Intelligent Data Analysis Summer 2014
C. Braune

Exercise Sheet 6

Exercise 20 Applications of Bayes Classifiers

For classifying emails as *spam* or *not spam* a naive bayes classifier can be used.
Answer the following questions:

1. Which attributes could be used for building a classifier? Which assumptions about the attributes do not hold?
2. You know, that emails are *spam* with a likelihood of 0.8. What additional a priori knowledge is needed for building a classifier? How can you obtain it? Do you (or the algorithm) need to understand the language the email was written in?
3. Decide whether the following email is *spam* or *not spam*:

   Dear Mr. Stonemeyr,

   we should urgently speak about the important topic of salary raises. Please meet me in my office tomorrow - urgently!

   With best regards,
   Mrs. Mörkl

   Hint: You know (at least now) that in real mails the following words are five times more likely to occur: Dear, Stonemeyr, boss, we, my, Mörkl.

   You also know that spam mail contain the following words five times more often than real mails: salary, important, urgently, raises, tomorrow.

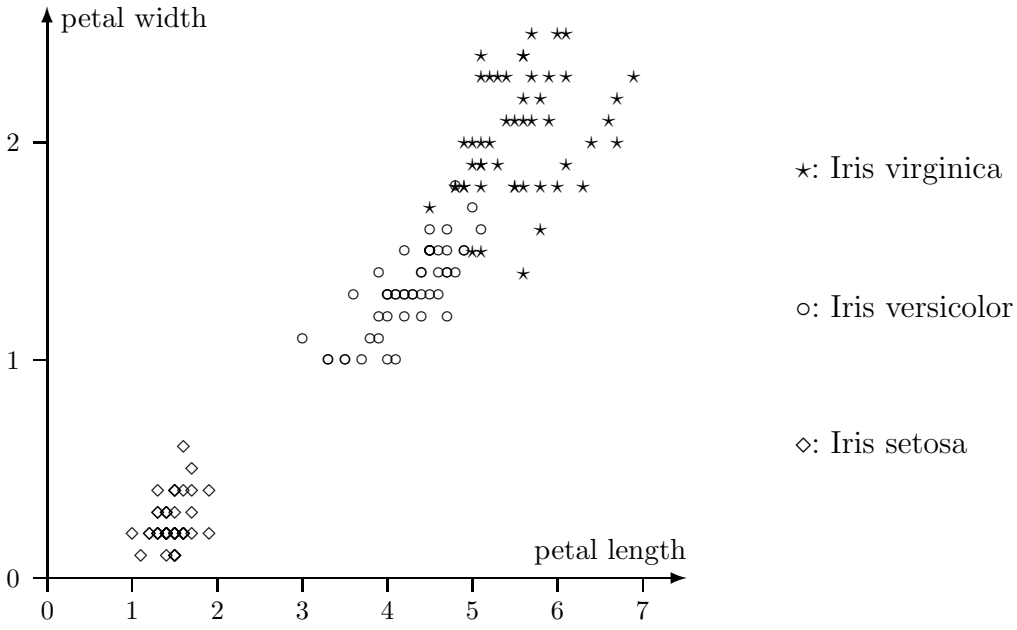
   All other words can be assumed equally likely in both types of mails.

Exercise 21 Decision Trees: Visualization

Consider the following decision tree for the Iris data (see the data set and the decision tree induction program that are available on the lecture page). The Iris data describe iris flowers by stating the petal length and width (in cm), the sepal length and width (in cm), and to which of the three types (classes) Iris setosa, Iris virginica und Iris versicolor each flower belongs.

```
dtree(iris_type) =
{ (petal_length|2.45)
  <:{ Iris-setosa: 50 },
  >:{ (petal_width|1.75)
      <:{ (petal_length|4.95)
          <:{ Iris-versicolor: 47, Iris-virginica: 1 },
          >:{ (petal_width|1.55)
              <:{ Iris-virginica: 3 },
              >:{ Iris-versicolor: 2, Iris-virginica: 1 }}},
      >:{ Iris-versicolor: 1, Iris-virginica: 45 }}};
```

Visualize this decision tree by drawing the regions, in which the different classes are predicted, into the following diagram:



Compare the result with the classification of the same data set with a naive or full Bayes classifier (see the lecture slides for such a visualization)!

Exercise 22 Induction of Decision Trees

In analogy to the example discussed in the lecture (prediction of a drug for a patient) to induce a decision tree for the following data set, which predicts the value of the attribute "Class", which states whether it is recommendable to play golf under the specified weather conditions. (If two or more attributes have the same quality, choose among them the attribute that is listed farthest to the left in the table.)

Outlook	Temp. (^oF)	Humidity (%)	Windy?	Class
sunny	85	85	false	Don't Play
sunny	80	90	true	Don't Play
overcast	83	78	false	Play
rain	70	96	false	Play
rain	68	80	false	Play
rain	65	70	true	Don't Play
overcast	64	65	true	Play
sunny	72	95	false	Don't Play
sunny	69	70	false	Play
rain	75	80	false	Play
sunny	75	70	true	Play
overcast	72	90	true	Play
overcast	81	75	false	Play
rain	71	80	true	Don't Play

Exercise 23 Induction of Decision Trees

A_1	F	T	T	T	F	F	F	T
A_2	F	T	F	F	T	F	T	T
A_3	F	F	F	T	T	T	T	F
C	F	F	T	T	T	F	T	F

Consider the data set shown on the right, which comprises three binary attributes A_1, A_2, A_3 and one (also binary) class attribute C ("T" means "true", "F" means "false"). If you process this data set with the procedure that was presented in the lecture, the resulting decision tree is more complex than necessary. (What would be a better, that is, simpler decision tree?) Which problem of greedy attribute selection becomes obvious with this example?