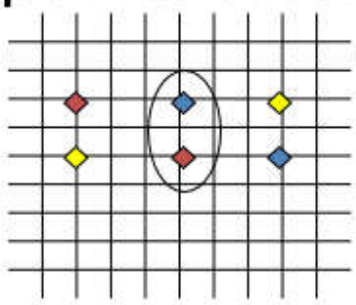
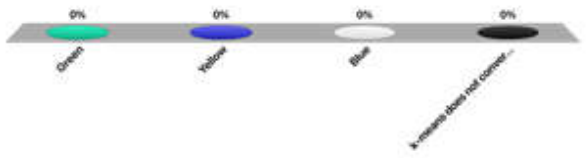
Suppose we have a dataset of pictures and we want to cluster them. Which partitioning algorithm seems more appropriate?

A. k-medoids
B. k-medians
C. k-means
D. none of the above

# What will be the color of the middle points after convergence (k=3)?



A. Green
B. Yellow
C. Blue
D. k-means does not converge



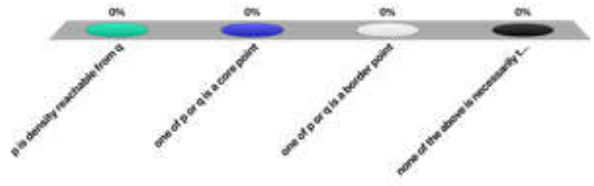
## If p and q are density connected, then …

A. p is density reachable from q

B. one of p or q is a core point

C. one of p or q is a border point

D. none of the above is necessarily true



In density-based clustering, which points can belong to multiple clusters?

A. Core points

B. Border points

C. Outliers

D. None

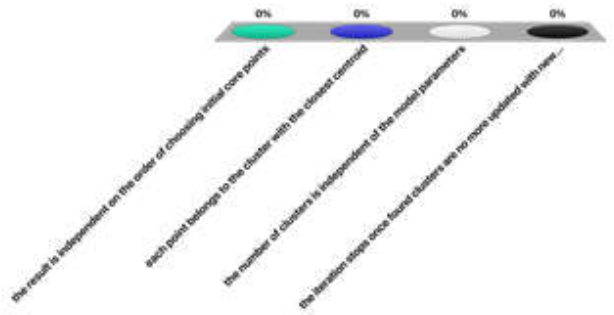
# When executing DBSCAN …

A. the result is independent on the order of choosing initial core points

B. each point belongs to the cluster with the closest centroid

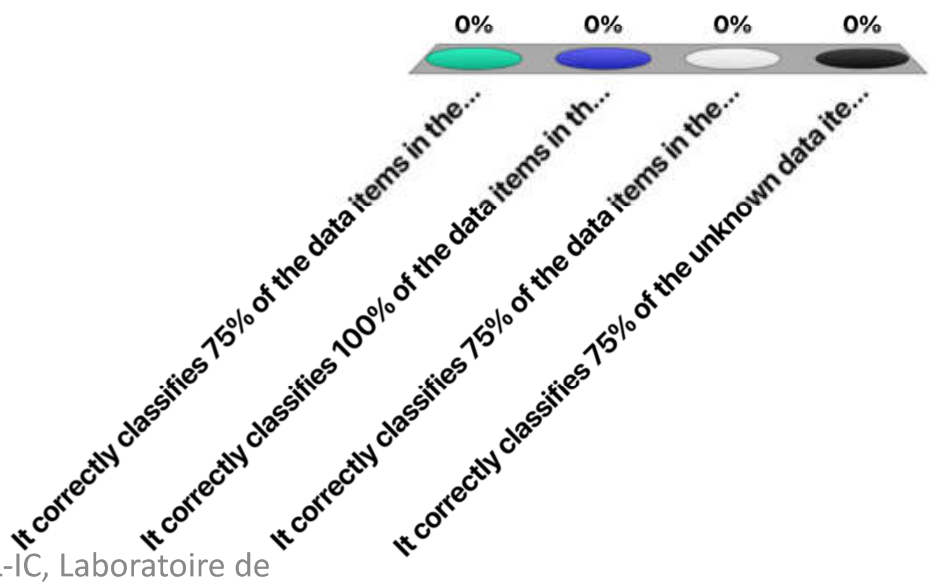
C. the number of clusters is independent of the model parameters

D. the iteration stops once found clusters are no more updated with new points



If a classifier has 75% accuracy, it means that ...

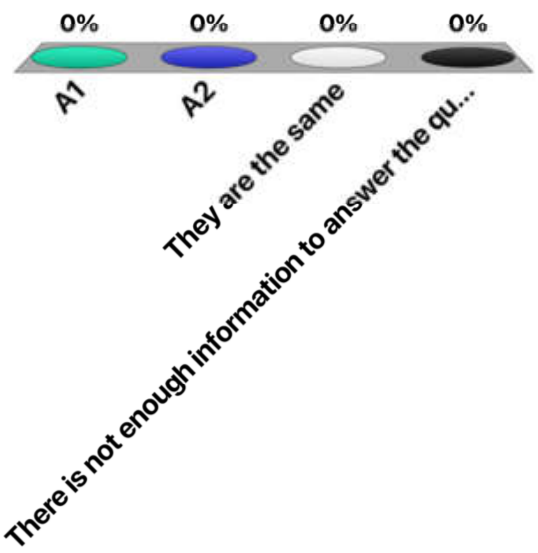
A. It correctly classifies 75% of the data items in the training set
B. It correctly classifies 100% of the data items in the training set but only 75% in the test set
C. It correctly classifies 75% of the data items in the test set
D. It correctly classifies 75% of the unknown data items



Given the distribution of positive and negative samples for attributes $A_1$ and $A_2$, which is the best attribute for splitting?

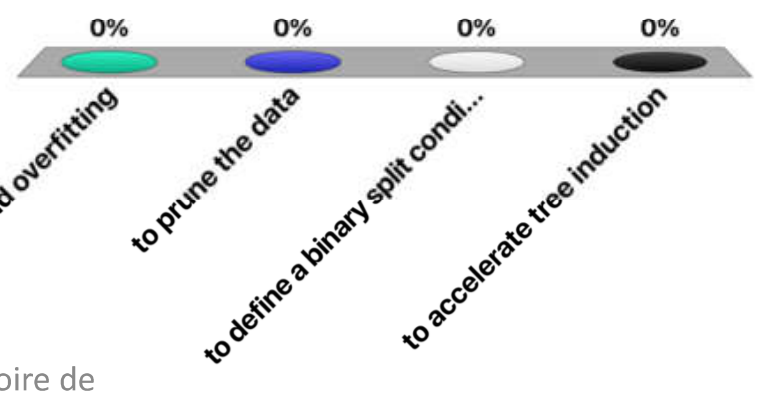
| $A_1$ | P | N |
|---|---|---|
| a | 2 | 2 |
| b | 4 | 0 |
| | | |
| $A_2$ | P | N |
| x | 3 | 1 |
| y | 3 | 1 |

A. A1
B. A2
C. They are the same
D. There is not enough information to answer the question



When splitting a continuous attribute, its values need to be sorted …

A. to avoid overfitting
B. to prune the data
C. to define a binary split condition
D. to accelerate tree induction



The computational cost for constructing a RF with K as compared to constructing K decision trees on the same data

A. is identical
B. is on average larger
C. is on average smaller