DAT630 2017 autumn, Final Exam - Answer Key

1 Info

2 Summary statistics (2p)

Which one is more useful for continuous data?

- Percentiles
- Frequencies

3 Summary statistics (3p)

Explain why statistics like AAD or MAD are less sensitive to outliers than variance.

Answer: Variance involves squares of differences with mean, and such squared differences are more affected by outliers than just the differences.

- 1p Mention underlying mean calculation (which is sensitive to outliers)
- 2p Compare with variance in terms of square differences vs differences

4 Similarity (2x2p)

$$\begin{aligned} x &= (1,0,0,1,1,0,1,1,0,1) \\ y &= (1,1,0,1,0,0,1,0,1,1) \end{aligned}$$

Calculate the similarity of the above two data binary vectors.

- Jaccard similarity: 0.5
- Cosine similarity: 0.666

5 Probability (3p)

Suppose the fraction of undergraduate students who smoke is 12% and the fraction of graduate students who smoke is 25%. If one-fifth of the college students are graduate students and the rest are undergraduates, what is the probability that a student who smokes is a graduate student?

Probability (between 0 and 1): 0.342

6 OLAP (3p)

In an OLAP data cube we have time as one of the dimensions. Which operation do we need to perform if we want to go from years to months?

- Drill-down
- Roll-up
- Slice
- Dice

7 Classification (2p)

Consider a training set that contains 100 positive and 400 negative examples. We have the following candidate rules:

- R1: A ⇒ + (covers 4 positive and 1 negative examples)
- R2: B ⇒ + (covers 30 positive and 10 negative examples)
- R3: C ⇒ + (covers 100 positive and 90 negative examples)

Determine which are the best and worst candidate rules according to Rule accuracy.

- R3 is best R1 is worst
- R2 is best R3 is worst
- R2 is best R1 is worst
- R3 is best R2 is worst
- R1 is best R2 is worst
- R1 is best R3 is worst

8 Classification (8p)

x	0.5	3.0	4.5	4.6	4.9	5.2	5.3	5.5	7.0	9.5
y	−	−	+	+	+	−	−	+	−	−

Classify the data point x = 5.0 according to its 1-, 3-, 5-, and 9-nearest neighbors (using majority vote). *(2 points each, -1 if incorrect!)*

- 1-nearest neighbor: +
- 3-nearest neighbor: -
- 5-nearest neighbor: +
- 9-nearest neighbor: -

9 Classification (10p)

	t1	t2	t3	t4	t5	t6	t7	class
Doc 1	2	0	1	2	0	2	4	C1
Doc 2	0	0	0	0	3	2	2	C3
Doc 3	3	4	0	2	0	0	2	C2
Doc 4	4	0	3	1	1	1	0	C3
Doc 5	1	0	0	3	1	2	0	C2
Doc 6	0	1	1	0	3	4	1	C1

$$P(t|c) = \frac{n(t,c)+1}{\sum_{t'} n(t',c)+|C|}$$

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

Hint: Notice that the description says Laplace smoothing! This is not to be confused with Jelinek-Mercer smoothing, what we typically use in Language Modeling retrieval. Laplace smoothing is computed using the formula shown above, where $n(t, c)$is the number of times term $t$appears with class c, and $|C|$ is the total number of classes.

Answer the following questions:

- What is the prior class probability for C2? $P(C2) = 0.333$
- What is the (smoothed) probability of term "t4" belonging to C2? $P(\text{“}t4\text{”}|C2) = 0.285$
- What is the probability of a new document "t1" belonging to C1? $P(C1|\text{“}t1\text{”}) = 0.041$
- What is the probability of a new document "t1 t4 t5" belonging to C3? $P(C3|\text{“}t1\ t4\ t5\text{”}) = 0.002$
- Which class will document "t4 t4 t5" be classified to? $P(c|\text{“}t4\ t4\ t5\text{”})$ is the highest for C2

10 Classification (3p)

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

Answer: 10

11 Clustering (3p)

50 data objects, denoted as $x_1, ..., x_{50}$, are clustered by an algorithm in the following 7 clusters:

$$
\begin{aligned}
C1 &= \{x_{46}, x_{18}, x_{43}, x_{34}, x_{37}, x_{29}, x_{14}, x_{27}, x_{50}, x_{15}, x_8\} \\
C2 &= \{x_{32}, x_9, x_{44}, x_{28}, x_{39}, x_{23}, x_3\} \\
C3 &= \{x_{16}, x_{42}, x_{24}, x_5\} \\
C4 &= \{x_{21}, x_{40}, x_{10}, x_{31}, x_{19}, x_{47}, x_{49}, x_{38}, x_4\} \\
C5 &= \{x_{22}, x_{30}, x_{17}, x_{13}, x_{11}, x_{33}, x_{34}, x_6, x_{12}, x_{45}\} \\
C6 &= \{x_2, x_{48}, x_7\} \\
C7 &= \{x_{26}, x_{25}, x_{36}, x_{35}, x_1, x_{20}\}
\end{aligned}
$$

What type of clustering is this?

- Hierarchical, non-exclusive, partial
- Partitional, exclusive, complete
- Hierarchical, non-exclusive, complete
- Partitional, non-exclusive, partial

12 Clustering (12p)

Points

	w	x	y	z
P1	3	1	0	4
P2	2	0	1.5	1
P3	1	3	4.5	5
P4	2.5	2.5	5	3.5
P5	6	3	2	0

Centroids

	w	x	y	z
C1	3	5	4	4.5
C2	1.5	2.5	1	2
C3	2	3.5	1	0

You are given five 4-dimensional data points and three initial cluster centroids.
Perform the first iteration of K-means clustering using the Euclidean distance

Which cluster the following data points get assigned to?
(In case a data point is of equal distance to more than one cluster, pick the cluster with the lowest index.)

- P1 is assigned to C2
- P3 is assigned to C1
- P5 is assigned to C3

Provide the new (updated) cluster centroids:
We are not interested in all values, only in those where there is an input field in the cell!

	w	x	y	z
C1		2.75	4.75	4.25
C2	2.5	0.5		2.5
C3	6		2	0

13 Minhash (2p)

Consider the following matrix:

Row	D1	D2	D3	D4
1	0	1	1	0
2	1	0	0	1
3	1	1	0	1
4	0	0	1	0
5	0	1	0	0
6	1	0	1	0

Perform a minhashing for the above data, with the new order of rows: 4, 6, 1, 3, 5, 2.
Which of the following statements are correct? *(1 point for each correct)*
Note: More than one options may be correct

- The minhash value for D3 is 1
- The minhash value for D2 is NOT 3
- The minhash value for D1 is 6
- The minhash value for D4 is 2
- The minhash value for D4 is NOT 3

14 Minhash (5p)

Consider the following matrix:

Row	D1	D2
1	1	1
2	0	0
3	0	1
4	1	0

Using the following two hash functions h1(x) = 3x + 1 mod 4; h2(x) = 5x + 2 mod 4;
What is the approximate Jaccard similarity of D1 and D2 based on their minhash signatures?

Answer: 0.333

15 Retrieval (8p)

Compute retrieval scores using the BM25 algorithm.

- The collection row shows the number of documents that contain the given term; the collection contains 1000 documents in total.

	T1	T2	T3	T4	T5	T6	length
Doc1	3	0	2	1	10	5	21
Doc2	4	3	3	2	1	7	20
Collection	100	50	80	93	100	25	1000

$$BM25(q,d) = \sum_{t \in q} \frac{f_{t,d} \cdot (1+k_1)}{f_{t,d} + k_1(1 - b + b\frac{|d|}{avgdl})} \cdot idf_t \qquad idf_t = \log \frac{N}{n_t}$$

- The average document length in the collection is 50.
- The BM25 parameters are k1 = 1.25 and b = 0.8.
- Use base10 logarithm for the computations!

Answers:

The query is a single term, "T2".

- BM25 score of Doc1: 0
- BM25 score of Doc2: 2.406

The query is a single term, "T2 T2 T5".

- BM25 score of Doc1: 2.108
- BM25 score of Doc2: 6.175

16 Retrieval (3p)

There are two document collections, A and B. The variance of the document lengths in A is much higher than in B. If we change parameter b of the BM25 retrieval model from 0.5 to 1, what impact will that have on the retrieval scores?

- The retrieval scores will change to the same extent
- The retrieval scores on collection A will change more than on collection B
- The retrieval scores on collection B will change more than on collection A

17 Retrieval (3p)

When using Language Modeling for document retrieval, we rewrite the probability $P(d|q)$ using Bayes' theorem and then drop $P(q)$ from the denominator. Give a brief justification of whythis can be done.

Answer: $1/P(q)$ becomes a multiplicative constant since (marginal) $P(q)$ is independent from d.

- 1p Mention it's dividing by a constant
- 2p Mention $P(q)$ is independent from d (or marginal)

18 Retrieval evaluation (8p)

Evaluate two retrieval systems in terms of DCG@5 and NDCG@10 on a given search query.
The table above contains the rankings generated by the two systems as well as the ground truth. Documents are judged on a 4point scale: nonrelevant (0), poor (1), good (2), excellent (3). The DCG formula to be used is also shown for your reference.

- What is DCG@5 for System A? 3.861

System A ranking	10, 7, 9, 8, 2, 1, 3, 4, 5, 6
System B ranking	3, 2, 1, 4, 5, 7, 8, 10, 9, 6
Ground truth	excellent: 1, 7 good: 2 poor: 3 (the rest are non-relevant)

$$DCG@p = rel_1 + \sum_{i=2}^{p} \frac{rel_i}{\log_2 i}$$

- What is DCG@5 for System B? 4.893
- What is NDCG@10 for System A? 0.693
- What is NDCG@10 for System B? 0.78

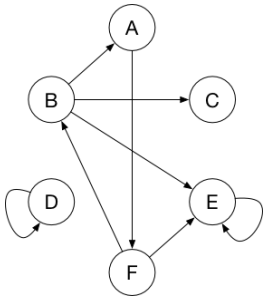
19 Retrieval evaluation (2p)

Given the following rankings and ground truth, decide which of the two systems has a higher Average Precision at rank cutoff 5.

System A ranking	1, 2, 4, 3, 6, 8, 10, 5, 9,7
System B ranking	2, 7, 5, 8, 4, 9, 1, 6,10, 3
Ground truth	2, 3, 4, 8, 10, 12

- They have the same AP
- System A
- System B

20 PageRank (12p)



Compute the PageRank values for the above graph for the first two iterations.
The probability of a random jump (i.e., the parameter q) is 0.2.

	Iteration 0	Iteration 1	Iteration 2
A	0.167	0.1	0.079
B	0.167	0.122	0.122
C	0.167	0.1	0.079
D	0.167	0.188	0.197
E	0.167	0.3	0.394
F	0.167	0.188	0.126

21 Coding (4p)

Write Python code that returns the matching document IDs for the Boolean query: "too AND (easy OR simple) AND NOT over".

```
if __name__ == "__main__":
    docs = all_docs()
    print(docs)  # prints {0, 1, 2, 3, 4}
    q_docs = match("search_term")
    print(q_docs)  # prints {1, 2}
```

You need to solve this by using the two provided python functions:

- `all_docs()` returns a set with all document IDs in the collection
- `match(term)` returns a set with the documents IDs that contain term

See the illustration above for an example. In this case, the collection contains 5 documents (IDs from 0 to 4) and documents 1 and 2 contain the term "search_term".

Answer:

```
res = match("too").intersection(match("easy").union(match("simple")).difference(match("over"))
print(res)
```

- 1p Query "too"
- 1p Query "easy OR simple"
- 1p Query "NOT over"
- 1p Full query, i.e., concatenating the three subqueries with AND