Department of Information and Computing Sciences, Faculty of Science, UU.
Made available in electronic form by the TBC of A–Eskwadraat
In 2009-2010, the course INFODM was given by Dr. A. J. Feelders.

Data Mining (INFODM)
July 8, 2010

You are allowed to consult 1 A4 sheet with notes written on both sides. You are allowed to use a calculator. Always show how you arrived at the result of your calculations. If you are a native speaker, answers in Dutch are preferred.

Question 1. Multiple Choice (16 points)

For the following questions, zero or more answers may be correct.

a) Which of the following statements about classification trees are correct?

1. In growing a tree, the misclassification error on the training sample never goes up when we expand one of its leafs.
2. In growing a tree, it is always possible to continue splitting until each leaf node contains examples of a single class.
3. When used to compute the impurity reduction of a split, the gini-index and entropy sometimes prefer different splits.
4. When C4.5 (J48) has to classify a new case whose value for the split attribute is missing at a given node, it sends the case to the child node with the highest probability (i.e. relative frequency).

b) Which of the following statements about frequent pattern mining are correct?

1. If all the subsets of size $k-1$ of a k-itemset are frequent, then the itemset itself must also be frequent.
2. All maximal frequent itemsets are closed.
3. From just the set of all *maximal* frequent itemsets and their support, one can infer all frequent itemsets and their support.
4. For an association rule, if we move one item from the right-hand-side to the left-hand-side of the rule, then the confidence will never go down.

c) Which of the following statements about linear regression/classification are correct?

1. Logistic regression produces a classifier with linear decision boundary that minimizes the number of classification errors on the training sample.
2. The Naive Bayes classifier makes the assumption that the attributes are independent given the class variable.
3. In linear regression, $\mathcal{R}^2$ is a number between -1 and 1 that measures the proportion of variation in y that is explained by the model.
4. Linear regression can only be applied when the expected value of y is a *linear* function of the input (predictor) variable z.

d) Which of the following statements about clustering are correct?

1. We don?t want the clusters that are found by a clustering algorithm to depend on the unit of measurement of a variable. For numeric data, we can prevent this from happening by subtracting the mean from each variable, so we get a new variable with zero mean.
2. In the DBScan algorithm, a *core* point is a point that has at least a specified number of points (MinPts) within a given radius (Eps).

3. In agglomerative hierarchical clustering, we can use single-linkage (MIN), complete-linkage (MAX) or average-linkage (Group Average) to compute the dissimilarity between clusters. The first step of the algorithm (the first merging of clusters) is the same regardless of the method we use to compute the dissimilarity between clusters.
4. In k-means clustering, selecting the value of k that produces the smallest Sum of Squared Errors (SSE) is not suited as a method to determine the number of clusters present in the data.

Question 2. Frequent Itemset Mining (14 points)

Given are the following frequent 3-itemsets:

$\{a, b, c\}$ $\{a, b, d\}$ $\{a, b, e\}$ $\{a, c, d\}$ $\{a, c, e\}$ $\{b, c, d\}$ $\{b, c, e\}$ $\{b, d, e\}$ $\{c, d, e\}$

a) List all candidate 4-itemsets obtained by the candidate generation procedure of the Apriori algorithm.

b) List all the candidate 4-itemsets that are pruned in the candidate pruning step of the Apriori algorithm.

We say a rule quality measure $\mathcal{Q}$ is symmetric if

$$\mathcal{Q}(A \rightarrow B) = \mathcal{Q}(B \rightarrow A),$$

for all itemsets A and B.

c) Is *Lift* a symmetrical measure? lf your answer is *Yes*, give a proof. If your answer is *No*, give a counterexample.

Question 3. Sequence Mining (15 points)

Consider the following data sequence:

$$\mathbf{d} = < \{a, b, c\}\{b, d\}\{b, c, d\}\{a, b\}\{c, d, e\} > .$$

Assume that the elements of $\mathbf{d}$ occur on consecutive time points. The following timing constraints are given:

- mingap = 0
- maxgap = 3
- maxspan = 5
- window size (ws) = 1

For each of the sequences below, determine whether, under the given timing constraints, they are valid subsequences of $\mathbf{d}$:

a) $< \{a, b\}\{c, d\}\{e\} >$

b) $< \{b\}\{b\}\{b\}\{b\} >$

c) $< \{a\}\{a\}\{b\} >$

d) $< \{a, b, c, d\}\{a, b, c, d\} >$

e) $< \{a, c\}\{e\} >$

Question 4. Regression (15 Points)

We are given the following observations on x and y:

x	0	1	3	4	11
y	0	11	23	38	89

We want to see if a linear model gives a reasonable lit of the data, so we estimate the model

$$\mathbb{E}[y|x] = w_0 + w_1 x$$

with least squares.

a) Compute the least squares estimates of w_0 and $w - 1$.

b) Predict the value of y for $x = 7$ using the result you obtained under a).

c) Predict the value of y for $x = 7$ using k nearest neighbour with $k = 3$.

Question 5. Clustering (20 points)

We are given the following data on 4 objects:

object	x_1	x_2
1	2	2
2	8	6
3	6	8
4	2	4

a) Cluster this data into two clusters, using the k-means algorithm. Use squared Euclidian distance as the distance measure. To initialize the algorithm, put objects 1 and 3 in one cluster, and objects 2 and 4 in the other cluster. Show the steps of the algorithm clearly. Give the value of the k-means error function after convergence.

b) What is the value of the error function in the optimal solution for $k = 4$?

To perform hierarchical clustering, we compute the squared Euclidian distance between each pair of objects, and put them in a distance matrix:

	1	2	3	4
1	0			
2	52	0		
3	52	8	0	
4	4	40	32	0

c) Perform agglomerative hierarchical clustering of the objects using the given distance matrix and single-linkage (MIN). Show the result in a dendrogram. The dendrogram should clearly show the order in which the clusters are merged, and the distance between the merged clusters. It does not have to be to scale.

Question 6. Classification Trees (20 points)

We are given data on two binary attributes A and B, and a binary class label. The possible values of A and B are T (for True) and F (for False). The right part of the table below contains counts of the number of records with the different value combinations. For example, there are 2 records of the negative class with $A = T$ and $B = F$.

		Class	
A	B	+	-
T	T	3	1
T	F	1	2
F	T	0	0
F	F	0	3

a) Compute the quality of a split on A, using the Gini-index.

b) Compute the quality of a split on B, using the Gini-index.

c) Which split is preferred? Why?

d) A researcher proposes the following impurity measure for binary classification problems:

$$i(t) = p(+|t) \times (1 - p(+|t)),$$

where $i(t)$ denotes the impurity of node t, and $p(+|t)$ denotes the relative frequency of the positive class in node t.

Is this a good impurity measure? Motivate your answer.