ML Assessment

Kedadry Yannis

October 26, 2023

Question 1

0.1

The first property implies that the impurity is maximal when a node contains an equal proportion of all possible classes. Such a distribution is the most uncertain.

The second property implies that the impurity is minimal when all the samples in a node belong to a unique and identical class.

The last property ensures that the order of class labels will not modify the impurity. This is indeed the behavior expected in a classification problem.

0.2

Let G be the Gini index formula for $K = 2$. We have $G = 1 - (p_0(t)^2 + p_1(t)^2)$

For the third property:

$$G = 1 - (p_0(t)^2 + p_1(t)^2) = 1 - (p_1(t)^2 + p_0(t)^2) \tag{1}$$

For the second property:

$$\begin{aligned} \text{By definition of } P\text{: } p_0 + p_1 &= 1 \\ p_1 &= 1 - p_0 \\ G = 1 - (p_0(t)^2 + (1 - p_0(t)^2)) \\ G &= 2p_0(t) - 2p_0(t)^2 \\ G &= 2p_0(t)(1 - p_0(t)) \end{aligned}$$

The only roots are then $p_0(t) = 0$ and $p_0(t) = 1$ so we have $\phi(0, 1) = 0$ and $\phi(1, 0) = 0$ as the only minimas.

For the first property:

$$\begin{aligned} \partial G p_0 &= 2 - 4p_0 \\ 2 - 4p_0 = 0 <=> p_0 &= 1/2 \\ \phi(1/2, 1/2) &\text{ is local a local maxima} \end{aligned}$$

Task 3

The results are almost equal; the only difference is that I used integers as candidates output so I couldn't find the 113.5 and found 113 instead.

Question 2

Let's note D the initial data set and B_i the i_th bootstrap set. We have $|D| = |B_i| = N$. Let X be a random variable that chooses uniformly and independently an observation in the data set.

We have $X \sim \mathcal{B}(N, 1/N)$. Then, we have:

$$\begin{aligned}
\mathcal{P}(XinB_j) &= 1 - P(XnotinB_j) \\
&= 1 - \binom{0}{N} \frac{1}{N}^0 (1 - \frac{1}{N})^{N-0} \\
&= 1 - (1 - \frac{1}{N})^N \\
&= 1 - \frac{N-1}{N}^N \\
\lim_{N\to\infty} \mathcal{P}(XinB_j) &= \lim_{N\to\infty} 1 - \frac{N-1}{N}^N \\
&= \lim_{N\to\infty} 1 - \frac{1}{\frac{N}{N-1}^N} \\
&= \lim_{N\to\infty} 1 - \frac{1}{1 + \frac{1}{N-1}^N} \\
&= \lim_{N\to\infty} 1 - \frac{1}{\exp} \\
&= 1 - \frac{1}{\exp} \\
&\approx \frac{2}{3}
\end{aligned}$$

Task 5

The linear regression is not working very well because the output is probably not linear to the data.

Question 3

According to MDA, the most important predictor is Glucose (expected for diabetes diagnoses). Glucose is also the highest parameter obtained from logistic regression. Moreover, the parameter is positive which implies that if the Glucose rises, the probability of having diabetes rises as well.

Question 4

We can observe that the majority of the weight is put on one observation which is probably the outlier. We can conclude that if we have an outlier, all the weights will be put on this point and the AdaBoost will therefore focus on that particular point.

Question 5

The idea is to use a random sub-sample of the input data like in the stochastic version of the gradient descent. By doing so, outliers will appear way less often than inliers and reduce their impacts on the AdaBoost focus over them.

Question 6

The XGBoost algorithm first sorts the data according to features' values. It doesn't use a greedy algorithm which is too demanding in memory but instead uses an approximate algorithm to find a good split. The algorithm is aware of sparsity so it is efficient on spare data as well. It uses system specifications to be fast.

The LightGBM algorithm reduces the number of features by using the sparsity of high dimensional data. It also reduces the number of instances by discarding the ones with small gradients. This is one of the fastest boosting algorithms because it is one of the lightest. A trade-off is that it loses a bit of precision compared to XGBoost for example which is slower.

Task 10

The XGBoost is of course the most accurate, followed by the AdaBoost, the LightGBM, and finally the linear regressor. These results are expected because the LightGBM is faster but loses in precision while the Linear regression is the least accurate because the problem is not linear.