

Knowledge Discovery in Databases II

SS 2019

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 Potential of inconsistencies in different domains

Given attribute values for $A_1 \in \mathbb{N}$, attribute values for $A_2 \in \{\text{red, green, blue}\}$, and attribute values for $A_3 \in \{x \in \mathbb{Z} | x^2 < 25\}$. What is the number of mutually distinct elements n in a data set which contains a feature space with one of the following attributes:

- A_1
- A_2^3
- $A_2 \times A_3$
- $A_1 \times A_2 \times A_3$
- $A_2^k \times A_3^l, k, l \in \mathbb{Z}$

Exercise 2-3 Subspace Inconsistency

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

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 χ^2 -statistics

Calculate if there is a dependency between your chances of passing the KDD2 exam and visiting the Exercises.

Grade	took part	enjoyed summer
1	10	5
2	25	8
3	35	7
4	20	16
5	12	48

Exercise 2-5 Greedy Forward Selection (option)

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.

- Download `FS_template.py` and the data set `iris.arff` from the homepage and analyse the code.
- Implement the method `class_counter` building up a dictionary containing the number of occurrences of the elements in `label_list` in `labels`.
- 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.
- 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`).
- Now change the code, so that the feature selection is based on information gain instead of the χ^2 -statistics.