COMP30027 – MACHINE LEARNING, SEMESTER 1, 2019

Project 1: Gaining Information about Naïve Bayes

Leonardo Linardi, 855915

Question 1

As Information Gain (IG) is derived from subtracting the entropy of the instances before splitting (i.e. `H(R)`) and the weighted average of the entropy over the children after the split (Mean Information), IG does impact the overall classifier's behavior.

To illustrate that, for a given dataset, the maximum IG that it can have is `H(R)`, where all the children nodes only contain a single class in it, giving a Mean Information of zero.

And, for each dataset, we can calculate the average IG of all attributes after we split the dataset (with respect to each attribute), divide it by `H(R)` and multiply it by 100%, to capture the IG we get when compared to the maximum IG we can get. Afterwards, we compare it to the performance of the classifier by its accuracy. Figure 1 shows the result.

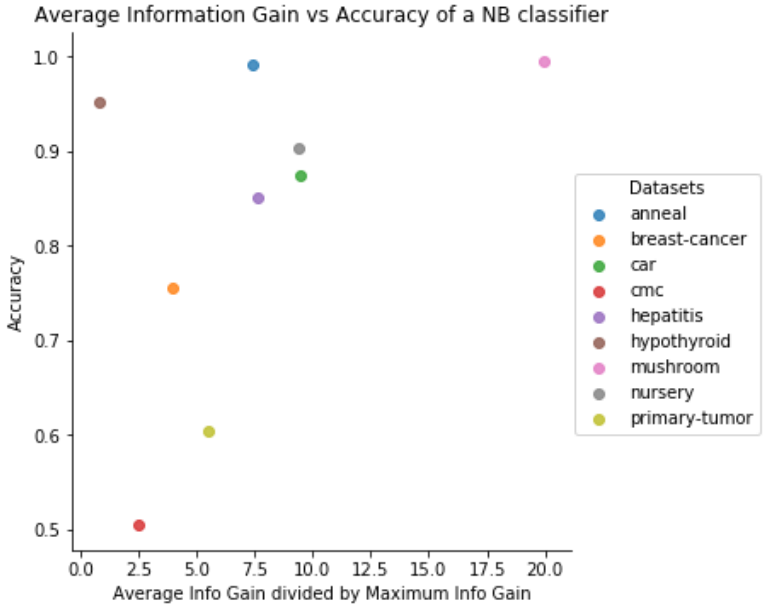

Figure 1. Average information gain vs accuracy of the classifier on various datasets.

We can see that as IG increases, accuracy tends to increase as well. Because a high IG can happen if Mean Information is low, which means that entropies are also low, and the instances are more predictable. But there are 2 datasets that doesn't correspond to this trend, which is the `primary-tumor` and `cmc`.

This could happen because `primary-tumor` and `cmc` have a quite high `H(R)`, `3.6437400563509668` and `1.5390345832497478` respectively. Which means the instances

are evenly distributed with respect to its labels and the classifier has a higher probability of predicting the wrong label, given their even distribution.

Question 5

One of the advanced smoothing regimes, i.e. the add-k smoothing, is going to be implemented on the Naïve Bayes Classifier and compared to other regular smoothing regimes (i.e. Epsilon, Laplace) and not smoothing at all.

Firstly, we are going to decide what constant `k` that would give the highest accuracy to the classifier. Figure 2 shows the accuracy of the classifier on different values of `k`.

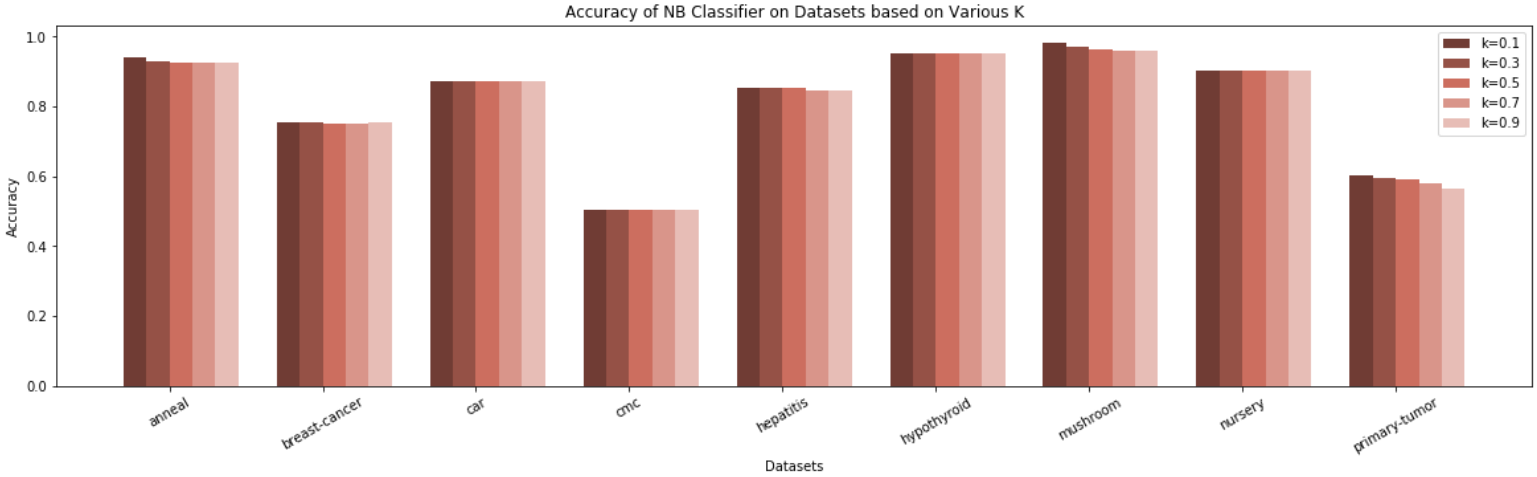

Figure 2. Accuracy of the classifier towards different choices of `k` for add-k smoothing.

Although the accuracy looks similar, but on certain datasets (i.e. `anneal`, `mushroom`, `primary-tumor`) smaller `k` turns out to be slightly better. So, we use `k` = 0.1 for add-k smoothing and compare it towards other smoothing regimes.

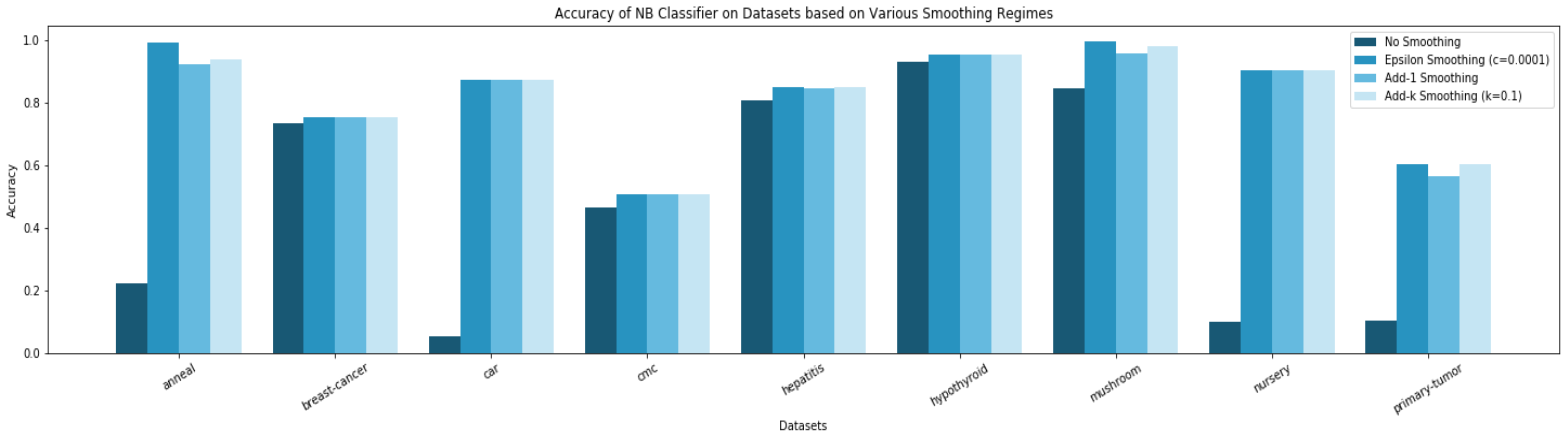

Figure 3. Accuracy of the classifier towards different smoothing regimes.

From Figure 3, we can observe that add-k smoothing has slightly higher accuracy (on some datasets i.e. `anneal, mushroom, primary-tumor`) when compared to add-1 smoothing.

Furthermore, we can clearly see that not smoothing at all would give the worst accuracy. Because without smoothing, this means that for every new attribute value found from the test instances would have a conditional probability of 0. And would result in a very poor score for that label, which often leads to wrong predictions.

Moreover, epsilon smoothing tends to give higher accuracy when compared to add-k smoothing, for instance, on datasets `anneal, hepatitis, mushroom.` This means epsilon smoothing tends to give the right score for the classifier to determine the class, resulting in a higher overall accuracy.