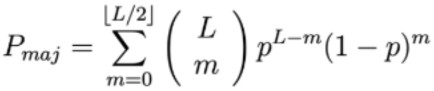
DATA MINING: Quiz 4

Student's name: Marc Nebot Moyano

1. Assume you have 5 *independent* classifiers, each of them with an accuracy of 0.7. Compute which is the accuracy for the *Majority Vote* algorithm for those 5 classifiers.

I'm going to use this formula:



$$P_{maj} = \sum_{m=0}^{\lfloor L/2 \rfloor} \binom{L}{m} p^{L-m}(1-p)^m$$

= 0,84

2. Briefly explain if each of the following claims is true or not and why:
   a. The larger the number of iterations in the bagging method, the lower the variance of results and the larger the accuracy obtained

   This is true because this is the explanation for bagging, taking samples from different methods and eventually reducing the variance.

   b. Boosting cannot be applied to support vector machines because the linear combination of hyperplanes is another hyperplane.

   False, it can be applied especially when svm become weak learners.

   c. When the "a" parameter in random forests is set to the number of features, random forest is equivalent to bagging with decision trees.

   True, since by assigning the parameter "a" we partition the nodes, which is what bagging does, taking into account all possible

d. Diversity of classifiers is the source of success in meta-method. In order to ensure this diversity, we always train classifiers with different training datasets.

True, we can use more methods to obtain diversity in our results, for example bagging with naive bayes or any other method

3. When implementing the main loop of the *Adaboost* procedure, what should we do when the error produced by the classifier on the training set (feed with a set of examples according to the current iteration weights) is equal to 0? Briefly explain why you think so.

   [✓ a.] Stop the boosting iterations and return the weighted ensemble of classifiers built until that moment.

   b. Return that last classifier as the final classifier.

   c. Remove that classifier and continue the boosting loop until the limit number of iterations is achieved.

   d. Reduce the confidence on that classifier (with respect to its theoretical confidence) and continue the boosting loop until the limit number of iterations is achieved.

   e. Boosting cannot be applied in that case.

When it is equal to 0 this happens in the first round so we have to stop the boosting to go back to the one who has made the perfect score.

4. After building a support vector machine with a linear kernel with a given C, we found the number of support vectors is very large. If we want to decrease the number of support vector, what should we do? Explain why.

   a. Decrease the C value
   [✓ b.] Increase the C value
   c. Change to the RBF kernel
   d. Try a Polynomic kernel
   e. None of the above

   When we increase the value of C we increase the margin constraint bar so the execution time will also increase. By increasing C we make the margin narrower so there will be fewer support machines to fit.

5. In the last few years, Artificial Intelligence has advanced a lot. Believe it or not, in the attached file "ChatGPT answers about SVMs.pdf" you will find a dialog I had about Support Vector Machines with ChatGPT. ChatGPT is an amazing chat bot developed by OpenAI that has been trained on a lot of textual data of different types (but without internet access). Its answers have really surprised me for their clarity, expressiveness and knowledge of the topic. However, ChatGPT answers are known to be not always correct (even when it gives convincing explanations... which turn out to be wrong). Your goal is to detect the answers that are wrong (if any) in the SVMs dialog. You have to write here the number of the questions (in red in the pdf) you think are wrong together with the correct answers.

6. Briefly explain if each of the following claims is true or not and why:

   a. In the *apriori* algorithm, given a rule, we will say that it is a good rule if its support and its confidence are above the required thresholds

   True, a good rule will have a good association with an apriori algorithm when it finds all the possible rules provided that its support is greater than or equal to the minsup threshold and its confidence is greater than or equal to the minconf threshold, these being the thresholds of each parameter.

   b. The support required for rules should be always independent of the elements that belong to the itemset.

   False, if they are independent we will not know whether this rule is permissive or restrictive because it will be totally arbitrary.

   c. While finding frequent itemsets, in the main iteration of the algorithm, the itemsets below minimum support of iteration *"i"* should be kept to do pruning in iteration *"i+1"*

   True, they have to maintain the pruning on each one as an apriori pruning if it finds an itemset that is infrequent its superset should not be generated or tested.

   d. The *apriori* algorithm can learn causal rules that explain the behavior of customers.

   True, these rules are based on a context, so if this context is very concrete, its rules will also be very concrete, therefore it can learn this type of rules.

(More space for answers if you need it)