Johannes Gutenberg-Universität Mainz
Institut für Informatik
Data Mining
Prof. Dr. Stefan Kramer

Machine Learning - Sheet 1

30.04.2020
Deadline: 07.05.2020 - 18:00

Task 1: Decision Tree *(20 Points)*

Read pages 55-60 of the book Machine Learning [1], and make yourself familiar with ARFF files (`http://www.cs.waikato.ac.nz/ml/weka/arff.html`). Use the Python skeleton code to implement a basic decision tree algorithm, as described in Section 3.4 (page 55).

1. (*3 points*) Implement an ARFF file parser in the `parser` method.
2. (*2 point*) Implement the data structure of the decision tree (inner nodes, leaves, the actual decision method) by completing class `Node` and `DecisionTree`. Also, complete `print_recursive` method that prints the resulting tree in an indented format.
3. (*3 points*) Implement the method `entropyOnSubset` that takes three arguments: a dataset $D = [inst_0, \ldots, inst_{N-1}]$, a list of indices $I = [i_0, i_1, \ldots, i_{m-1}]$ that describes a subset of the dataset, and a class attribute C. The subset of the dataset is then defined as

   $$S := \{inst_{i_j} | 0 \leq j < m\}.$$

   The entropy of S relative to the C-wise classification is then

   $$H(S) := -\sum_{v \in values(C)} p_v \cdot log_2(p_v),$$

   where p_v is the proportion of S belonging to class v. The value $H(S)$ is what the `entropyOnSubset`-method should return.
4. (*3 points*) Let D, I, S and C be as above. Implement the method `informationGain` that takes D, I, C and an additional attribute A as arguments, and returns

   $$InformationGain(S, A) := H(S) - \sum_{v \in values(A)} \frac{|S_v|}{|S|} \cdot H(S_v),$$

   where S_v are those instances that take value v at attribute A.
5. (*4 point*) Implement an `attributeSelection` method that performs the attribute selection for a given node, reusing the `informationGain` method.
6. (*5 points*) Implement methods `trainModelOnSubset` and `trainModel`, and test your implementation on the Weather dataset (`weather.nominal.arff`).

Please note that we will use this implementation in later exercises!

References

[1] Tom M. Mitchell. *Machine learning.* McGraw Hill series in computer science. McGraw-Hill, 1997.