Introduction to Machine Learning Course

Short HW2: Classification: Introduction

Submitted individually by Thursday, 24.11, at 23:59.

We'll receive late submissions for additional 24 hours, deducing 5 points from the grade.

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. **Prove/Refute:** Given a tree node $v = \{(x_i, y_i)\}_{i=1}^{m}$ that includes $m \in \mathbb{N}$ examples and given a binary feature a of the examples in v. Assume we split according to a, meaning that we create two subsets $v_{a=T} \triangleq \{(x_i, y_i) \in v \mid x_i[a] = T\}$ and $v_{a=F}$. Then, the entropies of both subsets cannot increase, i.e., it holds that $\left(H(v_{a=T}) \leq H(v)\right) \wedge \left(H(v_{a=F}) \leq H(v)\right)$.

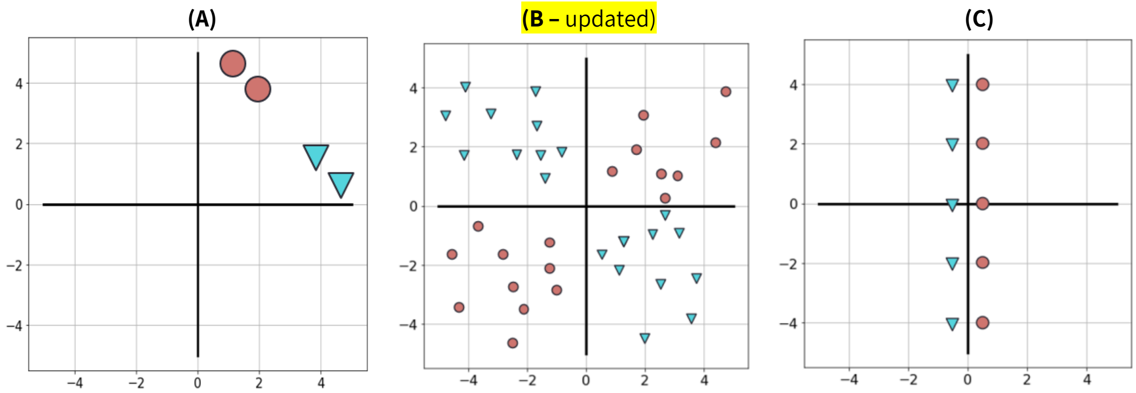
   Note I: We defined in the tutorial $H(v) = H\left(\frac{1}{|v|}\{(x,y) \in v \mid y = 1\}\right) \triangleq H(p_v) = -p_v \log_2 p_v - (1-p_v)\log_2(1-p_v)$.

   Note II: This question does not focus on a specific algorithm or a splitting criterion, but rather on properties of the entropy.

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 all datapoints are rotated by the same unknown angle θ (around the origin). That is, each 2-dimensional data point $\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ is transformed into $\mathbf{Q}\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ where $\mathbf{Q} \in \mathbb{R}^{2\times 2}$ is some rotation matrix.

Without knowing the exact angle θ, 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.