Brief Introduction to Algorithmic Data Analysis in English

Esther Galbrun
Autumn 2020
Plots below depict the decision boundary of binary classifiers. Associate each classifier to its decision boundary.

\[ \begin{align*}
    a) & \quad \text{classifier} \\
    b) & \quad \text{classifier} \\
    c) & \quad \text{classifier} \\
    d) & \quad \text{classifier} \\
    e) & \quad \text{classifier}
\end{align*} \]

\( i \) \( k \)-NN \( ii \) decision tree \\
\( iii \) naive Bayes \( iv \) linear SVM \\
v) kernel SVM radial basis function
Q4.2: Precision and recall

Plots below depict the decision boundary of binary classifiers. Dots represent the ground-truth, with the positive class in red. Associate each classifier to its performance on this data.

```
a)  
   b)  
   c)  
   d)  
```

```
Precision  Recall
1       0.5
0.5     1
```

(i)    (ii)    (iii)    (iv)
How much time is necessary to carry out 10-fold cross-validation if the training procedure is quadratic in the number of training instances, whereas the prediction is done in constant time for any given instance, and the available dataset contains $n$ instances?
Consider two classifiers $A$ and $B$. On one data set, a 10-fold cross validation shows that classifier $A$ is better than $B$ by 3%, with a standard deviation of 7% over 100 different folds. On the other data set, classifier $B$ is better than classifier $A$ by 1%, with a standard deviation of 0.1% over 100 different folds. Which classifier would you prefer on the basis of this evidence, and why?
An analyst has trained a decision tree on a dataset. The model has high accuracy on the training data but the accuracy drops sharply on the test data.

In order to improve the performance of the model, you recommend to

a) increase the depth of the tree
b) increase the minimum size of leaves
c) subsample the training data