Homework 3 - Solution

Task

Consider the following classifiers:

1. classification tree of the depth 2 (so called "stump")
2. classification tree of the depth 3
3. logistic regression
4. SVM with RBF (radial basis function) kernel and g=1
5. random forest with 100 trees
6. nearest neighbours classifier with number of neighbours set to 5

For each of these classifiers paint:

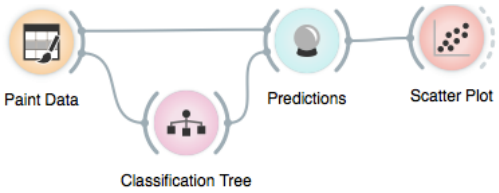
A. a data set where the classifier finds the "right" decision boundary

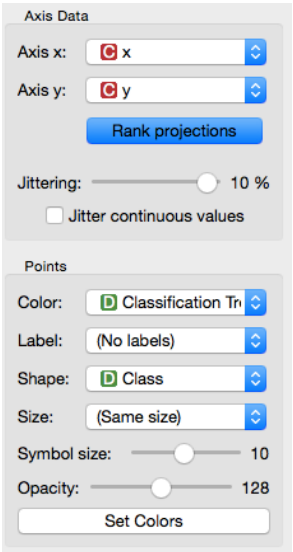
B. a data set where the classifier failed to find the "right" decision boundary

Demonstrate A and B through screen shots of a Scatter Plots.

Solution

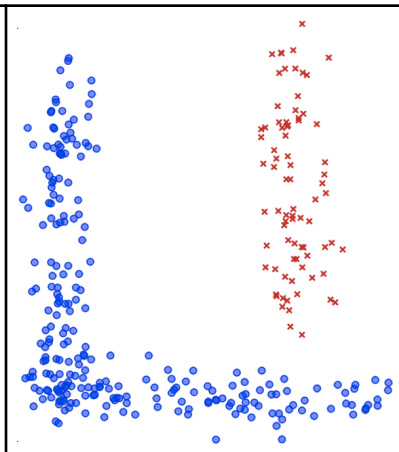
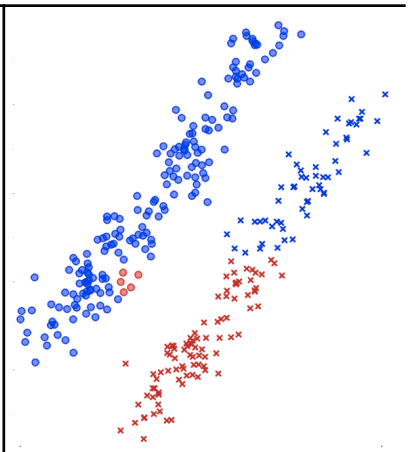
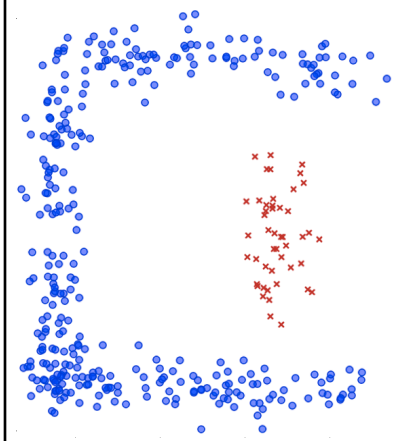
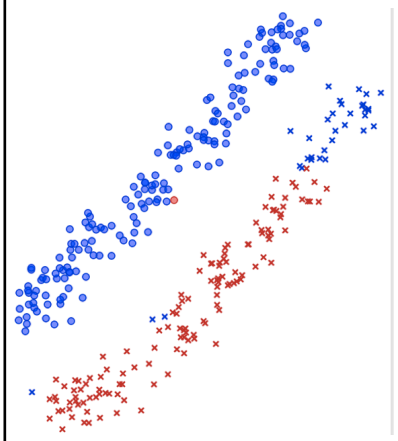
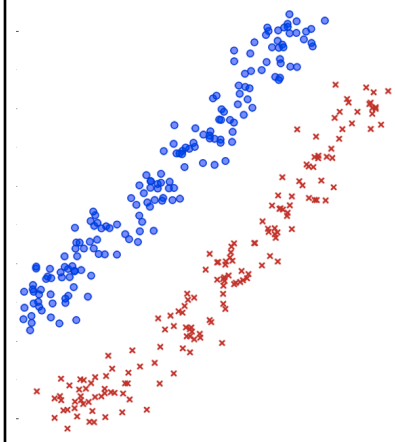
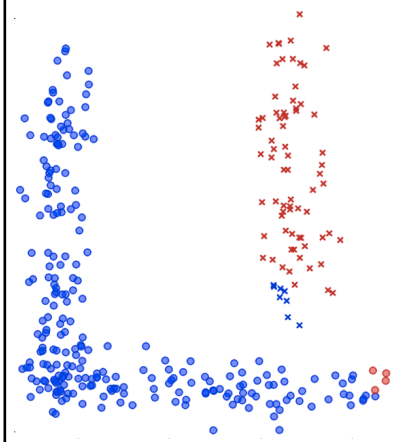
We need this schema. We set the scatter plot to color the points according to the classification and to use shapes for representing the original class. Ideally, we see red crosses and blue circles.

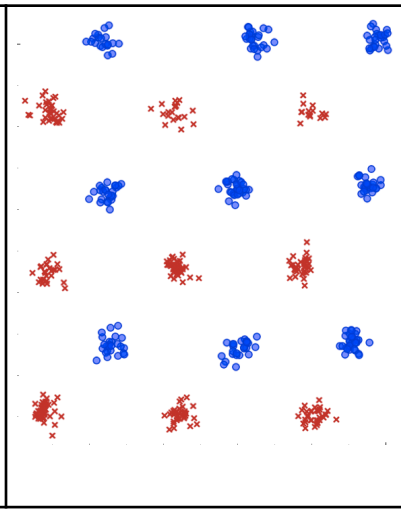
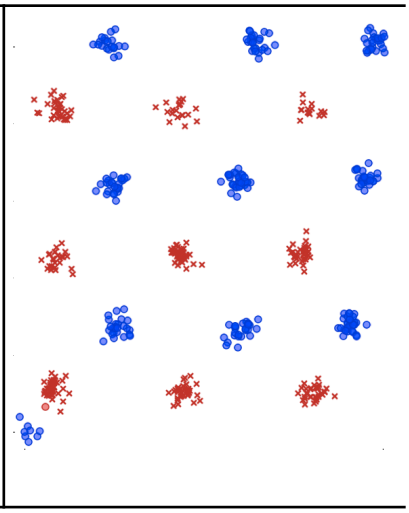




The homework has many possible solutions. Here are a few that demonstrate some interesting cases

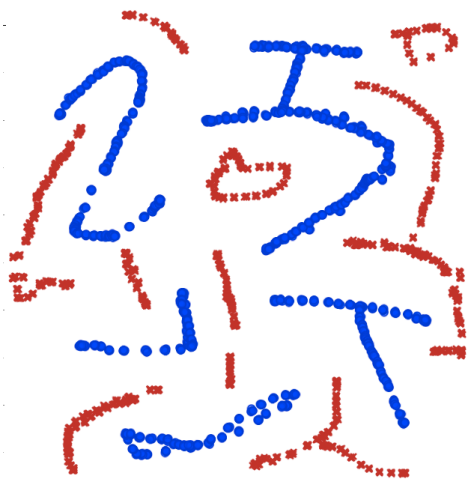
	works	does not work

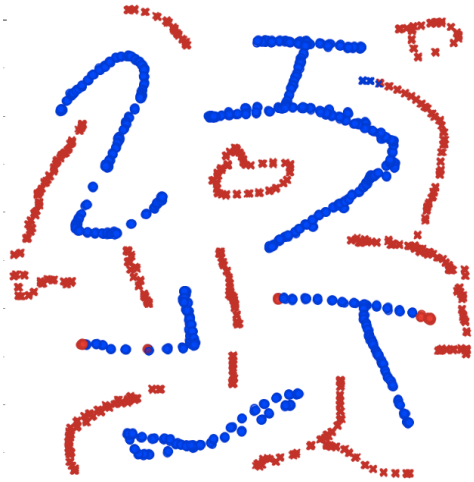
Classification trees of depth 2 will cut the space twice, with both cuts perpendicular to the axes. It can't make a diagonal cuts - not even a single one.		
Classification trees of depth 3 will cut the space three times, with both cuts perpendicular to the axes. It can handle the c-shaped drawing here; the two-level couldn't do it. On the other hand, no tree can (efficiently) handle separation along sloped lines.		
Logistic regression draws a single line at arbitrary slope without a blink. On the other hand, the line it draws is always straight, and it's always a single line. It made a good effort with the picture on the right, though: it found a line that almost separates between the reds and blues.		

SVM with RBF kernels, random forests with 100 trees and k-nearest neighbours will draw complex boundaries that can separate anything. On the right picture, at the bottom left, there is something they do not do: the red circle is so close to the other group that they consider it a noise and put in the other group.		

Whether the model will correctly classify the example in the last picture depends upon the strength of regularization (smoothing) it uses. Stronger the regularization prevent the algorithm from making the model more complicated just to cover a few additional, weirdly located points — it treats such points as noise.

This is easiest to observe with the Nearest neighbors classifier. If the number of neighbors considered when making the prediction is small, the classifier correctly classifies all points. If the number is increased, the points lying closer to the regions covered with the other color are misclassified.





The behavior of the model does not depend only on its type but also on its settings.