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

   **Solution: Model stacking:**

   - Use a collection of arbitrary learners (preferably different ones)
   - Fit all learners on the training data
   - Fit a "combiner" model on the training data using all the predictions of the other models as additional inputs.
   - The "combiner" is often a single-layer Logistic Regression classifier.

   (b) Boosting

   **Solution:**
   **Boosting:**

   - Use a collection of weak learners (high-bias/low-variance) such as very shallow decision tress (Decision Stumps)
   - Train the learners sequentially
   - Let the current learner focus on those data points, that the earlier learners got wrong (*the hard data points*).

   (c) Bagging

   **Solution: Bagging (Bootstrap Aggregating):**

   - Used to combine a collection of high-variance/low-bias predictors into a single predictor with less variance while keeping low-bias
   - Draw new training data sets from the training data
   - Fit one learner on each subset
   - Aggregate the predictions of the learners (e.g. plurality vote in classification, weighted mean in regression)
   - Popular algorithm: Random Forests

2. Describe the conceptual relationship between Bagging and Random Forests.

> **Solution:** Bagging is a ensemble learning strategy that creates $T$ training datasets $D_t$ of size $n$ from a training dataset $D$ of size $n$ by sampling from $D$ with replacement. On each of the bootstrap samples $D_t$ a model is fitted. The predictions of the $T$ models are then aggregated using plurality voting for classification or by computing the mean for regression tasks.
>
> A random forest builds on this idea using the following setting:
>
> - All models are decision trees
> - During split evaluation at each node, a subset of $m$ attributes is selected randomly. Only these $m$ attributes are evaluated as possible split candidates.

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.

> **Solution:** If we split on client, all obtained successor nodes would be pure. They contain only a single example of one class. However, when we would use such a decision tree for predicting the risk level of a new client there are two major problems:
>
> - The client id and the risk level are (hopefully) completely unrelated. The prediction of risk level for a client based on the client's id works only for this exact client from the training set. The classifier simply remembers the training dataset. Such a decision rule is worthless for any new client that we haven't seen before. -¿ The classifier does not generalize at all.

- Estimating the class probability based on a single observation in the leaf node is highly unreliable.

(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?

**Solution:** .

3b)

$$\mathrm{ent}(D) = -\Big(\underbrace{\tfrac{1}{2}}_{p(\text{risk}=\text{low})}\log_2 \underbrace{\tfrac{1}{2}}_{} + \underbrace{\tfrac{1}{2}}_{p(\text{risk}=\text{high})}\log_2 \underbrace{\tfrac{1}{2}}_{}\Big) = 1 \quad ; \; |D| = 8$$

$\boxed{\text{RECURSION 1:}}$

- **IG(D, License):**
  - License = "0 1. Feb" $\rightarrow \overbrace{\{1,4,6\}}^{\text{clients}}$ $\quad |D_{\{1,4,6\}}| = 3$

    $\mathrm{ent}(D_{\{1,4,6\}}) = -\left(\frac{1}{3}\log_2\frac{1}{3} + \frac{2}{3}\log_2\frac{2}{3}\right) \approx 0{,}92$
  - License = "02. Jul" $\rightarrow \{2,7,8\}$ $\quad |D_{\{2,7,8\}}| = 3$

    $\mathrm{ent}(D_{\{2,7,8\}}) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) \approx 0{,}92$
  - License = ">7" $\rightarrow \{3,5\}$ $\quad |D_{\{3,5\}}| = 2$

    $\mathrm{ent}(D_{\{3,5\}}) = -\left(\frac{1}{2}\log_2\frac{1}{2} + \frac{1}{2}\log_2\frac{1}{2}\right) = 1$

  $\hookrightarrow$ $\mathrm{IG}(D, \mathrm{License}) = 1 - \mathrm{ent}(D, \mathrm{License})$

  $= 1 - \left(\frac{3}{8}\mathrm{ent}(D_{\{1,4,6\}}) + \frac{3}{8}\mathrm{ent}(D_{\{2,7,8\}}) + \frac{2}{8}\mathrm{ent}(D_{\{3,5\}})\right)$

  $\approx 0.06$

- **IG(D, Gender):**
  - Gender = "m" $\rightarrow \{1,2,5,6,8\}$ $\quad |D_{\{1,2,5,6,8\}}| = 5$

    $\mathrm{ent}(D_{\{1,2,5,6,8\}}) = -\left(\frac{2}{5}\log_2\frac{2}{5} + \frac{3}{5}\log_2\frac{3}{5}\right) \approx 0{,}97$
  - Gender = "f" $\rightarrow \{3,4,7\}$ $\quad |D_{\{3,4,7\}}| = 3$

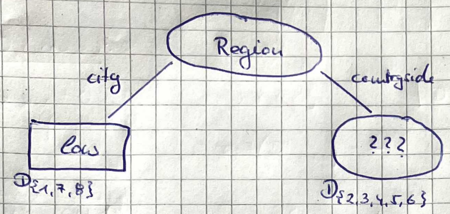
    $\mathrm{ent}(D_{\{3,4,7\}}) = -\left(\frac{2}{3}\log_2\frac{2}{3} + \frac{1}{3}\log_2\frac{1}{3}\right) \approx 0{,}92$

  $\hookrightarrow$ $\mathrm{IG}(D, \mathrm{Gender}) = 1 - \left(\frac{5}{8}\mathrm{ent}(D_{\{1,2,5,6,8\}}) + \frac{3}{8}\mathrm{ent}(D_{\{3,4,7\}})\right)$

  $\approx 0.05$

- **IG(D, Region):**
  - Region = "city" $\rightarrow \{1,7,8\}$, $|D_{\{1,7,8\}}| = 3$, $\mathrm{ent}(D_{\{1,7,8\}}) = 0$
  - Region = "countryside" $\rightarrow \{2,3,4,5,6\}$, $|D_{\{2,3,4,5,6\}}| = 5$, $\mathrm{ent}(D_{\{2,3,4,5,6\}}) \approx 0{,}722$

  $\hookrightarrow$ $\mathrm{IG}(D, \mathrm{Region}) = 1 - \left(\frac{3}{8}\cdot 0 + \frac{5}{8}\mathrm{ent}(D_{\{2,3,4,5,6\}})\right) \approx 0{,}55$




$|D_{\{2,3,4,5,6\}}| = 5$

$ent(D_{\{2,3,4,5,6\}}) \approx 0{,}722$

**RECURSION 2:**

- $IG(D_{\{2,3,4,5,6\}}, License)$:
  - License = "~~07.Jul~~ 01.Feb" $\to \{4,6\}$ $ent(D_{\{4,6\}}) = 0$
  - License = "02.Jul" $\to \{2\}$ $ent(D_{\{2\}}) = 0$
  - License = ">7" $\to \{3,5\}$ $ent(D_{\{3,5\}}) = 1$

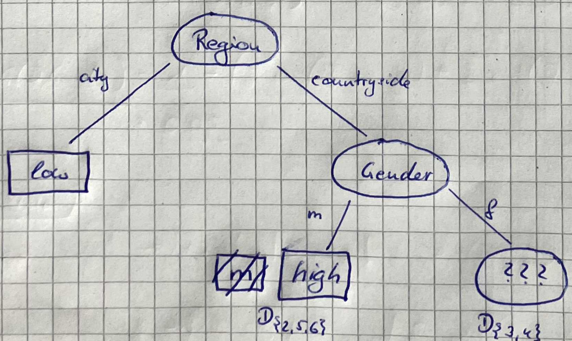
  ↳ $IG(D_{\{2,3,4,5,6\}}, License) = 0{,}722 - (0 + 0 + \frac{2}{5} \cdot 1) = 0{,}322$

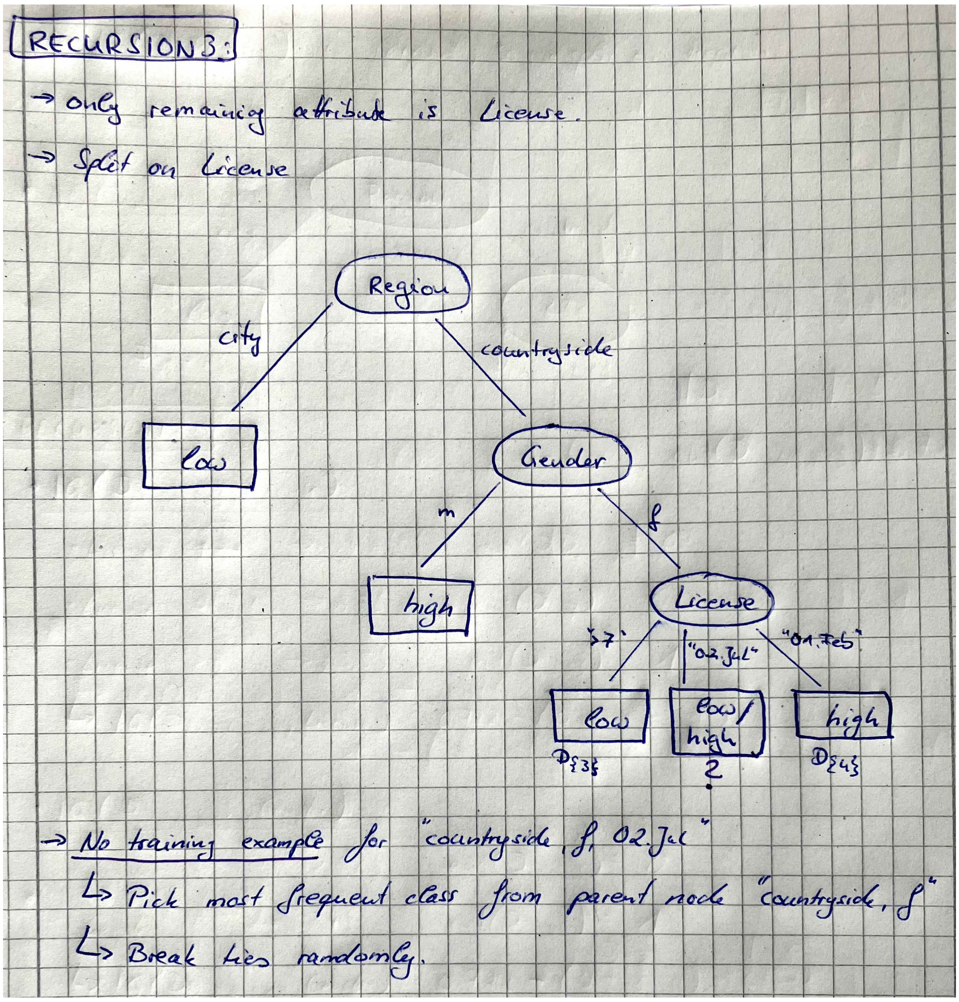
- $IG(D_{\{2,3,4,5,6\}}, Gender)$:
  - Gender = "m" $\to \{2,5,6\}$ $ent(D_{\{2,5,6\}}) = 0$
  - Gender = "f" $\to \{3,4\}$ $ent(D_{\{3,4\}}) = 1$

  ↳ $IG(D_{\{2,3,4,5,6\}}, Gender) = 0.722 - (0 + \frac{2}{5} \cdot 1) = 0.322$

→ $IG(D_{\{2,3,4,5,6\}}, License) = IG(D_{\{2,3,4,5,6\}}, Gender)$ → randomly pick one.

→ Let's use Gender






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?

**Solution:** Any dataset that has at least three points on a line parallel to the coordinate axes where the middle point is from the opposite class.

The CART algorithm can handle continuous data.