Universitetet i Stavanger

DAT630 - 2016 Fall - Trial Part I

Questions	Type	Grading
1 DAT630-2016-trial info	Writing assignment	Manual score
2 Summary statistics	Composite	Automatic score
3 Attribute types	Composite	Automatic score
4 Similarity	Composite	Automatic score
5 Classification	Composite	Automatic score
6 Classification	Numeric entry	Automatic score
7 Classification Evaluation	Composite	Automatic score
8 Classification	Writing assignment	Manual score
9 Clustering	Composite	Automatic score
10 Coding	Simple choice	Automatic score

DAT630 - 2016 Fall - Trial Part I

Exam start time:	15.11.2016 04:30	PDF created	14.11.2016 23:01
Exam end time:	15.11.2016 06:30	Created by	Krisztian Balog
		No. of pages	7

Section 1

1 QUESTION

DAT630-2016-trial info

You can use

- Calculator
- All written (printed) material
- All electronic material brought on a pendrive (PDFs, slides, MS Excel files, python code, etc.)
- Any program available on the PC (MS Excel, Adobe Acrobat, etc.)
- **No online resources**

Scoring of multiple choice questions:

- 2 or 3 points if correct
- 0 if unanswered
- -1 if incorrectly answered

For all computations, provide numbers up to 3 digits after the dot (e.g., 0.7, 0.25, 0.333).

If you have any comments about the exam, write them here

2 QUESTION

Summary statistics

$$\text{variance}(x) = s_x^2 = \frac{1}{m}\sum_{i=1}^{m}(x_i - \bar{x})^2$$

Compute the range and variance of the following data: *17 5 3 9 49 53 11*. (5x2p)
For variance use the formula above. All results are to be *rounded* to one decimal place.

Range:
Median:
Mean:
Variance:
Absolute Average Deviation (AAD):

3 QUESTION

Attribute types

Classify the attribute *student ID* as binary, discrete, or continuous. Also classify it as qualitative (nominal or ordinal) or quantitative (interval or ratio). (2x2p)

Select an alternative

- Continuous
- Discrete
- Binary

Select an alternative

- Quantitative
- Qualitative

4 QUESTION

Similarity

$$\mathbf{x} = (1, 1, 0, 1, 0, 0, 1, 0)$$
$$\mathbf{y} = (1, 0, 0, 1, 0, 0, 1, 1)$$

Calculate the similarity of the above two data binary vectors. (2x2p)

Jaccard similarity: [ ]
Cosine similarity: [ ]

5 QUESTION

Classification

	t1	t2	t3	t4	t5	class
Doc 1	1	0	2	2	0	C1
Doc 2	3	0	0	2	3	C3
Doc 3	0	4	2	0	0	C2
Doc 4	1	0	0	1	1	C1
Doc 5	0	0	0	2	1	C3
Doc 6	0	1	1	4	3	C3
Doc 7	0	2	1	0	0	C2
Doc 8	1	0	1	2	3	C3

Train a Naive Bayes classifier given the document-term matrix and class labels in the table above. Use Laplace smoothing for computing term probabilities.

Answer the following questions (5x2p)

- What is the prior class probability for C2?
  P(C2) = [ ]
- What is the (smoothed) probability of term "t4" belonging to C2?
  P("t4"|C2) = [ ]
- What is the probability of a new document "t1" belonging to C1?
  P(C1|"t1") = [ ]
- What is the probability of a new document "t1 t4 t5" belonging to C3?
  P(C3|"t1 t4 t5") = [ ]
- Which class will document "t4 t4 t5" be classified to?
  P(c|"t4 t4 t5") is the highest for [ ] (C1, C2, C3)

6 QUESTION

Classification

Assume a multiclass classification problem with 5 categories.
Using the one-against-one strategy, how many binary classifiers are needed in total? (3p)
Answer:

7 QUESTION

Classification Evaluation

	actual label	predicted label
Instance 1	N	Y
Instance 2	Y	Y
Instance 3	Y	Y
Instance 4	N	Y
Instance 5	N	N
Instance 6	N	N
Instance 7	Y	Y
Instance 8	Y	N
Instance 9	N	N
Instance 10	N	Y

Given the actual class labels and the predicted class labels for 10 instances in the table above, evaluate the classifier in terms of Precision, Recall, and F1-measure. (3x2p)

- Precision:
- Recall:
- F1-measure:

8 QUESTION

Classification

Explain with your own words what overfitting means in practical terms for a decision tree classifier. How can it be avoided? (4p)
(Note: Copy-paste answers from the slides or from the book will not be accepted.)

Fill in your answer here

9 QUESTION

Clustering

	x_1	x_2	x_3	x_4
P1	2	0	5	2
P2	3	4	5	9
P3	1	5	8	1
P4	7	3	1	2

The table shows the vector representation of four data points that we want to cluster using the K-means method with k = 2. Assume that we use the dot product between vectors as the similarity measure between them. If we select points 3 and 4 as the initial centroids, what will be the cluster centroids in the next iteration? (10p)
Mind that the dot product is similarity and not a distance metric!

	x1	x2	x3	x4
Centroid 1				
Centroid 2				

10 QUESTION

Coding

```
# Compute the distance between two clusters.
# - sim is the similarity matrix between the data points.
#   For any two data points i and j 0 <= sim[i][j] <= 1.
# - c1 and c2 are the list of data points (indices)
#   belonging to each cluster; c1 and c2 are both non-empty.
def cdist(sim, c1, c2):
    s = 1
    for i1 in c1:
        for i2 in c2:
            if sim[i1][i2] < s:
                s = sim[i1][i2]
    return 1-s
```

The above code computes the distance between two clusters based on a similarity matrix of data points. Which linkage function does it implement? (3p)

Select an alternative:

- Single link
- Complete link
- Group average
- None of the above