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

Machine Learning - Sheet 1

27.04.2017
Deadline: 04.05.2017 - 10:00

Task 1: Boolean Functions *(4 Points)*

Give decision trees to represent the following Boolean functions:

(a) $A \wedge \neg B$

(b) $A \vee (B \wedge C)$

(c) A xor B

(d) $(A \wedge B) \vee (C \wedge D)$

(e) M of N: a function that depends on variables $X_1, \ldots, X_N$, and returns `true` if and only if exactly M of the N variables are `true`.

Fill the gaps in the provided file `Exercise_01_01.java`, test your code.

Task 2: Literature *(0 Points)*

Read pages 55-60 of the book Machine Learning [1].

Task 3: Entropy *(7 Points)*

An information source emits four symbols A, B, C, D, which occur with probabilities $p(A), \ldots, p(D)$.

1. Suppose that $p(A) = 0.5$, $p(B) = 0.3$, $p(C) = 0.1$, $p(D) = 0.1$. Compute the entropy $H(p)$.
2. For which probabilities is the entropy H maximal? For which is it minimal?
3. Explain in your own words what the entropy is.
4. Design binary code words for the symbols. Fill the gaps in the provided file `Exercise_01_03.java`.
5. How is the entropy used for the decision tree construction?

Task 4: ARFF Files *(0 Points)*

Make yourself familar with ARFF files (see `http://www.cs.waikato.ac.nz/ml/weka/arff.html`).

Task 5: Information Gain *(9 Points)*

Implement the following methods in Java (`Exercise_01_05.java`):

1. 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 describe 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.

2. 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.

3. Test your implementation and try it out on the Weather dataset (`weather.nominal.arff`). Hint: The provided class `Dataset` has a method `load` that can load ARFF datasets.

References

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