framework renowned for its speed and efficiency. It stands out for its ability to handle large datasets. LightGBM employs a histogram-based approach for building decision trees. LightGBM is designed for efficiency, making it a popular choice for both research and production environments.