Data Mining
(Mining Knowledge from Data)

Model Evaluation

Magda Friedjungová

Classification

- Classification has 2 stages:
  - Learning (training)
  - Recalling (using, applying the model)

Simple Classification

- Data about characteristics of cars
- Classification into two classes:
  - cars from America (origin 1)
  - cars from Europe and Japan (origin 2, 3)
- Create a chart that distributes cars by individual attributes x=mpg and y=weight.
- The colour of car will show the class.

Classification according to one input

- Make classifier which will decide on attribute “mpg” to which of two classes the car belongs.
- Calculate misclassified cars and calculate the percentage of classification success.
- Do the same for the classifier which declares that all cars come from America (belong to class 1).

Classification Rating 1

- Calculate frequencies in individual classes.
- Calculate TP, TN, FP and FN
  - true = correct classification
  - false = incorrect classification
  - positive = car is classified as American car
  - negative = car is classified as non-American car

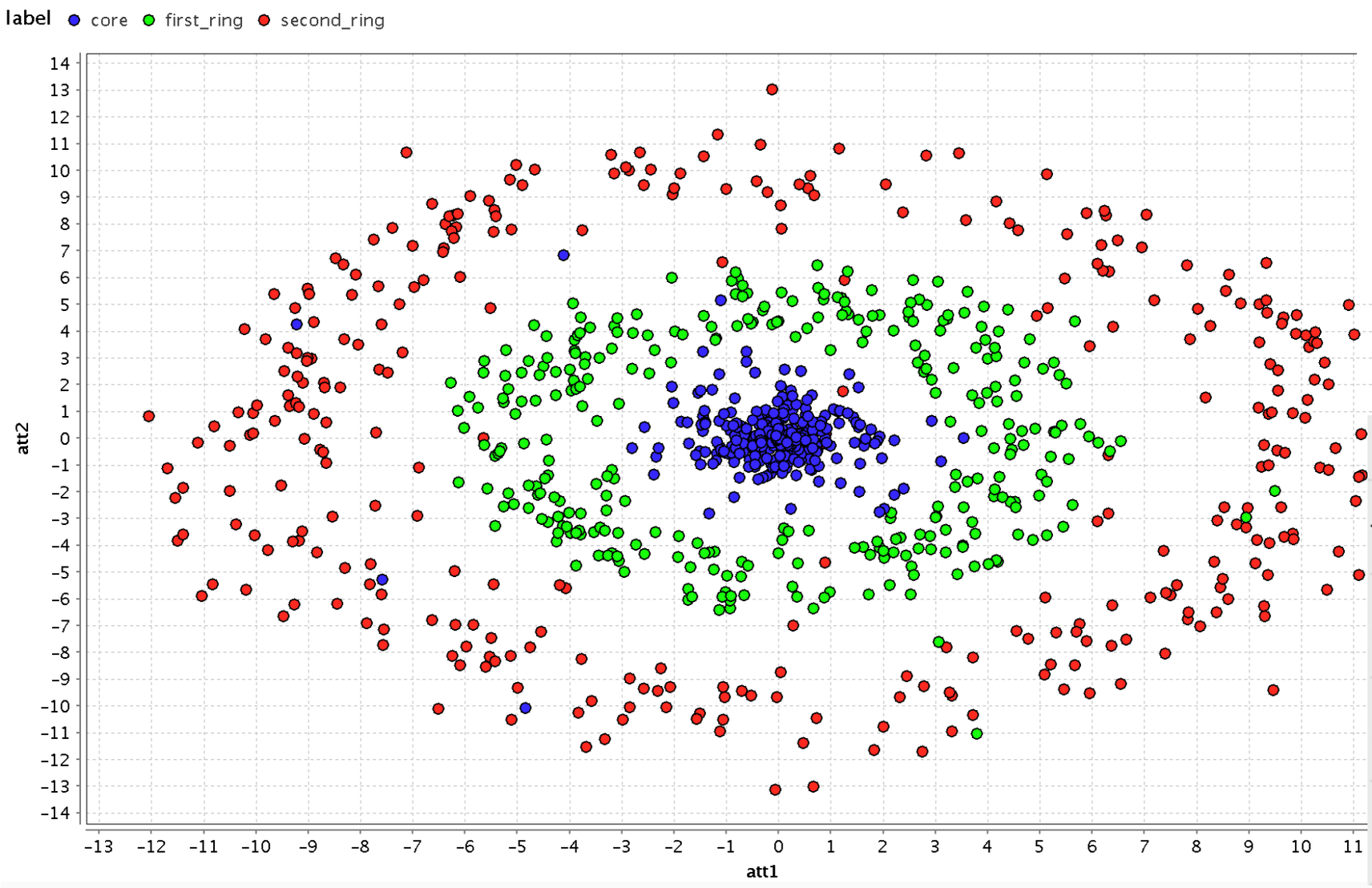
classification \ origin	American car	Non-American car
American car	TP	FP
Non-American car	FN	TN

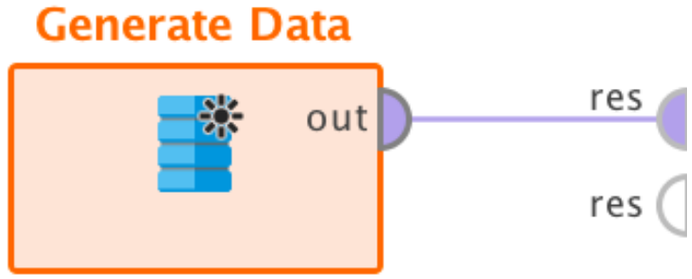
Classification Rating 2

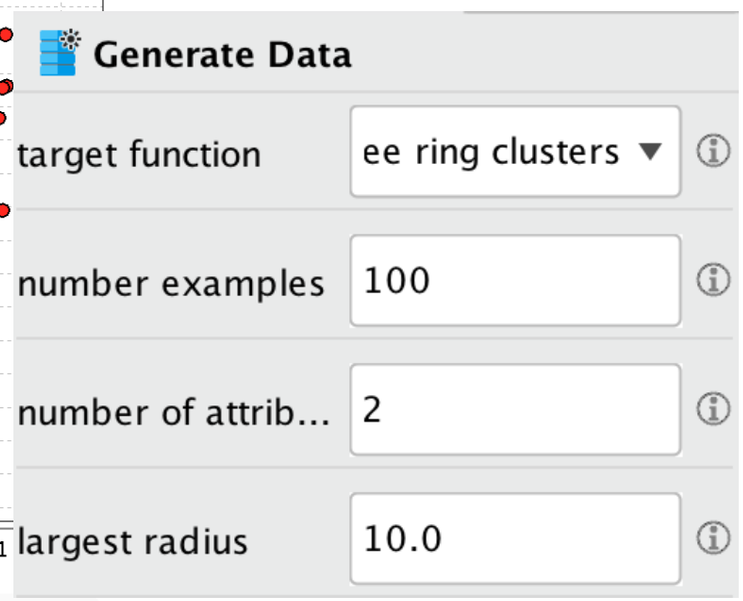
- Calculate the FN rate, FP rate, sensitivity and specificity.
- FP rate = percentage of non-American cars classified as American cars.
- FN rate = percentage of incorrectly classified cars.
- Specificity = probability of correctly classified non-American cars.
- Sensitivity = probability of classification of American car as American car.

The ratio of training and testing data

- Determine the best k (k-NN) for the “three rings” data.







Fix k and change the number of instances

- Change k from 10 to 5000.
- Watch the progress of errors on the training and testing data.

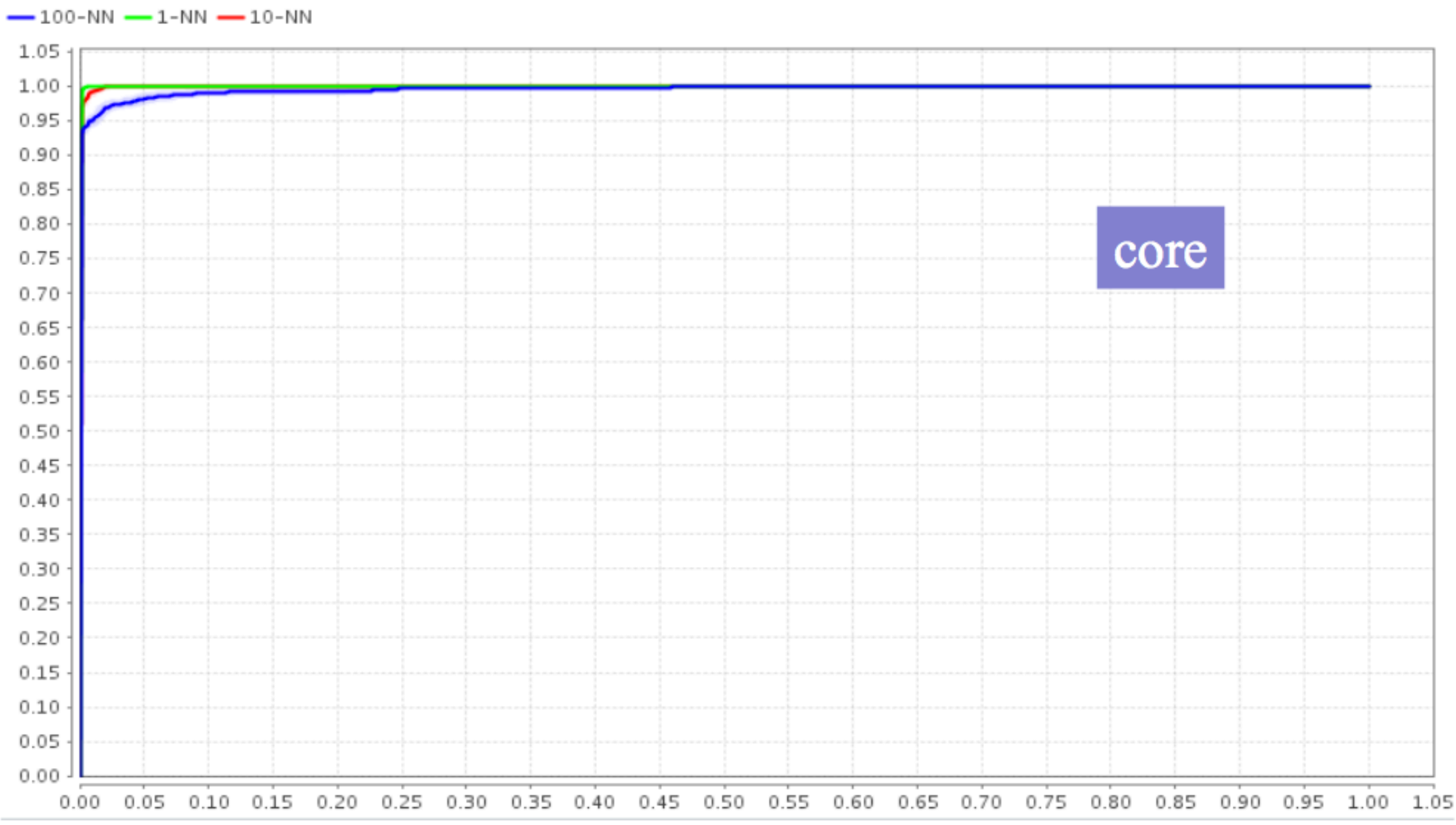
Learning curve

- For the same k and 1000 instances, change the ratio of the number of instances of training and test sets from 0.01 to 0.99.
- Repeat the experiment 100 times and plot the graph for all values.
- Interpret the result.

What k is optimal?

- Determine the best k (k-NN) for the “three rings” data.
- Try to use “x-validation” instead of “split validation”.
- Try to use “bootstrap validation” 100 times.
- Interpret the confusion matrix.

Calculate the ROC for the best algorithms and compare them



Do it for other classes as well

- First and second ring