

Short HW2: Classification: Introduction

Submitted individually by Thursday, 04.05, at 23:59.

You may answer in Hebrew or English and write on a computer or by hand (but be clear).

Please submit a PDF file named like your ID number, e.g., 123456789.pdf.

Bonus (maximal grade is 100): Writing on a computer (using LyX/LaTeX, Word + Equation tool, etc.) = 2 pts.

Decision trees

Here you will show that greedy TDIDT algorithms do not guarantee “optimal” trees.

1. Propose a dataset with binary features and a binary target label, such that ID3 (with no stopping rule) returns a decision tree of depth 3 or more (not counting the root level but counting the leaves) even though there exists a decision tree of depth 2 which fits the dataset perfectly. Just to be clear, the depth is the length (in edges) of the longest (directed) path in the tree.

You should:

- 1.1. Explicitly write such a dataset with 3-4 binary features, one binary target label, and 5-8 distinct examples (no two examples with identical features). The data should be in a tabular form like in the dry run in Tutorial 03.
- 1.2. Manually run ID3. Include the required entropy and information gain calculations (like in Tutorial 03). Draw the resulting tree. Make sure the tree's depth is at least 3.
- 1.3. Show a tree of depth 2 which perfectly fits the dataset (i.e., empirical error should be zero).
- 1.4. Consider running ID3 with `max_depth=2` on your dataset (when facing a tie – predict True). What is the empirical error of the resulting tree? Explain (no need to actually rerun ID3, think why).

2. We now wish to prove that the information gain, i.e., $IG(v, a) = H(v) - \frac{|v_{a=T}|}{|v|} H(v_{a=T}) - \frac{|v_{a=F}|}{|v|} H(v_{a=F})$, is non-negative for any set of examples v and a binary attribute a .

Recall our definition of $v_{a=T} \triangleq \{(x_i, y_i) \in v \mid x_i[a] = T\}$ and $v_{a=F} \triangleq \{(x_i, y_i) \in v \mid x_i[a] = F\}$.

See the following dialogue, where we (really) asked [ChatGPT](#) to prove this (we shortened it a little):

Question: could you prove that the information gain in id3 cannot be negative?

Answer: Yes, I can prove this using mathematical reasoning.

In the ID3 algorithm, the information gain (IG) of an attribute A with respect to a dataset D is defined as:

$$IG(A) = Entropy(D) - \sum (|D_v| / |D| * Entropy(D_v))$$

where $Entropy(D)$ is the entropy of the original dataset D, $|D_v|$ is the number of instances in the subset of D that have a specific value v for the attribute A, and $Entropy(D_v)$ is the entropy of that subset.

Entropy is a measure of the randomness or unpredictability of the target variable, and it is calculated as:

$$Entropy(D) = - \sum (p(c) * \log_2(p(c)))$$

where $p(c)$ is the proportion of instances in D that belong to class c.

Now, let's assume that the IG for a given attribute A is negative, i.e., $IG(A) < 0$.

Then we have:

$$Entropy(D) - \sum (|D_v| / |D| * Entropy(D_v)) < 0$$

Rearranging the terms, we get:

$$Entropy(D) < \sum (|D_v| / |D| * Entropy(D_v))$$

Since both $Entropy(D)$ and $Entropy(D_v)$ are non-negative, the right-hand side of the inequality is also non-negative.

Therefore, we have:

$$Entropy(D) \geq 0.$$

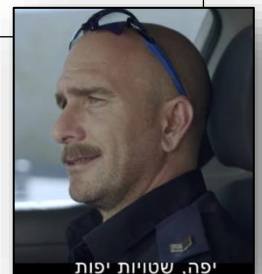
This contradicts the assumption that $IG(A) < 0$, because

$Entropy(D) - \sum (|D_v| / |D| * Entropy(D_v))$ is the value of $IG(A)$.

Thus, we can conclude that the information gain in the ID3 algorithm is always non-negative and cannot be negative.

- 2.1. Attached is a quote from “ארץ נהדרת” that best describes the answer above.

Briefly explain what the problem with the above “proof” is.



- 2.2. Let us really prove this.

Let v be an arbitrary set of examples and a be a binary attribute (feature).

Assume the classes are also binary.

The entropy is thus: $H(v) = H\left(\frac{1}{|v|} \{(x, y) \in v \mid y = 1\}\right) \triangleq H(p_v) = -p_v \log_2 p_v - (1 - p_v) \log_2 (1 - p_v)$.

(Notice that like in the tutorial, we use H with a set v or with a probability p_v interchangeably.)

Prove: $IG(v, a) \geq 0$.

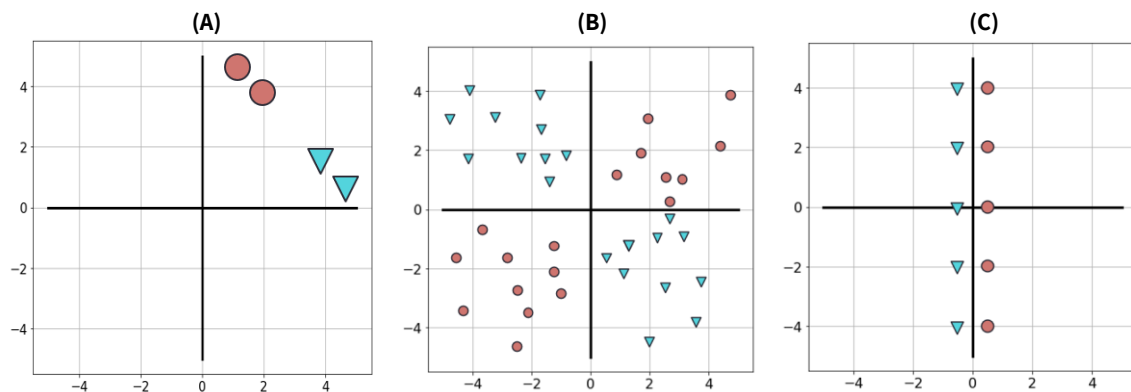
You can use the following property without proving it.

Property: It holds $\forall \beta_1, \beta_2, \alpha \in [0, 1]$ that: $-\alpha H(\beta_1) - (1 - \alpha) H(\beta_2) \geq -H(\alpha \beta_1 + (1 - \alpha) \beta_2)$.

Separability

3. Following are 3 training sets in the \mathbb{R}^2 feature space with 2 classes (blue/red).

Assume no dataset has two points in the exact same coordinates.



Following are 4 models.

- i. kNN with $k = 1$ (where a point is not considered a neighbor of itself)
- ii. kNN with $k = 3$ (where a point is not considered a neighbor of itself)
- iii. Homogeneous linear model
- iv. Decision tree with at most 4 leaves

(We only consider nodes that split according to a threshold rule on one feature, e.g., $x_1 \geq 5$.)

3.1. For each model above, write which datasets this model can perfectly fit (i.e., with 0 training error) and which datasets it cannot. Write your answers in a table, like in the example below. When you say a model cannot perfectly fit a certain dataset, explain why in 1-2 sentences (without drawings).

Example for a table (answers are random):

Model / dataset	(A)	(B)	(C)
i.	Yes	Yes	Yes
⋮	⋮	⋮	⋮
iv.	Yes	No. Because trees are green.	Yes

3.2. Now assume that the 1st feature from all the datasets above is scaled by an unknown factor $\alpha > 0$.

That is, each 2-dimensional data point (x_1, x_2) is transformed into $(\alpha x_1, x_2)$.

Without knowing the exact factor α , answer for each of the 4 models:

- Might your answers for that model change?
 - If not, briefly explain why.
 - Otherwise, the answers for which datasets might change? Briefly explain why.

Answer for example:

- i. Answers unchanged because this is the best model ever.

⋮

- iv. Answers on datasets (A), (C) might change because this and that.