Quiz 1
MINERIA DE DADES

Antoni Cifre Vicens

15-29 December 2020.

Answer the following questions in the spaces reserved for this use.

1. (1.75pt) Write whether the following problems can be solved using supervised data mining algorithms for classification or not. In case it is not possible, explain very briefly why not.

   (a) Given a dataset describing houses sold in a given city with the sell price, predict the sell price for a new house.

   Yes.

   (b) Given a dataset with information about the outcomes of football matches in the Spanish league, predict the outcome of a match in the English league.

   No, because the data is not correlated with each other.

   (c) Given a dataset with information about the outcomes of football matches in the Spanish league to predict the winner of the league.

   Yes.

   (d) Given a dataset with pictures of hand gestures and their meaning, recognize moving hand gestures in real time.

   Yes.

2. (1pt) You are given a data set on cancer detection. You've build a classification model and achieved an accuracy of 96%. Why shouldn't you be happy with your model performance? What can you do about it?

   The decision taken is not to try the appropriate one, in this case, what should be improved is the recall to avoid leaving a cancer patient to be detected.

3. (1.5pt) Given the following confusion matrices generated on the same testing data, show accuracy for both models. Explain also which model you think is better and why.

(a) Model 1

	Predicted positive	Predicted negative
True positive	51	101
True Negative	40	428

$$Accuracy = \frac{51 + 428}{51 + 428 + 40 + 101} = 0.772$$

$$Recall = \frac{51}{51 + 101} = 0.335$$

$$Precision = \frac{51}{51 + 40} = 0.56$$

$$F = \frac{2 * R * P}{R + P} = 0.419$$

(b) Model 2

	Predicted positive	Predicted negative
True positive	61	91
True Negative	80	388

$$Accuracy = \frac{61 + 388}{61 + 388 + 80 + 91} = 0.724$$

$$Recall = \frac{61}{61 + 91} = 0.401$$

$$Precision = \frac{61}{61 + 80} = 0.432$$

$$F = \frac{2 * R * P}{R + P} = 0.415$$

To make the decision to select one of the two models, we need to know what objective to solve the model has.

In the case that we want to solve a typical problem, we can say that model 1 is the best, but if we have a critical problem that we need the best possible recall, we will select model 2.

4. (1.25pt) When building a classifier using any supervised methods, should we find the best k value for the k-fold cross-validation method in order to obtain the best accuracy? Explain why.

Finding the highest value for k is not only to obtain better precision, but it also helps us to obtain better defined boundaries or to reduce noise sensitivity. But not all supervised methods use this k to improve precision, it can also be used to improve the value of f-measure or they simply do not use such as Naive Bayes

5. (1.75pt) Mark the true sentences and briefly explain your answer.

(a) In general, when training a classifier using the k-nn algorithm on an unbalanced training dataset, the best choice for k is to use high values.

False, having an unbalanced training dataset a small k is better in order to improve the recall of the unbalanced class.

(b) In order to use the k-nn method is enough to have a clean dataset without missing values and containing only numerical attributes.

True, with such a set it would be enough, but if you want to improve, you just need to apply feature selection.

(c) In the k-nn algorithm, the distance-weighted parameter is more relevant when k is large than when k is low.

Creo que False.

(d) In general, the larger the value of k, the better the accuracy because we have more a more robust estimator.

The estimator is generally more robust, but it does not mean that the accuracy is better. There is always a level too high for k.

6. (0.5pt) Why Naive Bayes algorithm is called 'naive'?

This model assumes that each input variable is independent.

7. (0.75pt) Answer if each of the following sentences about the Naive Bayes algorithm is true or not.

   (a) In general, when using Naive Bayes algorithm, the larger the number of features on the dataset, the better the performance

   False

   (b) The smoothing technique is used to reduce the impact of the assumption of independence of features in the dataset.

   Is used to remove noise from a data set.

   (c) When computing the conditional probability of a numerical feature with respect to the class, we always use the normal distribution.

   True

8. (0.75pt) To reduce overfitting of a Decision Tree, mark which of the following method can be used:
   (a) Increase minimum number of examples allowed in leafs
   (b) Increase depth of trees
   (c) Set a threshold on the minimum information gain to split a node

9. (0.75pt) Which of the following are disadvantages of Decision Trees?
   (a) A Decision tree is not easy to interpret
   (b) Decision trees is not a very stable algorithm
   (c) Decision Trees will overfit the data easily if it perfectly describes the training dataset