Machine Learning Exercise (SS 23)

Assignment 06: Decision Tree (Solution)

Mojtaba Nayyeri
Mojtaba.Nayyeri@ipvs.uni-stuttgart.de

Akram Sadat Hosseini
Akram.Hosseini@ipvs.uni-stuttgart.de

Nadeen Fathallah
Nadeen.Fathallah@ipvs.uni-stuttgart.de

Rodrigo Lopez Portillo Alcocer
rodrigo.lopez-portillo-alcocer@ipvs.uni-stuttgart.de

Tim Schneider
timphillip.schneider@ipvs.uni-stuttgart.de

Osama Mohammed
osama.mohammed@ipvs.uni-stuttgart.de

Daniel Frank
daniel.frank@ipvs.uni-stuttgart.de

This assignment sheet consists of three theoretical tasks.

Submit your solution in ILIAS as a single PDF file.[1] Make sure to list your full name and immatriculation number at the start of the file. Optionally, you can *additionally* upload source files (e.g. PPTX files). Remember to fill out the exercise slot and exercise presentation polls linked in ILIAS. If you have any questions, feel free to ask them in the excercise forum in ILIAS.

Submission is open until Monday, 12th of June 2022, 11:59 AM.

[1] Your drawing software probably allows to export as PDF. An alternative option is to use a PDF printer. If you create multiple PDF files, use a merging tool (like pdfarranger) to combine the PDFs into a single file.

Decision Tree

The following table shows data about mushrooms. Each mushroom is described by three attributes: Cap Color, Gill Color and Sticky Texture. The target category is Poisonous.

Cap Color	Gill Color	Sticky Texture	Poisonous?
Brown	White	Yes	Yes
Orange	White	Yes	Yes
Red	White	No	Yes
Yellow	White	No	Yes
Brown	Yellow	No	Yes
Orange	Yellow	Yes	No
Red	Yellow	Yes	No
Yellow	White	Yes	No
Brown	White	Yes	No
Yellow	White	No	No

Table 1 Mushroom Classification Dataset.

1. Compute the root split of a decision tree for this data set. Make your split decision based on maximizing information gain.
   Note: Use $\log_2$ as logarithm. Please see the additional table provided at the end of this question for $\log_2$ computation hints.
2. Plot the resulting decision stump after the execution of the optimal root-node split. The stump should clearly indicate the classification decision associated with each leaf node.
3. Provide the confusion matrix of your decision stump on the training data. Compute precision, recall and F1-score for class "Yes".

Provide step-by-step solution, even when the computation might be trivial. You may skip repetitive calculations, yet, point to the corresponding step-by-step calculation.

x	$\approx \log_2(x)$	x	$\approx \log_2(x)$
$\frac{1}{3}$	-1.6	$\frac{3}{7}$	-1.2
$\frac{2}{3}$	-0.6	$\frac{4}{7}$	-0.8
$\frac{1}{4}$	-2.0	$\frac{3}{4}$	-0.4
0.5	-1	1	0

Table 2 $\log_2$ table

1.

$$H(\text{root}) = -(0.5 \log 0.5 + 0.5 \log 0.5) = 1 \qquad (1)$$

Cap Color	Yes	No	H
Orange	1	1	1
Red	1	1	1
Brown	2	1	$-(\frac{2}{3}\log\frac{2}{3}+\frac{1}{3}\log\frac{1}{3})\approx 0.93$
Yellow	1	2	≈ 0.93

$IG = 1-(\frac{2}{10}+\frac{2}{10}+\frac{3}{10}0.93+\frac{3}{10}0.93)\approx 0.04$

Gill Color	Yes	No	H
White	4	3	$-(\frac{4}{7}\log\frac{4}{7}+\frac{3}{7}\log\frac{3}{7})\approx 0.97$
Yellow	1	2	≈ 0.93

$IG = 1-(\frac{7}{10}0.97+\frac{3}{10}0.93)\approx 0.04$

Sticky Texture	Yes	No	H
Yes	2	4	≈ 0.93
No	3	1	$-(\frac{3}{4}\log\frac{3}{4}+\frac{1}{4}\log\frac{1}{4})\approx 0.8$

$IG = 1-(\frac{6}{10}0.93+\frac{4}{10}0.8)\approx 0.12$

Split root with feature "Sticky Texture".

2. (Sticky Texture)
   - If Yes $\Rightarrow$ Classify No
   - If No $\Rightarrow$ Classify Yes
3. Confusion matrix:

Predicted/Actual	Yes	No
Yes	3	1
No	2	4

$$\text{Precision} = \frac{3}{3+1} = 0.75$$
$$\text{Recall} = \frac{3}{3+2} = 0.6$$
$$\text{F1} = 2\cdot\frac{0.75\cdot 0.66}{0.75+0.66} = 0.7$$

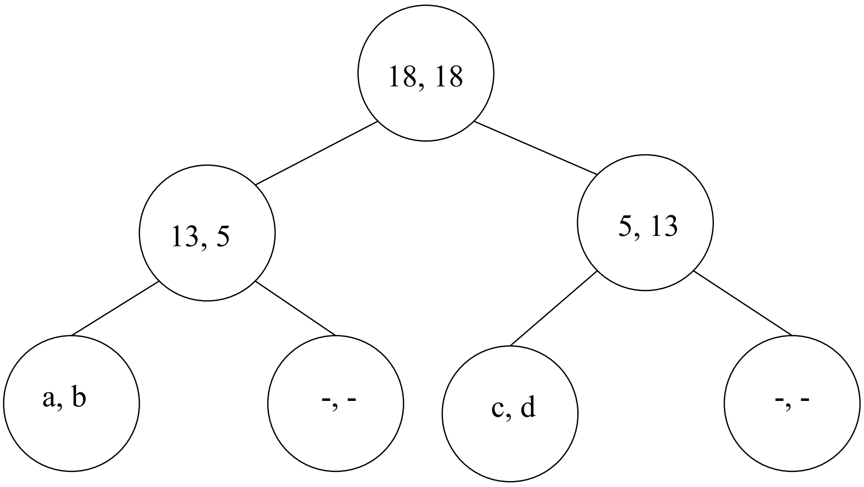

Figure 1

Decision Trees and Information Gain

Figure 1 illustrates a decision tree, with numerical values assigned to each node representing the count of examples belonging to different classes. For instance, the root node displays the numbers 18 and 18, indicating 18 examples in class positive and 18 examples in class negative. In order to save space, we have omitted the specific counts in the two right leaf nodes, as they can be inferred from the values present in the corresponding left leaf nodes.

1. What is the information gain of the split made at the root node?

   $H(\text{root}) = -(0.5\log 0.5 + 0.5\log 0.5) = 1$

   $IG(split) = H(root) - (\frac{1}{2}H(left) + \frac{1}{2}H(right)) = 1 - [\frac{1}{2} * (-\frac{13}{18}\log_2\frac{13}{18} - \frac{5}{18}\log_2\frac{5}{18}) + \frac{1}{2} * (-\frac{13}{18}\log_2\frac{13}{18} - \frac{5}{18}\log_2\frac{5}{18})] = 1 - 0.85 \approx 0.15$

2. Which values of a and b will result in the minimal information gain at this split, and what is the corresponding value?

   $IG(split) = H(Y) - H(Y|X) \geq 0$ since $H(Y) \geq H(Y|X)$

   $IG(split)_{min} = 0$ when $H(Y) = H(Y|X) = -(\frac{13}{18}) \times \log_2(\frac{13}{18}) - \frac{5}{18} \times \log_2(\frac{5}{18})$

   $$\Rightarrow \quad \begin{cases} a = 0 \\ b = 0 \end{cases} \text{or} \quad \begin{cases} a = 13 \\ b = 5 \end{cases} \Rightarrow IG = 0$$

3. Which values of c and d will yield the maximum information gain at this split, and what is the resulting value?

   $IG(split) = H(Y) - H(Y|X) \leq H(Y)$ when $H(Y|X) = 0$

   $H(left) = H(right) = 0 \Rightarrow$

   $$\begin{cases} c = 5 \\ d = 0 \end{cases} \quad \text{or} \begin{cases} c = 0 \\ d = 13 \end{cases} \Rightarrow IG = 0.852$$

Regression with Decision Trees and kNN

Decision trees and k-nearest-neighbors have applications beyond classification as they can also be employed for regression tasks. Research and explain the following two questions in sufficient detail:

1. What are the distinctions between constructing regression trees and classification trees? How is a prediction calculated in a regression tree?
   When constructing a regression tree, the initial steps are similar to building a decision tree. However, there are key distinctions between constructing a classification tree and a regression tree. In regression, the output variable consists of numerical values (continuous) whereas in classification, it involves categorical values (discrete). Since entropy is not appropriate for handling continuous variables, an alternative calculation method such as Mean Square Error (MSE) is used to compute predictions in regression trees.

2. How can k-nearest-neighbors (kNN) be utilized for regression purposes?
   To employ k-nearest-neighbors (kNN) for regression, the initial step involves identifying the k closest data points to the new data point. Subsequently, the average distance of these points is calculated, and by connecting these points, the regression line can be determined.

Please provide proper citations for your sources.