Ricardo Aler Mur

SELF-ASSESSMENT: Attribute selection and attribute generation, WITH QUESTIONS AND ANSWERS

1) List and justify various reasons why the selection of attributes is important

Answer. 1) the existence of redundant attributes: some classification algorithms such as Naive Bayes learn poorly in the presence of redundant attributes. 2) the existence of irrelevant attributes: although learning algorithms are able to some extent to ignore these attributes, their presence usually degrades the performance of the classifier, therefore it is convenient to remove them during the preprocessing phase. 3) Curse of dimensionality: as the number of dimensions grows, the number of data required to adjust the models grows very quickly (exponentially in the worst case). 4) Sometimes it is important to know which are relevant attributes (for example, which genes are relevant for predicting certain diseases).

2) If when training a linear classifier, the training dataset has as many attributes as training data instances, which would approximately be the percentage of correct answers in test?

Answer: Since a linear classifier is able to separate as much instances as the number of dimensions (even if the data is random), the classifier constructed with this training set will be almost arbitrary, bringing the percentage of correct answers for the test to the chance level. For instance, a linear classifier in two dimensions will be able to classify perfectly any training set containing up to 2+1 instances. In the same way, a linear classifier in 100 dimensions will be able to classify perfectly any training set containing up to 100+1 instances. Given that this is true of any training set, irrespectively of where positive and negative instances are located, the resulting classifier will be arbitrary and the expected success rate on a new (test) dataset will approach chance (i.e. for two-class classification problems, chance success rate is 50%).

3) Why exhaustive search is not feasible in most of the practical problems?

Answer: because the number of subsets that can be done with n attributes grows exponentially.

4) What is the main advantage of the Wrapper vs. Ranking? And its main disadvantage?

Answer: The Wrapper methods are able to detect attributes that work well together but not so well separately. The Ranking methods cannot do this, since the attributes are evaluated individually. The main disadvantage of Wrapper is the time it takes, since it involves launching a machine learning algorithm for each attribute subset to be evaluated.

5) What is the main advantage of Random Projections over PCA?

Answer: Its speed, along with the fact that if the number of target dimensions is high enough, their results are similar.