IR+DM

Jens Kristian R. Nielsen, Thomas D. Vinther

15. maj 2019

1 Abstract

Abstract: a paragraph of 3-10 lines text summarizing what you did and whether your solution works (i.e., to which extent do you consider it complete and sufficiently detailed? Which issues, if any, remain to be solved?), anything else that should be taken into account when assessing your solution.

2 Solution

2.1 A

Consider the data in the table about playing tennis. Apply the information gain based algorithm to obtain a decision tree. Hint: the lecture on Data Mining shows how to get started (and you can use the numbers for humidity and wind on the top level to see if you are on the right track).

We use the information gain algorithm as seen in the lectures, where we choose class to be our "target/ decision. And so we start to calculate the information gain, for building our decision tree by looking at the different splits as seen in the tables below:

$O\backslash C$	Yes	No
Sunny	3	2
Overcast	5	0
Rain	2	3

$T\backslash C$	Yes	No
Hot	2	2
Mild	4	2
Cool	3	1

$H\backslash C$	Yes	No
High	3	4
Normal	6	1

$W\backslash C$	Yes	No
False	6	2
True	3	3

2.2 B

Consider the data and its initial partitions (light blue, dark blue) depicted on the right side. Apply the k-means algorithm to find two clusters. Instead of the (usual) Euclidean Distance between points, use the Manhattan Distance. The Manhattan Distance between two points is the sum of absolute differences in each dimension (so basically, no need to square and take the square root as in Euclidean Distance). Formally, MD(x,y)=, where x=(x1, x2) is a point with 2 dimensions as is the case here. Using Manhattan Distance, it should be relatively easy to sketch the steps and results of k-means in (copies of) the figure.

2.3 C

C. Take the following three text examples:
"Unlike classification or prediction, which analyzes data objects with class labels, clustering analyzes data objects without consulting a known class label."
"Classification can be used for prediction of class labels of data objects. However, in many applications, prediction of missing values is performed to fit data objects into a schema."
"Sun Salutation, a ritual performed in the early morning, combines seven different postures. The sun, the life generator, is invoked by this Yogic exercise."
Using word frequency (simple word counts), Euclidean Distance and Cosine Similarity, as well as the following "stop words", which of these are most similar? Stopwords = { a, an, are, be, because, by, can, for, however, in, into, is, keep, many, not, of, or, rather, than, the, they, this, to, unlike, used, way, which, with, without }

3 Summary

Summary: a paragraph of 3-10 lines text summarizing what you have learned and what you would still like to learn: what have you realized when working with the hand-in material and what would you like to continue with?