Introduction to Machine Learning Course

Dry 2 – Classification: Introduction

Submitted individually by Tuesday, 27.04, at 23:59. Each day of delay costs 5 points.

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.

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 tree from the dry run in the tutorial is of depth 3.

   You should:

   1.1. Explicitly write such a dataset with 3-4 binary features, one binary target label, and 5-10 examples. 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.
   Draw the resulting tree. See the dry in tutorial 03 for reference.
   Make sure the tree's depth is at least 3.

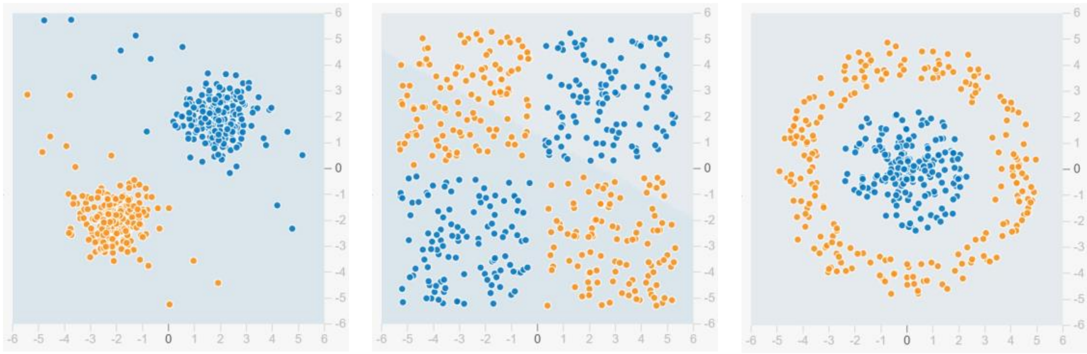
   1.3. Show a tree of depth 2 which perfectly fits the dataset (i.e., empirical error should be zero).

   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).

Separability

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

(A) (B) (C)



2.1. For each of the following models, choose all datasets from the above that the model can perfectly fit (i.e., with 0 training error).

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. kNN with $k = m - 1$ (where a point is not a neighbor of itself; m is the number of points)
iv. Linear SVM (i.e., with no kernel)
v. Decision tree with no stopping criteria.
vi. Decision tree with at most 2 leaves
vii. Decision tree with at most 4 leaves

2.2. Now consider all the datasets are rotated by the same unknown angle.

That is, each 2-dimensional data point $\boldsymbol{x}$ is transformed into $\mathrm{R}\boldsymbol{x}$, where R is some unknown (unitary) rotation matrix.

Without knowing the exact rotation angle, answer <u>for each of the 7 models</u> from 2.1:

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