Introduction to Algorithmic Data Analysis

Esther Galbrun

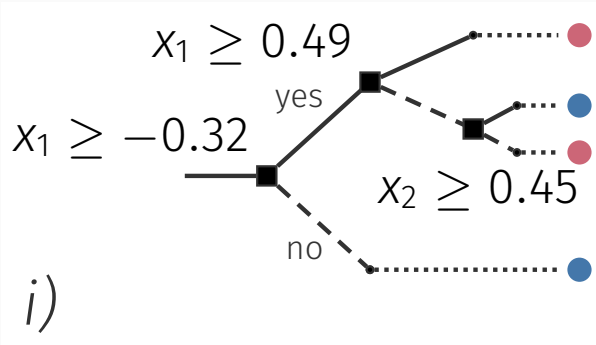
Autumn 2023

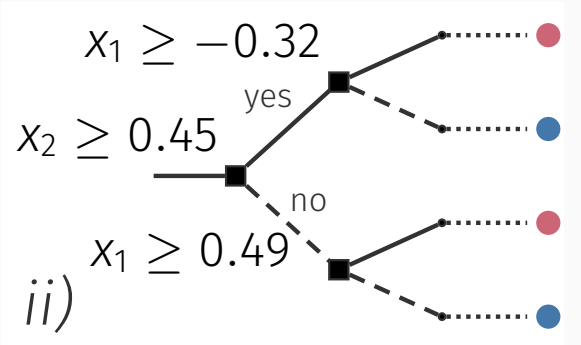
UNIVERSITY OF
EASTERN FINLAND

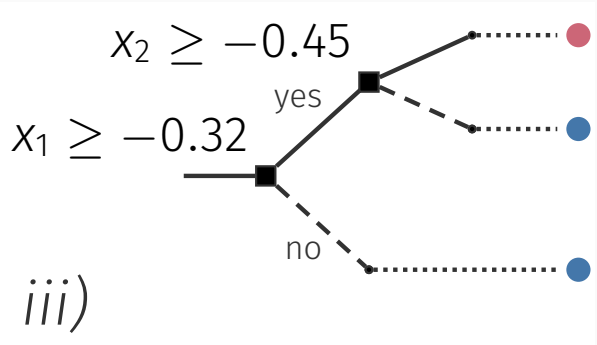
Last updated on 2023-07-04 10:14

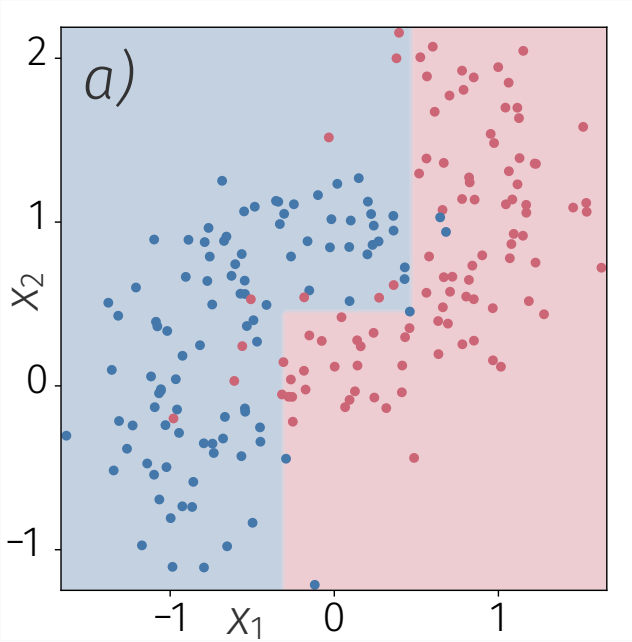
Q3.1: Decision trees

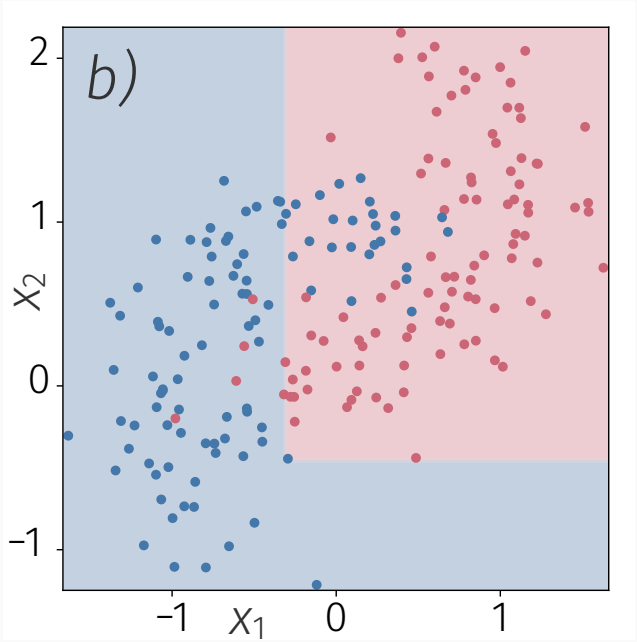
Plots below depict the decision boundary of decision trees.
Associate each tree to its decision boundary.











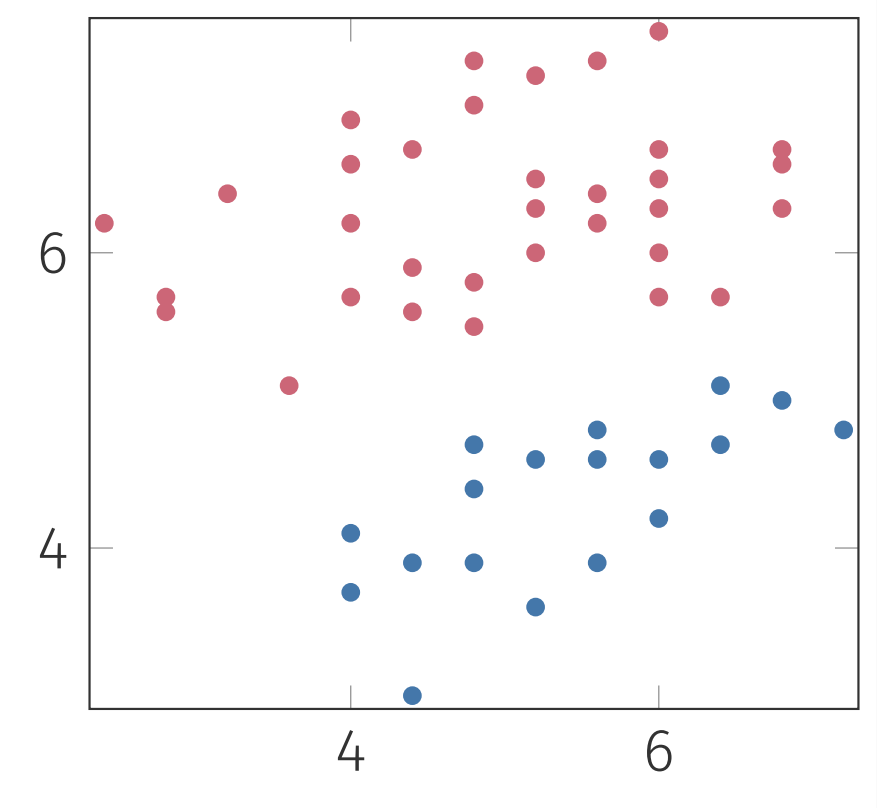
Q3.2: Splitting hairs

While growing a decision tree, we compare two possible splits.
We compute the *error rate*, *Gini index*, *entropy* and *information gain* for either one.
In fact, the four measures agree that the second split is better.
Prior to split, the counts are 8 32 and *entropy*= 0.722.
Can you identify which measure corresponds to which values?

	yes 0 18 no 8 14	yes 6 4 no 2 28
i)	0.200	0.150
ii)	0.202	0.214
iii)	0.520	0.508
iv)	0.255	0.213

Q3.3: Support vector machine of choice (i)

What type of support vector machine best suits this dataset?



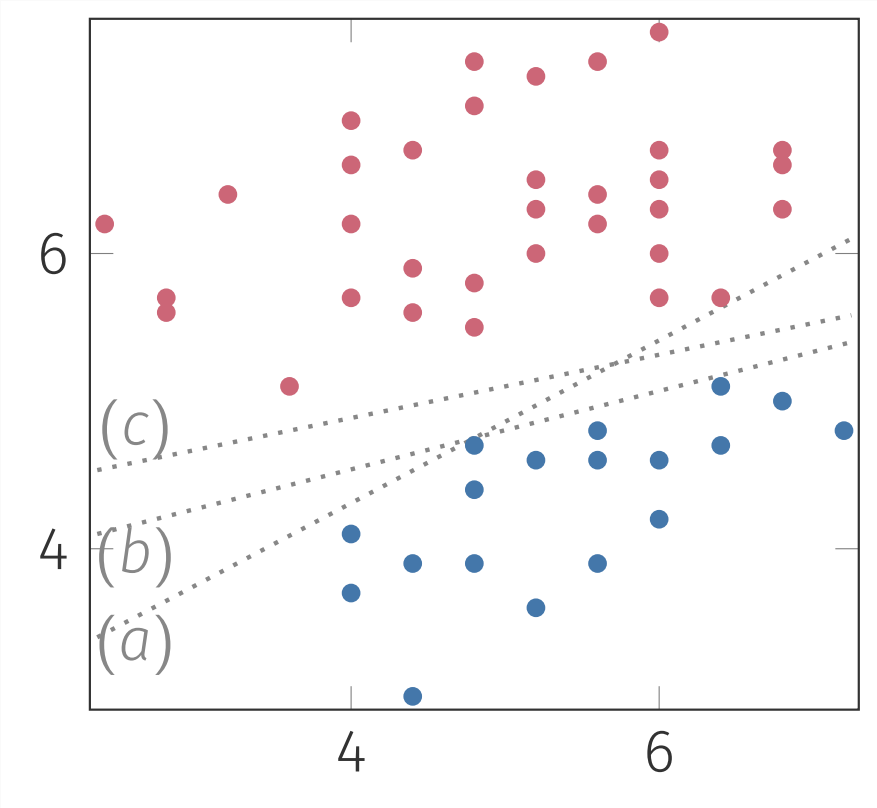
i) hard-margin linear SVM

ii) soft-margin linear SVM

iii) kernel SVM

Q3.4: Split space

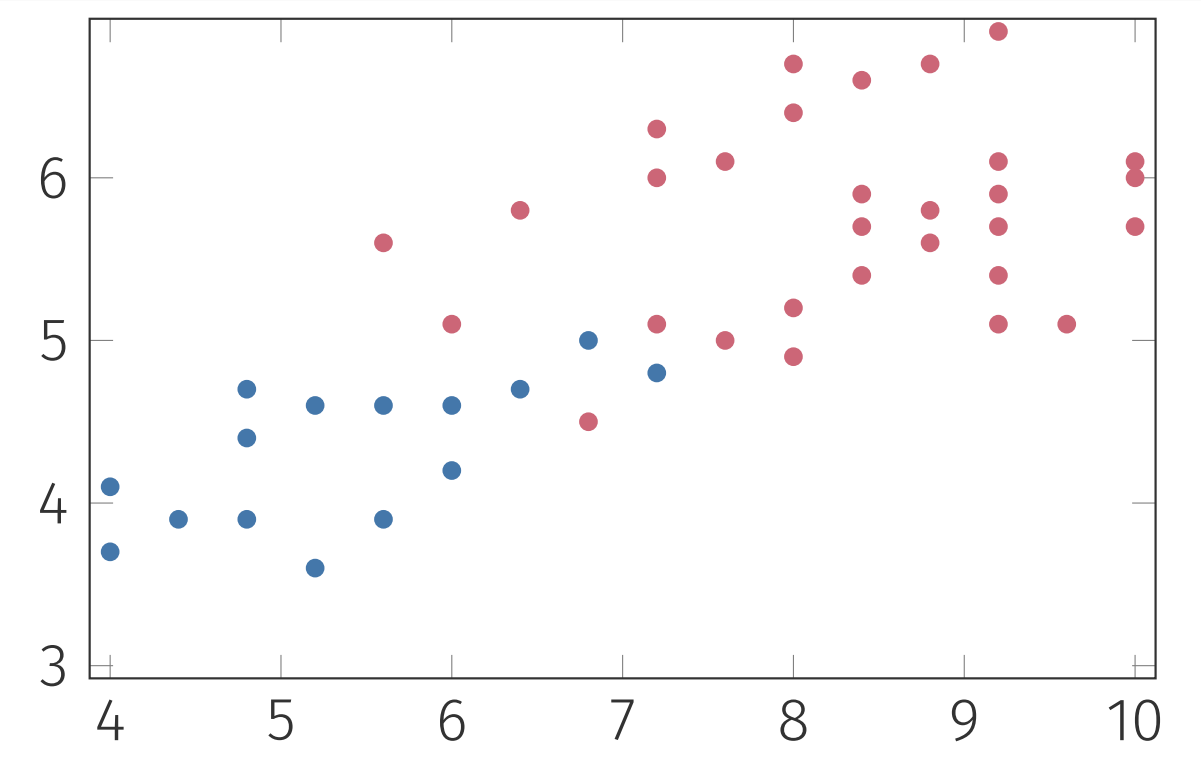
Which of the three lines corresponds to the decision boundary of a hard-margin linear SVM?



Q3.5: Support vector machine of choice (ii)

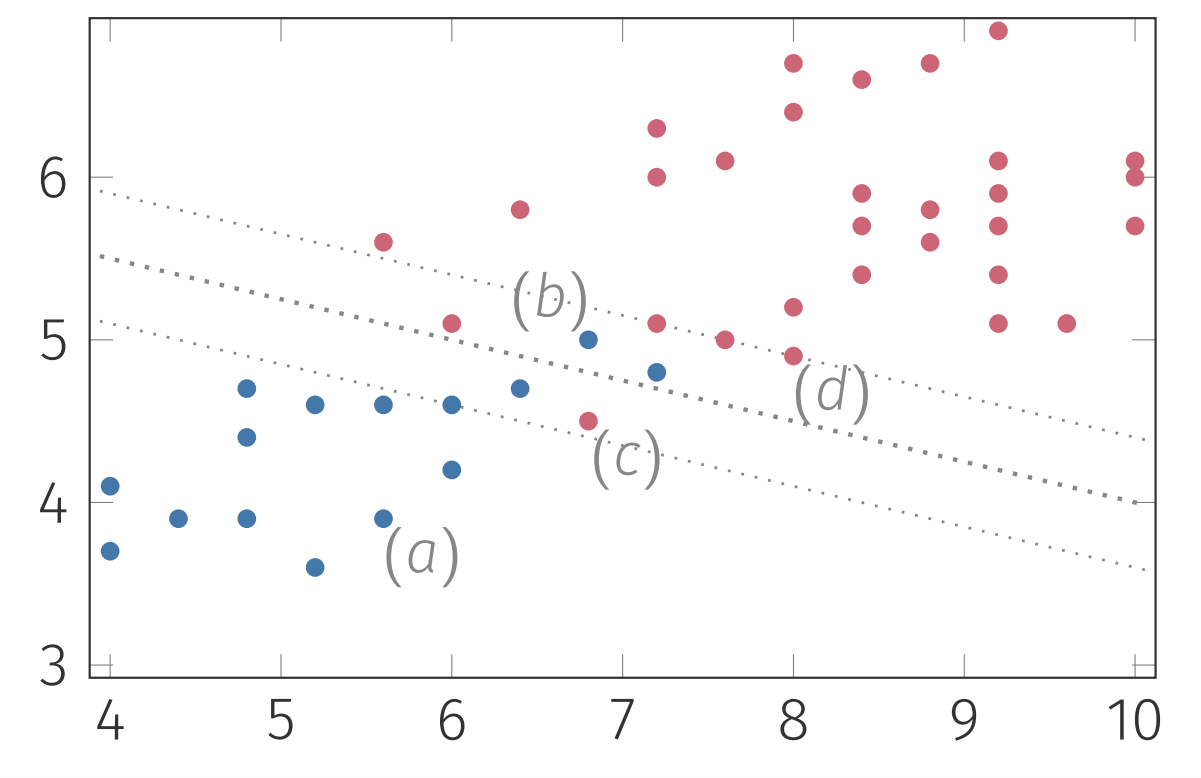
What type of support vector machine best suits this dataset?

i) hard-margin linear SVM
ii) soft-margin linear SVM
iii) kernel SVM



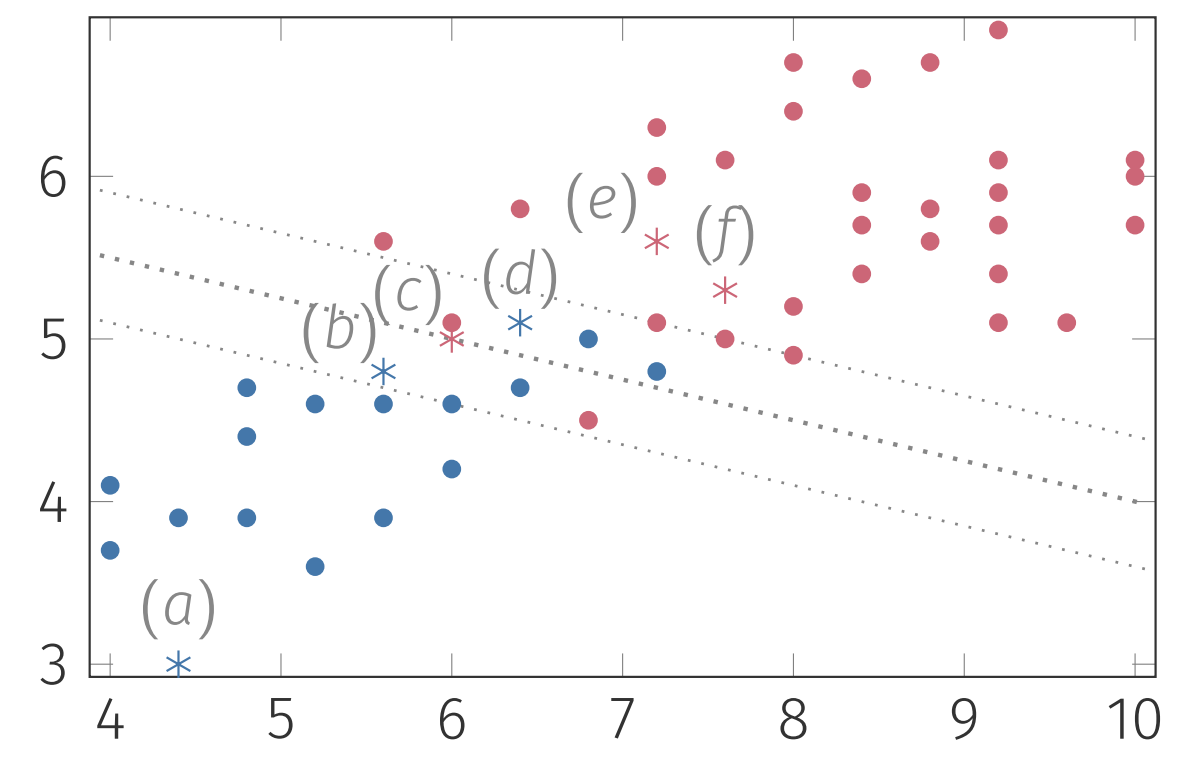
Q3.6: Support vectors

The decision boundary and margin learnt by a soft-margin SVM for this dataset are drawn as gray dotted lines.
Which of the following training data points are support vectors?



Q3.7: Prediction confidence

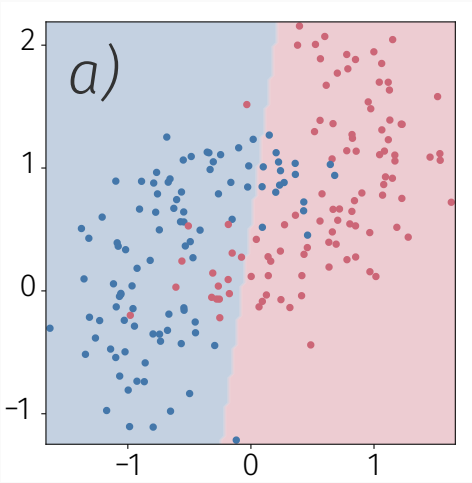
The decision boundary and margin learnt by a soft-margin SVM for this dataset are drawn as gray dotted lines.
Rank the following test data points from the least to the most confident prediction.

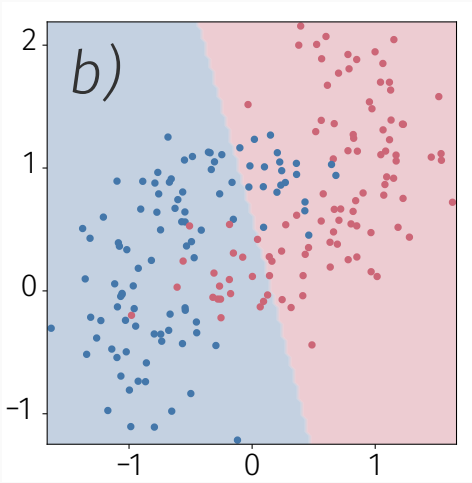


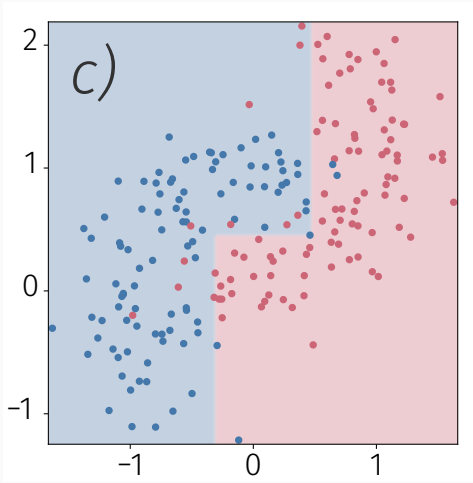
Q3.8: Decision boundaries

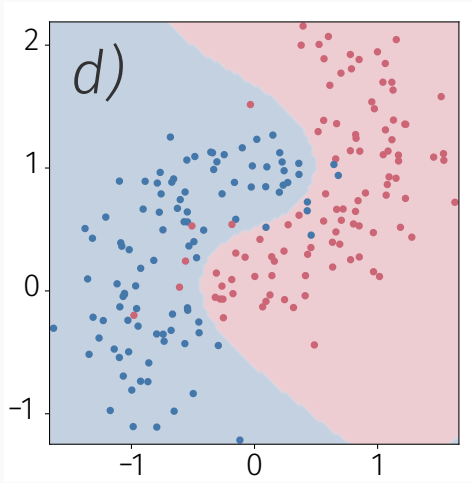
Plots below depict the decision boundary of binary classifiers on the training set.

Associate each classifier to its decision boundary.





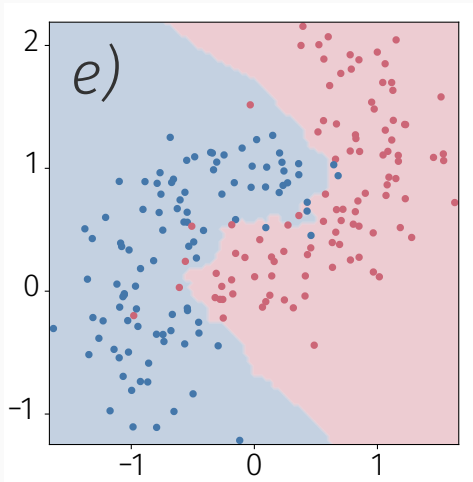




i) *k*-NN *ii)* decision tree

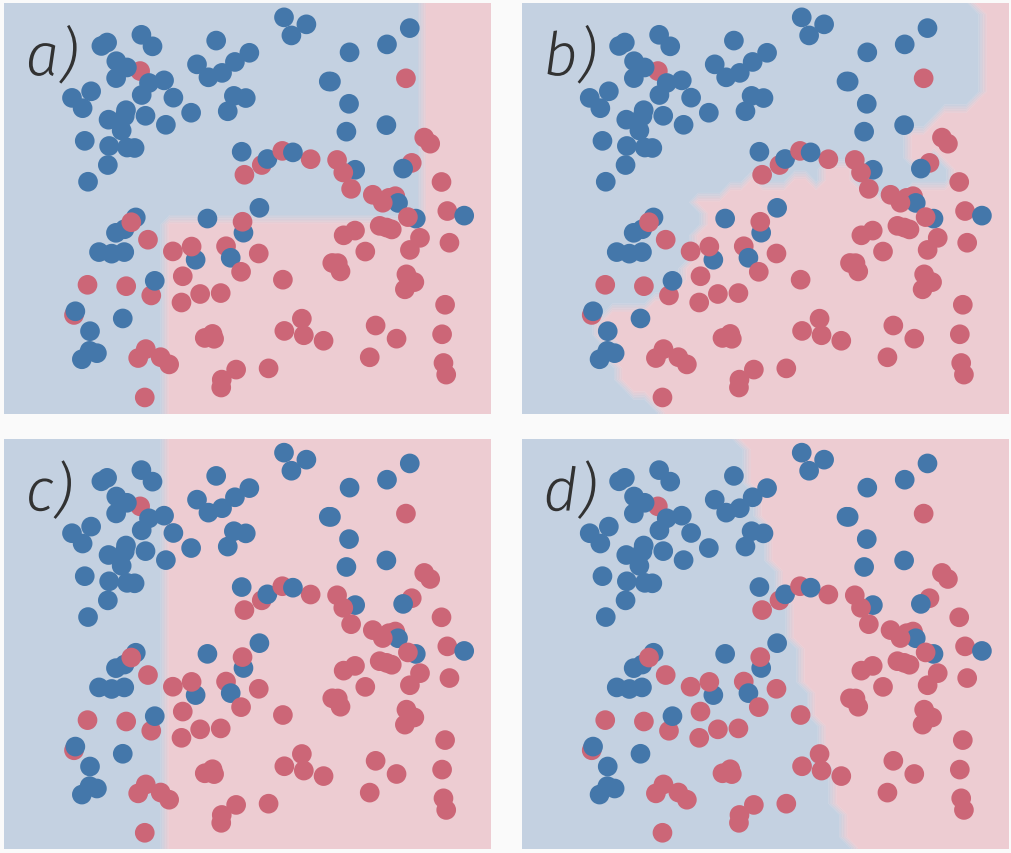
iii) naive Bayes *iv)* linear SVM

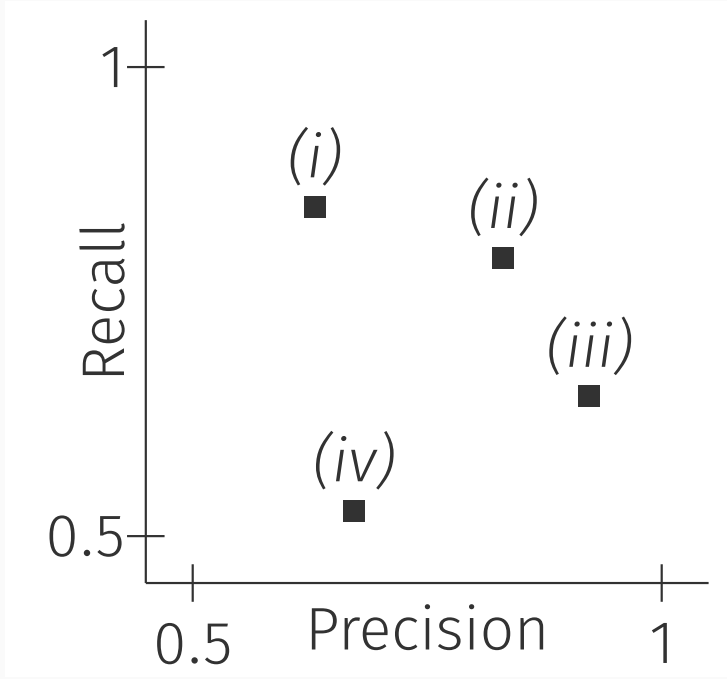
v) kernel SVM radial basis function



Q3.9: 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.





Q3.10: Cross-validation running time

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?

Q3.11: Significantly better

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?

Q3.12: Remedy prescription

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

i) increase the depth of the tree
ii) increase the minimum size of leaves
iii) subsample the training data