Maëlie LEBARON
214322IV

Open book test 2

ITI8730
November 11, 2021

1 Could labelled points belong to the decision boundary?

In the hard-margin case, no, as both cluster are separated enough to draw a line between them and the goal is to maximise the distance between the *separating plane* and both of the planes *that contain a point, respectively from each cluster*. To sum up, ideally, there is no point on the decision boundary.

In some case, the dataset is not that perfect and cluster can overlap, which mean that the decision boundary can *wrongly label* some points (if they are *"on the wrong side" of the boundary, comparing to the true label*. In other word, soft-margin case can allow this type of things, but in a ideal way, it should not happen. In the case were it happen, it means that the point is as likely to be from one label or another, so it can be considered as outlier.

2 Is the Gini index proportional to the Entropy value? (case of classification).

Gini index and entropy are proportional, but not strictly mathematically speaking. If we take as an example the case with 2 cluster and we plot the gini index and the entropy for each repartition of point between the 2 cluster, we can see that both curve are similar. There is approximately a factor 2 between both indicators. This seems logical as the intuitive principle behind Gini index and entropy is the same : compute the discriminative power of a feature.

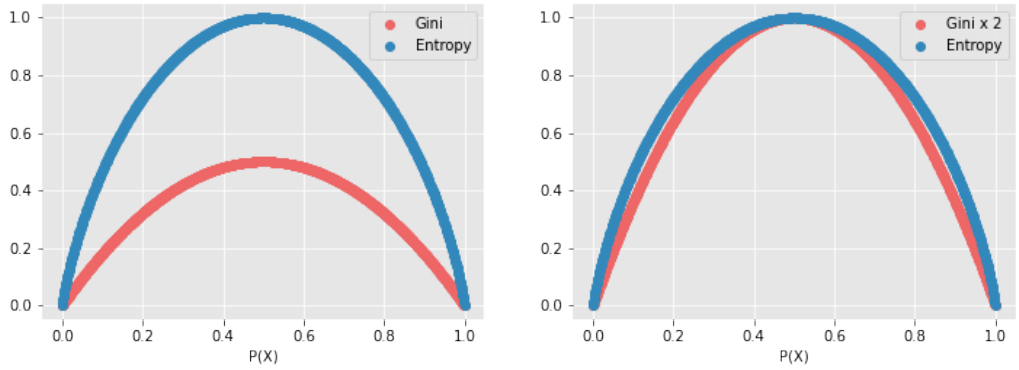

Figure 1: Gini index and entropy depending on the proportion of each dataset, *source: quantdare*

3 Would it be possible to generalize Fisher's score to handle a few variables at once?

It would be interesting to generalize Fisher's score in the case where sectioning two variables together would gives better results than sectioning them independently. Technically, it should be possible to generalize the fisher score to a set of variable instead of one, using joint to create new subset depending on the said features.

4 Why is it not enough to use one characteristic to describe the goodness of the classification model?

The different characteristics of the goodness of a classification model depends en different values and they do not compute the same things. For example, if we want to test for a certain type of disease (as covid-19), the number of fn as much more impact than the number of fp, but not all of the goodness characteristics will take fn in account (TNR). Each characteristics will gives us certain information about the tendencies of the false results (if we have more fn than fp for example), or the global tendency of the model to classify correctly a datapoint.

Also, sometimes, one characteristic can indicate an interesting goodness when other characteristics will mitigate this goodness. For example, we can have a really good accuracy, but with all the *falsely classified data* will be fn, which can lead to wrong interpretation of the data.

5 Which one among the classification model goodness parameters (accuracy, precision, recall specificity, f1-score) are less sensitive to the data balancing?

Accuracy is the less sensitive to data balancing as the formula is, in a simplified way :

$$\frac{nb\_of\_good\_label\_computed}{total\_number\_of\_label\_computed}$$

The other goodness parameters are more sensitive to the data balancing because :

- Precision relies on the number of classified positive by the algorithm
- Recall relies on the number of real positives
- Specificity, or TNR, relies on the number of real negatives

- F1-score is calculated based on recall and specificity (the number of classified negative, and therefore the number of negative in the original dataset, in a way, are not used)

6 Referring to the decision tree growing, which measure of uncertainty (Gini index or Entropy) is preferable in the case of categorical data?

Entropy is used in the information gain, which will compute the difference of entropy between and after the new split.

Gini index will indicate *success or failure* depending on the new split purity (or randomness).

Therefore, Gini index is more well suited for categorical data (and also, it is less computationally expensive).

7 Could Fisher score be used to compute information gain?

Fisher is usually computed before processing, as a way to gain a insight on the quantity of information a feature can gives us. It can be interpreted as the ratio of separation interclass to intraclass.

Information gain is the difference of information we have before and after performing an action on the dataset, it can be used in decision tree algorithm for example. Usually, information gain is computed using entropy, which goal is to gives us the discriminative value of each feature.

In one hand, we could use fisher score to calculate information gain as it also gives us an idea of the discriminative power for a feature. In the other hand, the fisher score depends on the intraclass separation, which is not as interesting when computing information gain.

8 Explain in your own words the meaning of the Bayes theorem.

The Bayes theorem is a probability theorem which gives us the likelihood to something to be depending on other likelihood. I can either be use to determine such probability or to determine if two variables are independent.

In the case of data classification, Bayes theorem is used in the *naïve Bayes model*. In the model we assume that each variable are independant given the class label (even if it is not

true). We then compute the probability of our datapoint to belong to a label knowing its feature, and we then label it as the highest probability label.

9 Explain the difference between the kernel function and the kernel trick.

The idea of the kernel trick is to use a higher dimensionality to separate linearly a non-linearly separable dataset in the original dimensionality. The thing is that to compute certain algorithm, as support vector machines, we need to use the dot product, which is not always easy to compute in high dimension spaces. To overcome this, we use a Kernel function, that can be use as the dot product in higher dimensional space, to replace the dot product.

To sum up, a kernel function in the function can be expressed as a dot product in higher dimensional spaces. The kernel trick is a trick used on non-linearly separable datasets, which use a kernel function to replace the dot product.

10 Does kernel function satisfy the axioms of the distance function?

Kernel function does not systematically satisfy the axioms of distances function. A kernel function does, however, typically satisfy the *symmetric and the non-negative axioms* of distances function, but not necessarily the other axioms.

To sum, kernel function satisfy 2 axioms of the distance functions most of the time, but it can also satisfy less or more than 2 axioms.

A good example we have seen in class is the RBF kernel function, which is not a distance function.

11 Does mean removal is the necessary prerequisite to apply PCA?

Mean removal is not a necessary prerequisite to PCA, but not doing it implies to keep the mean values aside to use them after and weight our results. In either way, the data needs to be modified using the mean.

In a general way, we should do mean-removal before applying PCA.

12 What is the difference between PCA and SVD?

Principal component analysis (PCA) and Singular value decomposition (SVD) are closely related.

The first difference to take into consideration is that they will not return the same thing : PCA will return one and only one set of basics vectors, for the rows, when SVD will return two sets of basics vectors, for the rows *and the columns*.

Also, SVD is really sensitive to mean-translation, as opposite for PCA. As a results, applying PCA on non mean-centered data will not gives the same set of basics vector as the set of *rows* vectors given by SVD.

13 Does Support Monotonicity Property required for the Apriori algorithm?

Support Monotonicity Property : $sup(J) \geq sup(I), \forall J \subset I$

The downward closure property which is implied by the support monotonicity property (as the support in a subset J is at least equal as the support in the subset I, and $sup(I)$ is superior to the minimum support level ($minsup$), which make it frequent, then $sup(J) \geq minsup$).

Therefore, we can not apply the Apriori algorithm without the support monotonicity property, as the prunning relies on the *the downward closure property*.

14 Explain the difference between the Association rule and Frequent pattern.

Frequent pattern are itemsets that have a support superior at minimum support level.

Association rule associate two itemset together that have a confidence superior to minimum confidence level, and the union of both itemset is a frequent itemset.

So the main difference between the association rule and the frequent pattern reside in the *confidence*. Let name the first itemset A, and the second B. The confidence of B knowing A (in the transaction) is : $conf(A \Rightarrow B) = \frac{sup(A \cup B}{sup(A)}$. This is a conditional probability: the association rule take in consideration the dependence between itemsets, when frequent pattern does not.

15 Explain in your own words what collective strength is.

The collective strength, in pattern recognition, is a way to classify the "quality" of a set a item we have grouped together. In the case seen in class with the *market basket data*, in would be for example :

We found a pattern with *cereal, milk, orange juice* that people tend to buy together.

The collective strength is computed on the proportion of violation of the itemset. The violation rate it the number of time this itemset is not completed (for example, someone who bought *cereal and orange juice* without buing *milk*).

The greater the collective strength, the less the itemset is violated, meaning that the relation between items in our itemset is more likely to be strong.

For example : *milk and cereal* should have a very strong collective strength, as they are consumed together most of the time, and *orange juice and cate food* should have a very weak collective strength, as they are, at first sight, not related at all.