Technische Hochschule Rosenheim
Prof. Johannes Jurgovsky

Exercise Sheet

Learning Goals

- Decision Trees
- Ensemble Learning
- ID3 Algorithm

1. Give a high-level description of the three ensemble learning variants discussed in the lecture.
   (a) Model Stacking
   (b) Boosting
   (c) Bagging
2. Describe the conceptual relationship between Bagging and Random Forests.
3. Working for a car-insurance company, your task is to predict the risk-class of a driver (applicant for an insurance contract) based on the following features:

   - License: Possession of driver's license (1-2 years, 2-7 years, >7 years)
   - Gender: male or female
   - Region: city or countryside

   You have the following data available for training your classifier.

client	License	Gender	Region	Risk
1	01. Feb	m	city	low
2	02. Jul	m	countryside	high
3	> 7	f	countryside	low
4	01. Feb	f	countryside	high
5	> 7	m	countryside	high
6	01. Feb	m	countryside	high
7	02. Jul	f	city	low
8	02. Jul	m	city	low

   (a) Explain why splitting on `client` has the highest Information Gain, so it looks like the perfect split, but why it still is the worst split possible.
   (b) Construct a decision tree based on the training data, using information gain as split strategy.
   Use the following notation:

   - Dataset D; number of classes C; attribute A with k different values
   - Entropy $ent(D) = -\sum_{c=1}^{C} p_c \log_2 p_c$
   - Conditional entropy $ent(D, A) = \sum_{i=1}^{k} \frac{|D_i|}{|D|} ent(D_i)$
   - Information gain $IG(D, A) = ent(D) - ent(D, A)$

   What problem do you encounter when splitting on `License`? How would you solve this?

4. Imagine some data described by two continuous attributes x_1 and x_2 varying between 0 and 1 and two class labels '+' and '-'. Draw a dataset where a decision tree using "value > number"- splits needs to split on x_1 multiple times to achieve a good result. Which one of the two decision tree algorithms is capable of representing such a split?