Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Peer Kröger
Yifeng Lu

Knowledge Discovery in Databases II
SS 2018

Exercise 2: Feature Selection

Exercise 2-1   Information Gain

Compute the information gain for each attribute in the following table, which attribute is the best?

gift ID	gender of the recipient	useful	beautiful	self-made	eatable	liked by the recipient
1	male	yes	yes	no	yes	yes
2	male	yes	yes	yes	no	yes
3	male	yes	no	yes	no	yes
4	male	yes	yes	no	no	yes
5	male	no	yes	yes	yes	yes
6	male	no	little	no	yes	yes
7	male	no	no	yes	no	no
8	male	no	yes	no	no	no
9	female	no	yes	no	yes	yes
10	female	no	yes	no	no	yes
11	female	no	little	yes	no	yes
12	female	no	no	no	no	yes
13	female	no	little	yes	no	yes
14	female	no	little	yes	no	yes
15	female	no	little	no	yes	no
16	female	no	little	no	no	no

Exercise 2-2   Greedy Forward Selection

The code template FS_template.py contains python code to read labeled feature vectors from an ARFF file (e.g. iris.arff) and compute the l best features either using Information Gain or χ^2-statistics.

(a) Download *FS_template.py* and the data set *iris.arff* from the homepage and analyse the code.

(b) Implement the method *class_counter* building up a dictionary containing the number of occurrences of the elements in *label_list* in *labels*.

(c) Implement the function *compute_entropy* for a given dictionary of labels and counts, and the sum over all counts *all* which computes the entropy in the dictionary.

(d) Implement the method *x2_statistics* for calculating this metric for a given split. The input consists of class dictionaries for both sides of the splits (*counter_l*, *counter_r*) and the number of elements of each side of the split (*all_l*, *all_r*).

(e) Now change the code, so that the feature selection is based on information gain instead of the χ^2-statistics.

(f) Weka provides a series of feature selection functions in "Select attributes". Compare your results with Weka. What you have implemented above corresponds to which attribute evaluator and search method in Weka?

Exercise 2-3   Subspace Selection by Inconsistency

Determine the most informative subspace using Branch-and-Bound in combination with the inconsistency criterium.

ID	attribute X	attribute Y	attribute Z	class
A	2	red	yes	1
B	3	red	yes	1
C	3	green	yes	1
D	4	green	yes	2
E	1	red	yes	2
F	1	green	yes	2

Exercise 2-4   Potential of inconsistencies in different domains

Given attributes $A_i \in \mathbb{N}$, attributes $B_i \in \{\text{red}, \text{green}, \text{blue}\}$, and attributes $C_i \in 0, 1$.

Is it possible for all n elements in a data set to be mutually distinct, when considering a feature space consisting of the following attributes:

- A_1

- B_1

- C_1

- $C_1 \times C_2 \times C_3$

- $B_1 \times C_2$

- $B_i^k \times C_j^l$

- $B_1 \times C_2 \times A_3$