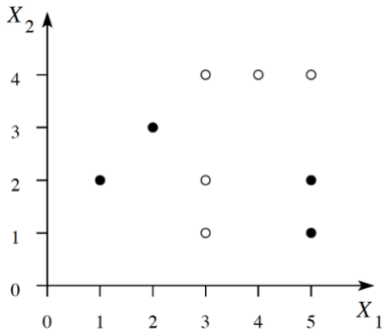
COMP4432 Machine Learning

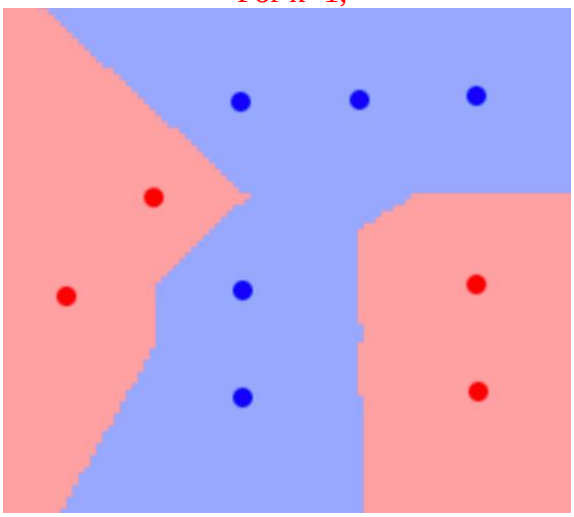
Tutorial Questions on k-Nearest Neighbor

1. For the dataset depicted below, draw the decision boundary formed by using *k*-nearest neighbor classifier when
   (a) *k*=1
   (b) *k*=3.
   Here, only a *rough* drawing (sketch) of the boundary is needed.

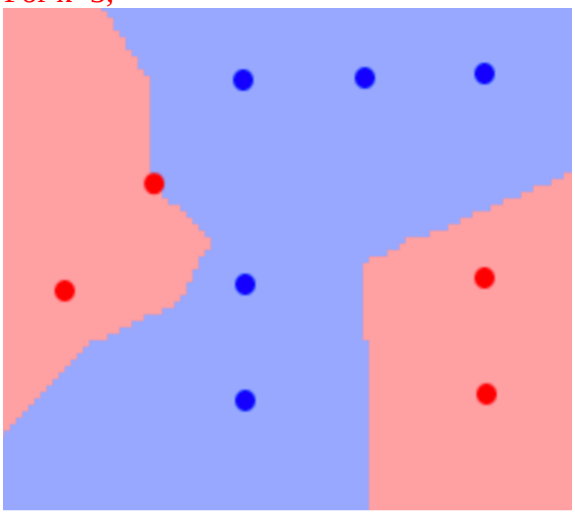


Answer

For k=1,



For k=3,



2. Suppose you are asked to enhance the performance of decision tree by combining it with the *k*-nearest neighbor method. Think about a conceptually sound, potentially workable solution and describe your idea. Comment particularly on the decision boundary of your hybrid model.

Answer

This is an open question and there exists quite a number of acceptable solutions. Some reference approaches are provided below.

This is quite a direct attempt by firstly using DT to coarsely divide the given data and then using kNN to carry out fine classification. A shallow tree, say a 20% grown tree, will be constructed and the data entered into the leaf node will then be applied with kNN classification wrt the data in that node. The decision boundary will be a combination of axis-parallel lines with zig-zag lines formed by kNN. You may refer to the paper by Buttrey and Karo:

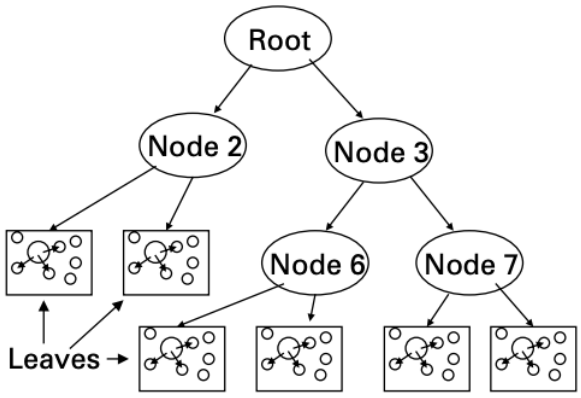


Fig. 1. Schematic of *k*-NN classification inside leaves. The tree is constructed in the usual way. Then a separate *k*-NN classifier is produced in each leaf. Each *k*-NN classifier may have a different *k*, a different set of included variables, and a different choice of scaling.

One potential advantage of this approach is about its speed. Rather than finding the k nearest neighbors from all samples which is very time-consuming when the dataset is big, this approach considers the leaf node of the constructed DT as a group and the nearest neighbor finding process can then be speeded up by considering the group representatives rather than the much larger number of raw samples.