Exam Data Mining

Date: 8 Januari 2018
Time: 14:00 - 17:00

General remarks

- A calculator is allowed. This also includes mobile phones turned to flight mode. You can only use your phone to use a calculator. We will check you phone for flight mode.
- The grades will be published within four weeks on the door of room 110.
- Your answers can be in English or Dutch.
- Cheating in any form will have serious consequences.

Question 1. Short questions (15 point)

Give short, pertinent answers to the following questions.

(a) The purpose of discretisation is to turn a numeric attribute into a nominal one. Give a disadvantage and an advantage of this operation.

(b) What is the name of the standard repository of datasets that is often used for experimenting with data mining algorithms?

(c) Explain what is the difference between entropy and joint entropy.

(d) In essence, the Self-Organising map and the k-Means algorithm perform a very similar clustering task. When comparing the resulting clusterings that these algorithms produce, what is the most striking difference?

(e) What is the benefit of Kernel Density Estimation over histograms, for density estimation over a single numeric attribute?

(f) Explain what Leave-One-Out means.

Question 2. Frequent Pattern Mining (20 punten)

Given a transactional dataset with the following itemsets over $\{A, \ldots, E\}$:

tid	Items
1	$\{D\}$
2	$\{A, C\}$
3	$\{B, E\}$
4	$\{D, C\}$
5	$\{D, C\}$
6	$\{A, D, C\}$
7	$\{A, B, E\}$
8	$\{C, B, E\}$
9	$\{A, C, B, E\}$
10	$\{A, C, B, E\}$

(a) How many association rules can theoretically be derived from this dataset of 5 items?

(b) Give the definitions of a *maximal itemset* and a *closed itemset*.

(c) Given a minimal support $minsup = 0.3$, draw the itemset lattice and label each node with at least one of the following letters, where appropriate: *I*= infrequent itemset, *F*=frequent itemset, *M*=maximal itemset, *C*=closed itemset.

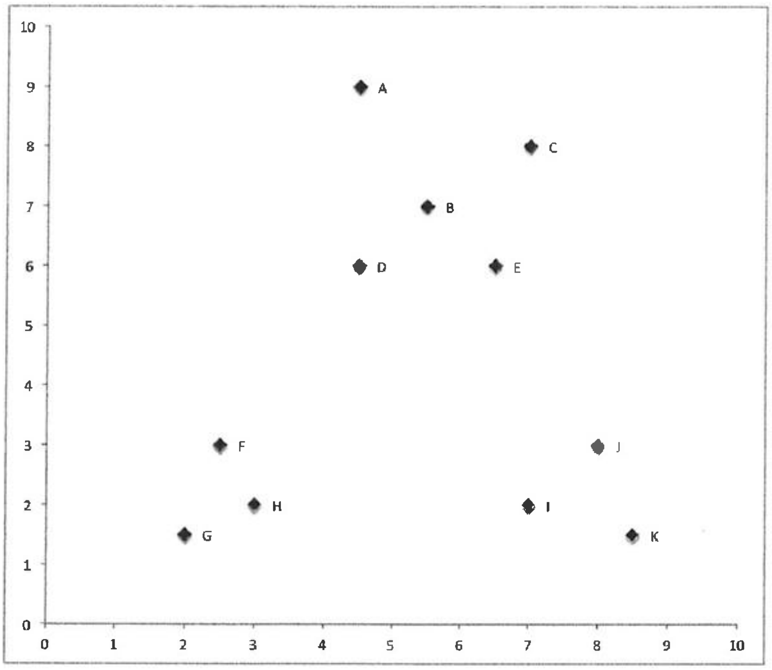
Question 3. Feature Selection (15 points)

(a) Explain the difference between a wrapper method and a filter method for feature selection.

(b) If you intend to apply C4.5 after the feature selection step, would you prefer a wrapper method or a filter method?

(c) 10-fold cross validation is the standard for validating a given classification algorithm and dataset. If feature selection is part of your classification set-up, it should be included inside the cross-validation procedure, instead of first selecting features. Explain why this is the case.

Question 4. Clustering (15 points)



a Consider the dataset given in the figure above. Provide a (rough) dendrogram as it would be produced using a bottom-up hierarchical clustering algorithm, using 'single link' distance between clusters. Since exact distances are hard to estimate from a picture, an approximate answer is acceptable.

b Describe how the following four settings compute the distance between between clusters:

- centroid
- single link
- complete link
- average link

c Assume we now cluster the data using k-Means, with $k = 3$. Give the approximate location of the final three centroids (in x, y coordinates).

d With the current dataset, the result of k-Means is quite stable at $k = 3$: each run will produce the same centroids (modulo ordering of the centroids). Give an example of a dataset that will *not* produce stable results.

Question 5. Maximally Informative k-Itemsets (15 points)

Assume a dataset is given that includes four binary attributes (items) $A, \ldots, D$. After inspecting the data, the joint entropy of the following itemsets is already computed:
$H(\{A, B\}) = 1.8$,
$H(\{A, D\}) = 1.2$,
$H(\{B, C\}) = 1.3$,
$H(\{A\}) = 1$,
$H(\{B\}) = 1$,
$H(\{C\}) = 0.6$,
and $H(\{D\}) = 0.3$.

a Give the tightest upper bound for the joint entropy $H(\{A, B, C, D\})$ that can be computed, given the available information.

b Is $\{A, B, C, D\}$ a miki with $k = 4$?

c Explain how the joint entropy $H(\{A, B, C, D\})$ is computed from the dataset.

d Give a greedy algorithm for finding approximate mikis, given a dataset D of width n and size N, and a itemset size k. The algorithm should test at most $O(kn)$ candidates. The algorithm can use a function *computeEntropy*(D, i) that returns the joint entropy of a candidate (where i is an itemset).

Question 6. Regression (20 points)

(a) Explain what the intuition is behind the definition of R^2, and how the possible values of the measure should be interpreted. If necessary, draw a diagram.

(b) A subgroup (in a regression setting) can also be interpreted as a (simple) regression model. Explain how this model works.

(c) Under what circumstances would you opt for a tree (regression or model) rather than a linear model? Give two reasons.

(d) Give the name of three algorithms for regression (hint: "regression trees" and "model trees" are not algorithms).

(e) In the algorithm that produces regression and/or model trees, what splitting criterion is used? Provide a formula or a description of how to choose the best split.