Machine Learning
Prof. Dr. Dominik Heider
Wintersemester 2023/24
Heinrich-Heine-Universität Düsseldorf

Submition of
Taha El Amine Kassabi (takas100)
Emre Gökcek (emgoe104)
Lutfi Orabi (luora100)
Ibrahim Shinahov (ibshi100)

Exercise 8

20. December 2023

Task 1

a) ID3 Decision Tree

Customer Number is not a relevant attribute

Recursion 0

Root Entropy

Yes: 3 No: 6

$$E([3,6]) = \frac{1}{3}\log 3 - \frac{2}{3}\log\frac{2}{3} \approx .918$$

Professional Status

Type	Yes	No	Entropy
Non-employed	0	2	$E([0,2]) = 0$
Civil Servant	1	1	$E([1,1]) = 1$
Employee	1	2	$E([1,2]) = .918$
Self-employed	1	1	$E([1,2]) = 1$

$$H_{avg}(Professional\ Status) = \frac{2}{9} + \frac{1}{3} * .918 + \frac{2}{9} \approx .75$$

$$IG(Termination, Professional\ Status) \approx .918 - .75 = .168$$

Contract Duration

Type	Yes	No	Entropy
Low	2	0	$H([2,0]) = 0$
Medium	0	4	$H([0,4]) = 0$
High	1	2	$H([1,2]) = .918$

$$H_{avg}(Contract\ Duration) = \frac{1}{3} * .918 \approx .306$$

$$IG(Termination, Contract\ Duration) \approx .918 - .306 = .612$$

Higher IG using Contract Duration.
$\Longrightarrow$ First step CD, second step PS

Since Low and Medium CD will produce pure nodes we only need to split High CD further.

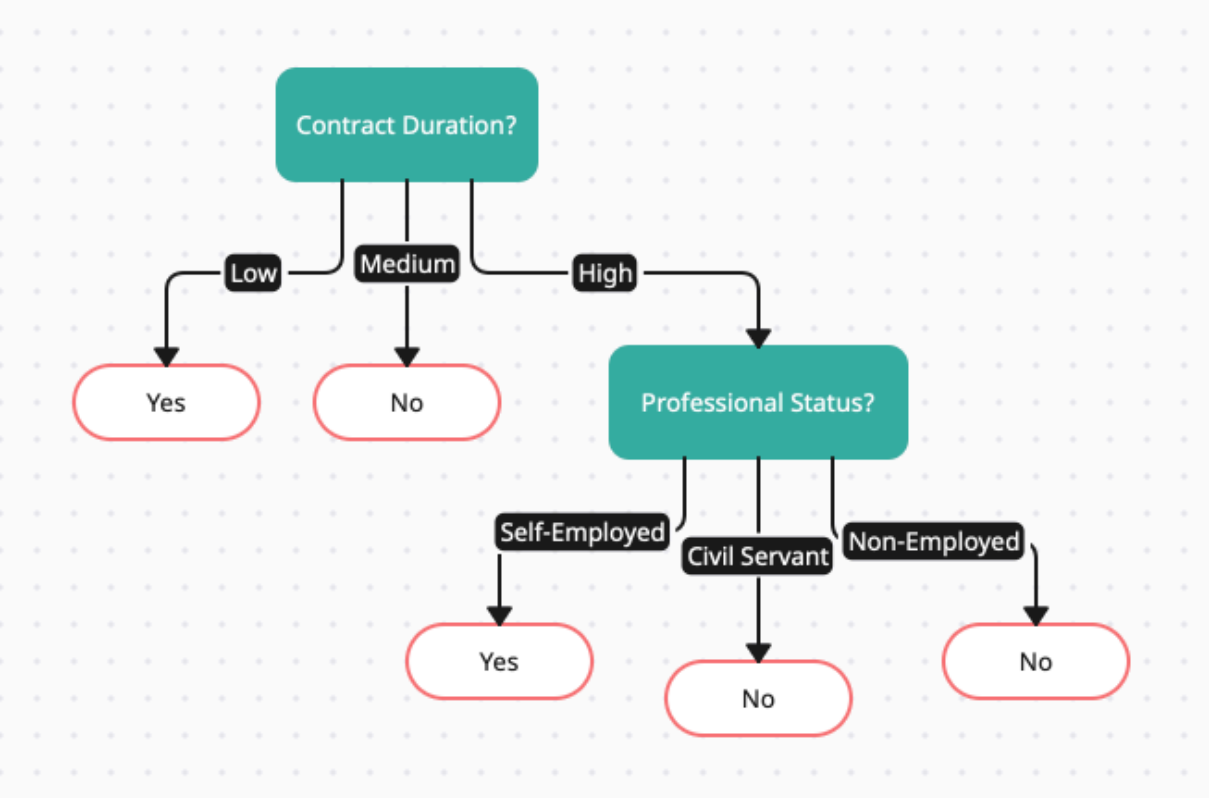
Recursion 1.1

Remaining Instances for High:

6	Non-employed	No
8	Civil Servant	No
9	Self-employed	Yes

Each has his own PS $\Longrightarrow IG(T, PS) = E(T) = .306$ as calculated before.

$\Longrightarrow$ Tree



b) Classify new

11 is at risk of termination. 12 not.
Both were immediately decided in the first step.

c) Decision Rules

```
IF Contract Duration IS Low THEN Termination IS YES
IF Contract Duration IS Medium THEN Termination IS NO
IF Contract Duration IS High AND Professional Status IS Self-Employed THEN
Termination IS YES
IF Contract Duration IS High AND Professional Status IS Civil Servant THEN
Termination IS YES
IF Contract Duration IS High AND Professional Status IS Non-Employed THEN
Termination IS NO
```

Yes. Usually High CD should mean no termination. Only Low CD.

d) IG vs GR vs GI

IG is the amount of entropy that would be reduced on average if the dataset is split on a specific attribute.
$IG(D,A) = H(D) - H_{avg}(D,A)$, where H_{avg} is the weighted average of the entropies. Id we split the data D using Attribute A.

GR is the ratio of Information Gain and Split Info (SI), where we correct for IG's preference of a higher number of categories. $SI(D,A) = -\sum_{S_i \in Split(S,A)} \frac{|S_i|}{|S|} \log_2 \frac{|S_i|}{|S|}$; $GR(D,A) = \frac{IG(D,A)}{SI(D,A)}$

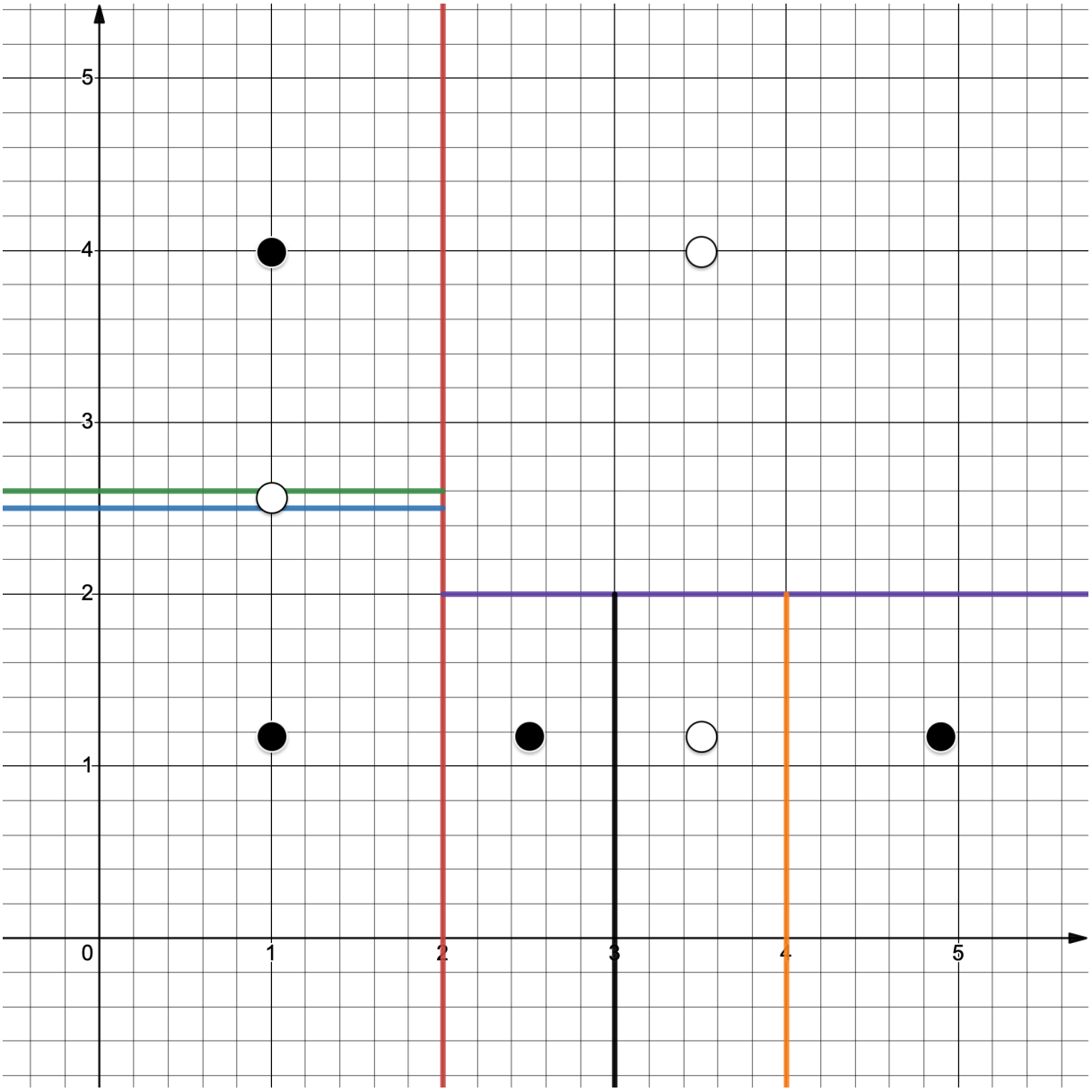
GI is 1 – the sum of squared probabilities. Generally it measures the probability, that a randomly chosen variable is classified incorrectly, if the data D is split using Attribute A.
$GI(D,A) = 1 - \sum_{S_i \in Split(S,A)} \left(\frac{|S_i|}{|S|}\right)^2$

e) Calculate GI

$$GI(D,\text{CD}) = 1 - \left(\left(\frac{|D.\text{CD} = \text{Low}|}{9}\right)^2 + \left(\frac{|D.\text{CD} = \text{Medium}|}{9}\right)^2 + \left(\frac{|D.\text{CD} = \text{High}|}{9}\right)^2\right)$$
$$= 1 - \left(\left(\frac{2}{9}\right)^2 + \left(\frac{4}{9}\right)^2 + \left(\frac{3}{9}\right)^2\right) = 1 - \frac{29}{81} = \frac{52}{81} \approx .642$$

Task 2

a) Regression 2D-DT into KoSy



b) Overfit

For $x \leq 2; 2.5 < y \leq 2.6$ we have white, but for any other y we have black. That seems non-sensical.
One could argue, that depending on the data $y \leq 2; 3 < x \leq 4$, could also be black. This would need further investigation. If It were only a few points, I would classify it black as to have a cleaner tree, since we would only have two checks: $x \leq 2 \vee y \leq 2$ is black, else white.

c) Define generalization capability

GC is the ability to perform well on unseen data.

Overfitting occurs when a model is excessively aligned with the training dataset to the point of incorporating noise, which can negatively affect its predictive accuracy.

Conversely, if the training data does not adequately capture the diversity of real-world scenarios, the model may be unable to make reliable predictions or classifications due to a lack of representative information.

Attempting to overly compensate for the second issue by setting overly broad boundaries can lead to underfitting. This occurs when the model is imprecise because it has not learned the intricacies of the data sufficiently.