Decision Trees

Einführung in die Computerlinguistik – Übung 04.12.2015

Decision trees – Christmas travel plans?

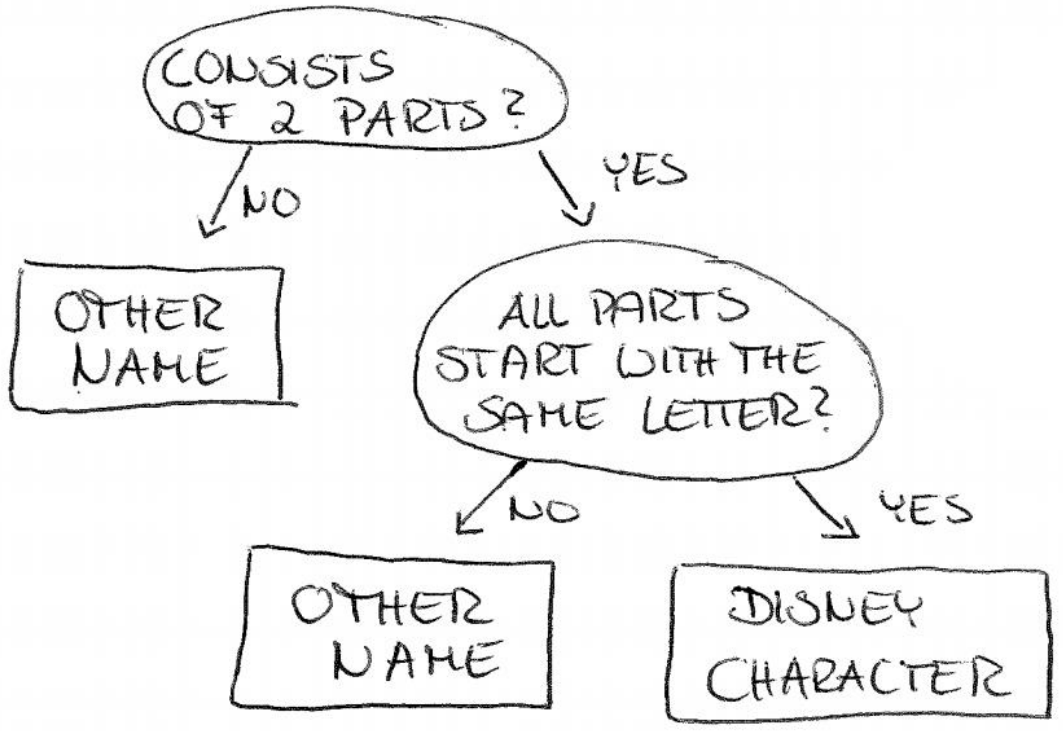
1. You want to create a decision tree in order to decide if you want to travel for Christmas or stay at home with your family.
    1. Name some possible features. Money, loves to travel, has company,…
    2. Suppose you use the following samples of friends for your decisions. Create a good decision tree! First: money or time

	Has money	Has free time	Loves family	Travel?
Jens	Yes	Yes	Yes	Yes
Anna	No	Yes	Yes	No
Jenny	No	No	Yes	No
Dennis	No	Yes	Yes	yes

    3. Is it always possible to get a decision with your tree? No, there is no clear solution; use probabilities!

Decision trees – Usage

1. Use the following decision tree to classify the names 1-5 as “Disney character” or “other name”.
   1. Mickey Mouse
   2. Angela Merkel
   3. Donald Duck
   4. Max Mustermann
   5. CIS



2. Which decision(s) is/are wrong?

   Max Mustermann

Write

1. Open console (Use MobaXterm on Windows)
2. Only if you are on your private computer
   ```
   ssh username@remote.cip.ifi.lmu.de
   ```
3. Login to my machine
   ```
   ssh tbd
   ```
4. Enable chat
   ```
   mesg y
   ```
5. Send answers or questions
   ```
   echo “bla bla" | write kannk
   ```

Decision trees – Code

1. Download “dtree.tar“ from the course page.

2. Go to the folder and unpack the archive.

3. Run:

```
python dtreebasic.py
```

Decision trees – Code

4. Look at the output. What does the program do?

   A decision tree to decide if the word is English or German

5. What is the meaning of “e”, what is the meaning of “g” in the output?

   “English” and “German”

6. Think of new features for the classification.

   Contains “sch”, contains “th”,…

Decision trees – Code

7. Modify the script such that the decision tree gets built with your features. Which accuracy do you get?
   `Hint: Look at lang_features.`

8. First without implementing it, which accuracy do you expect when training on the development set?

9. Now test your hypothesis. Have you been right?
   It can be 100 or less; if less it is because not everything can be clearly classified and probabilities are used