Quiz 2
MINERIA DE DADES

Antoni Cifre Vicens

15-29 December 2020.

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.

Prob	C1	C2	C3	C4	C5	MV
0,7	**0,3**	0,3	0,3	0,3	0,3	0,00243
	0,3	0,3	0,3	0,3	0,7	0,00567
suma MV	**0,3**	0,3	0,3	0,7	0,3	0,00567
0,16308	**0,3**	0,3	0,3	0,7	0,7	0,01323
	0,3	0,3	0,7	0,3	0,3	0,00567
Accuracy	**0,3**	0,3	0,7	0,3	0,7	0,01323
0,83692	**0,3**	0,3	0,7	0,7	0,3	0,01323
	0,3	0,3	0,7	0,7	0,7	0,03087
	0,3	0,7	0,3	0,3	0,3	0,00567
	0,3	0,7	0,3	0,3	0,7	0,01323
	0,3	0,7	0,3	0,7	0,3	0,01323
	0,3	0,7	0,3	0,7	0,7	0,03087
	0,3	0,7	0,7	0,3	0,3	0,01323
	0,3	0,7	0,7	0,3	0,7	0,03087
	0,3	0,7	0,7	0,7	0,3	0,03087
	0,3	0,7	0,7	0,7	0,7	0,07203
	0,7	0,3	0,3	0,3	0,3	0,00567
	0,7	0,3	0,3	0,3	0,7	0,01323
	0,7	0,3	0,3	0,7	0,3	0,01323
	0,7	0,3	0,3	0,7	0,7	0,03087
	0,7	0,3	0,7	0,3	0,3	0,01323
	0,7	0,3	0,7	0,3	0,7	0,03087
	0,7	0,3	0,7	0,7	0,3	0,03087
	0,7	0,3	0,7	0,7	0,7	0,07203
	0,7	0,7	0,3	0,3	0,3	0,01323
	0,7	0,7	0,3	0,3	0,7	0,03087
	0,7	0,7	0,3	0,7	0,3	0,03087
	0,7	0,7	0,3	0,7	0,7	0,07203
	0,7	0,7	0,7	0,3	0,3	0,03087
	0,7	0,7	0,7	0,3	0,7	0,07203
	0,7	0,7	0,7	0,7	0,3	0,07203
	0,7	0,7	0,7	0,7	0,7	0,16807

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 until it reaches its maximum possible.

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

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

   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.

      Normally they are always trained with the same data set to be able to group them and give a score to each classifier based on the score obtained.

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.

   We can select option **b** if we are satisfied with the confidence given initially, but in the event that we want to continue improving said confidence we could choose option **d** to improve the result a little more.

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

   If we make the c smaller, the Alpha is reduced, which also reduces the number of sv.

5. We have a linear SVM trained on a dataset of 100,000 observations. The SVM shows 60,000 supports. Comment each of the following statements:
   a. SVM is Ok. We probably will have a low error on the small test set
      It depends on the distribution of the data, if the clusters are divided we will have almost no errors, but in the case that the clusters are degraded between them, we could have errors, but most likely they are low

   b. We need to increase the number of supports to reduce the error in the test set. Therefore, we should increase C.
      In the case of increases C, the number of sv will increase but it could cause a bigger error in the test set.

   c. We have to reduce the number of supports to reduce the error in the test set. Therefore, we should increase the C parameter.
      In the case of increases in C, the number of sv will increase so it will not be possible to reduce it.

   d. We need to increase the number of supports to reduce the error in the test set. Therefore, we should decrease the C parameter.
      In this case we are looking for the same objective as with -b, in this case with a smaller C the error will also be smaller.

   e. We have to reduce the number of supports to reduce the error in the test set. Therefore, we should decrease the C parameter.
      If we decrease the value of C we will obtain a lower alpha, reducing the number of SV.

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
   b. The support required for rules should be always independent of the elements that belong to the itemset.
   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"
   d. The apriori algorithm can learn causal rules that explain the behavior of customers.

7. If the objective of the clustering algorithm is to find an assignment of individuals to classes so that the distance between different classes is maximized and at the same time minimize the distance between elements of the same class, reason why in the SSE only the distance between elements of the same class is considered.

   Because the SSE evaluates the error of each cluster individually, it does not compare the clusters among the others, in this way the value returned by SSE only depends on the distance between the elements of the cluster itself to obtain the error of each one.

8. We have a data set of about 40,000 customers and we want to study what types of clients we have in our business. You find a problem in some of these solutions? Reason if you find a problem or not.
   a. Run k-means with k = 10 to find 10 kinds of customers.
      It is very likely that the number of different clients is greater than 10, so increasing the k and then grouping the different clusters is better than starting with a too low k.

   b. Apply Ward's hierarchical clustering algorithm.
      it would be a good idea to visualize and group the clusters.

   c. Set k = 10 and find out from there how far away most of the elements are of the data set the k-th element and from this information apply the DBscan algorithm with appropriate parameters.

      It would be a good idea to eliminate people who do not adhere to any set and to be able to find patterns that are hard to find with k-means.

9. Reason if the following sentences are correct or not.
   a. If I know that my data does not have a uniform density, to apply the hierarchical with Link MAX is not suitable.

   b. If I know that my data has a uniform density, to apply the DBScan is a good idea

      Yes, the problem is when the density is too high.

   c. If I know that my data has a uniform density, to apply the kmeans algorithm is a good idea

      Yes, as long as the clusters are visible k-means can identify them.

   d. If I know that my data has non-spherical shapes, to apply the hierarchical with Link MIN is not suitable

      If the clusters are far apart, you can differentiate it.