University of Stavanger

Faculty of Science and Technology

Exam in subject:	**DAT550 Data Mining**
Date:	**May 23, 2019**
Duration:	**4 hours**
Allowed remedies:	**Fixed simple calculator, Attached lecture slides**
The exam consists of:	**6 exercises on 16 pages**
Contact during exam:	**Vinay Setty, tlf. 518 32760**
Remarks:	Your responses should be entered into Inspera Assessment. For some questions there is a negative point of -1 if your answer is wrong. Unless explicitly mentioned about the negative score, the minimum score you can get for a question is 0. For some multiple choice questions there are multiple correct answers.
Attachments:	Lecture slides

Question 1: Data Preprocessing (15/85)

(a) (5%) *Consider a reservoir sampling process with k units of memory for a stream of data elements. Which of the following statements are true?*

A. Each k-subset of the data stream is equally likely to be chosen as the sample

B. Probability that an ith element in the stream replaces an existing item in the reservoir is $\frac{1}{i}$

C. Reservoir sampling is done with replacement

D. The k sample elements are the true random samples at any point in the stream

E. The ith element has a higher probability of being included in the sample than jth element provided $i < j$, (ith element appears before jth)

Warning: This question has negative score of -1 for an incorrect answer. (minimum score -1)

Solution:

A. Each k-subset of the data stream is equally likely to be chosen as the sample

B. Probability that an ith element in the stream replaces an existing item in the reservoir is $\frac{1}{i}$

C. Reservoir sampling is done with replacement

D. The k sample elements are the true random samples at any point in the stream

E. The ith element has a higher probability of being included in the sample than jth element provided $i < j$, (ith element appears before jth)

(b) (2%) *Consider a reservoir sampling process with k units of memory for a stream of data elements. Without loss of generality, assume that the data elements are natural numbers. Can you use reservoir sampling to estimate the mean value (i.e., average) of the data elements in the stream?* Warning: This question has negative score of -1 for an incorrect answer. (minimum score -1)

A. True

B. False

Solution:

A. True

B. False

(c) (4%) *Consider a reservoir sampling process with* 10 *units of memory for a stream of data elements. What is the probability that an 100th element is not included in the sample?*

Solution:
0.9

(d) (2%) *The PCA of M^T is the same as the PCA of M ((Assuming M is matrix and is not symmetric).* Warning: This question has negative score of -1 for an incorrect answer. (minimum score -1)

A. True

B. False

Solution:

A. True

B. False

(e) (2%) *The PCA of $A \times B$, for arbitrary matrices A and B, is the product of the PCA of A and the PCA of B.* Warning: This question has negative score of -1 for an incorrect answer. (minimum score -1)

A. True

B. False

Solution:

A. True

B. False

Question 2: Simple Classification (19/85)

(a) (5%) *Consider the following dataset with 8 documents d_1 to d_8 and features/attributes f_1 to f_3.*

1. What is the entropy over the categories for these training instances d1 to d9? Recall that the entropy of a partition $\mathcal{T}$ is given as $H(\mathcal{T}) = -\sum_j \mathrm{P}(\mathcal{T}_j) \cdot$

Doc	f_1	f_2	f_3	Class
d1	2	0	0	c_1(Algebra)
d2	2	0	0	c_1(Algebra)
d3	0	0	0	c_2(Calculus)
d4	0	1	0	c_2(Calculus)
d5	0	2	0	c_3(Stochastics)
d6	0	2	0	c_3(Stochastics)
d7	0	1	1	c_3(Stochastics)
d8	0	2	1	c_3(Stochastics)

$\log_2 \mathrm{P}(\mathcal{T}_j)$. Note the log_2 has been chosen to make the calculations simple for you. You actually do not need a calculator.

2. Using the training set d1 to d9, suppose we want to construct a decision tree for the binary classification of the category c3 ("Sochastics"), i.e., the tree decides whether a new document belongs to c3 category or not, using binary splits. Determine the split with the highest information gain for binary split at the root level. Recall Information gain formula

$$G(k, k_1, k_2) = H(k) - \frac{|k_1|}{|k|} H(k_1) - \frac{|k_2|}{|k|} H(k_2)$$

$f_1 \geq 1$ $f_2 \geq 1$ $f_3 \geq 1$
$f_1 \geq 2$ $f_2 \geq 2$ $f_3 \geq 2$
$f_1 \geq 3$ $f_2 \geq 3$ $f_3 \geq 3$

Solution:

1.

$$H(\mathcal{T}) = -1 \cdot \left(-\frac{2}{8}\log_2\frac{2}{8} - \frac{2}{8}\log_2\frac{2}{8} - \frac{4}{8}\log_2\frac{4}{8} \right) H(\mathcal{T}) = +0.5 + 0.5 + 0.5 = 1.5$$

2. $f_2 \geq 1$

(b) (2%) When constructing a decision tree, we will minimise the error rate when we choose a split that maximises the entropy.

A. True

B. False

Solution:

A. True

B. False

(c) (5%) Norway has become a multi-cultural country. We are given the following observations about people who live in Norway (N) or other parts of the world ($\overline{N}$), people who like Asian food (A) or not ($\overline{A}$), and people who like Bollywood (i.e., Indian) movies (B) or not ($\overline{B}$). In total, 1000 people from all over the world were asked for this survey:

- 200 people live in Norway, 800 live elsewhere in the world.
- 30 percent of the people living in Norway like Asian food.

- 50 percent of all people not living in Norway like Asian food.
- 10 percent of the people living in Norway like Bollywood movies.
- 20 percent of all people not living in Norway like Bollywood movies.
- 8 percent of all people living in Norway like both Asian food and Bollywood movies.

What is the probability that a person who likes both Asian food and Bollywood movies lives in Norway? If necessary, make appropriate assumptions on independence or conditional independence among random variables.

Recall the Bayes' theorem:

$$P[X|Y] = \frac{P[Y|X]P[X]}{P[Y]}$$

Solution:

We need to compute $P[N|AB]$. By Bayes' Theorem, we have:

$$P[N|AB] = \frac{P[AB|N]P[N]}{P[AB]}$$

$P[AB|N]$ is given by the last item in the description. P[N] is given in the first item. For P[AB] we need to make an independence assumption about the random variables A and B, for otherwise we lack information. So we further derive:

$$P[N|AB] = \frac{P[AB|N]P[N]}{P[A]P[B]}$$

For P[A] and P[B] we need to compute the total probabilities, hence:

$$P[A] = P[A|N]P[N] + P[A|\overline{N}]P[\overline{N}]$$
$$P[B] = P[B|N]P[N] + P[B|\overline{N}]P[\overline{N}]$$

We can plug in these total probabilities into the equation for $P[N|AB]$. Then all probabilities in the right-hand formula are known. Substituting the values into the equation yields:

$$P[A] = 0.3 \cdot 0.2 + 0.5 \cdot 0.8 = 0.46$$
$$P[B] = 0.1 \cdot 0.2 + 0.2 \cdot 0.8 = 0.18$$
$$P[N|AB] = \frac{0.08 \cdot 0.2}{0.46 \cdot 0.18} = \frac{0.016}{0.0828} \approx 0.2$$

(d) (2%) The Naive Bayes algorithm is sensitive to irrelevant attributes.

A. True

B. False

Solution:

A. True

B. False

(e) (5%) Consider the task of building a classifier from random data, where the attribute values are generated randomly irrespective of the class labels. Assume the data set contains records from two classes, "+" and "-". Half of the data set is used for training while the remaining half is used for testing.

Suppose two-thirds of the data belong to the positive class and the remaining one-third belong to the negative class. And the classifier predicts each test record to be positive class with probability 2/3 and negative class with probability 1/3. What is the F1-score of such a classifier? Recall that the F1-score is the harmonic mean of precision and recall/sensitivity

Solution:

$$F1 - score = \frac{2 * (4/9 \cdot N)}{2 * (4/9 \cdot N) + (2/9 \cdot N) + (2/9 \cdot N)} = \frac{8/9 \cdot N}{12/9 \cdot N} = 8/12 = 2/3$$

0.6666

		Predicated	
		+	-
actual	+	2/3*(2/3*N)	1/3*(2/3*N)
	-	2/3*(1/3*N)	1/3*(1/3*N)

Question 3: Advanced Classification (12/85)

(a) (4%) Consider the linear regression cost function with regularization:

$$J(\theta) = \frac{1}{2m}\left[\sum_{i=1}^{m}\left(h_\theta\left(x^{(i)}\right) - y^{(i)}\right)^2 + \lambda\sum_{i=1}^{n}\theta_j^p\right]$$

Suppose if we increase p to a large value what is its effect on bias and variance? Assuming $\theta > 1$.

Warning: This question has negative score of -1 for an incorrect answer. (minimum score -1)

Solution:

A. Increases both bias and variance

B. Decreases both bias and variance

C. Bias and variance are unaffected

D. decreases variance, increases bias

E. Depends on the data

(b) (2%) If enough weak classifiers are been combined, AdaBoost will eventually reach zero training error.

Solution:

A. True

B. False

(c) (4%) Suppose you are asked to design a machine learning algorithm for an hospital. The patients come in, they take various tests and measurements, such as height, weight, medical history, lifesyle, temperature, blood pressure etc. And the goal is to predict if the patient has the risk of getting heart disease. Which model would you choose for this purpose? And which model you would not choose elaborate reasoning behind your decisions. Assume you have some past labelled training data. Some of the test like ECG (Electro Cardiogram) are more expensive so you only want to run this test if there is a 80% or higher confidence from your model that this patient is at the risk of heart disease. How would you adapt your model for this purpose?

Solution:

This is an open ended question, no perfect answer but they have to justify what they choose. For example, decision trees are good for interpretability, SVMs are better if the data is not linearly separable etc. For the confidence any model can be modified to produce confidence score.

(d) (2%) In k-fold cross-validation, what is the effect of the higher value of k (the number of folds) on the training error?

A. Higher

B. Lower

C. No change

D. Depends on the dataset

Solution:

Lower (because more training data)

A. Higher

B. Lower

C. No change

D. Depends on the dataset

Question 4: Clustering (17/85)

(a) (2%) *Which of these conditions are plausible termination conditions in K-Means?*

1. Reached a fixed number of iterations.
2. Assignment of observations to clusters does not change between iterations. Except for cases with a bad local minimum.
3. Cluster center do not change between successive iterations.
4. Error (RSS (Residual Sum of Squares)) falls below a threshold.

A. Only 1,2 and 3.

B. Only 1,2 and 4.

C. Only 1,3 and 4.

D. All of the above.

Warning: This question has negative score of -1 for an incorrect answer. (minimum score -1)

Solution:

A. Only 1,2 and 3.

B. Only 1,2 and 4.

C. Only 1,3 and 4.

D. All of the above.

(b) (5%) *What is the complexity of K-Means algorithm if t : number of iterations, k : number of clusters, n : number of objects to be clustered, d :number of attributes* Warning: This question has negative score of -1 for an incorrect answer. (minimum score -1)

A. $O(t * k * n^2 * d)$.

B. $O(t * k * n * d)$.

C. $O(t * k * (n + d))$.

D. $O(t * k^2 * n * d)$.

Solution:

A. $O(t * k * n^2 * d)$.

B. $O(t * k * n * d)$**.**

C. $O(t * k * (n + d))$.

D. $O(t * k^2 * n * d)$.

(c) (5%) *Consider the following statements pertaining to K-Means algorithm*

1. The number of clusters must be estimated in Step 1.

A. True.

B. False.

2. The ability to try and judge alternatives for cluster center is severely hampered as the number of points and pairs increase.

A. True.

B. False.

3. The K-Means algorithm "only works with real valued data".

A. True.

B. False.

4. The K-means algorithm tends to work "best when the clusters that exist in the data are of approximately unequal size".

A. True.

B. False.

5. Both attribute significance and clusters themselves cannot be fully explained using the K-Means algorithm.

A. True.

B. False.

Warning: This question has negative score of -1 for an incorrect answer. (minimum score -1)

Solution:

1. The number of clusters must be estimated in Step 1.

A. True.

B. False.

2. The ability to try and judge alternatives for cluster center is severely hampered as the number of points and pairs increase.

A. True.

B. False.

3. The K-Means algorithm "only works with real valued data".

A. True.

B. False.

4. The K-means algorithm tends to work "best when the clusters that exist in the data are of approximately unequal size".

A. True.

B. False.

5. Both attribute significance and clusters themselves cannot be fully explained using the K-Means algorithm.

A. True.

B. False.

(d) (5%) Consider the following matrix: Perform a minhashing for the above data,

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

with the new permutation (order of rows): 4, 6, 1, 3, 5, 2. Which of the following

statements are correct? (Note: More than one options may be correct) Warning: This question has negative score of -1 for an incorrect answer. (minimum score 0)

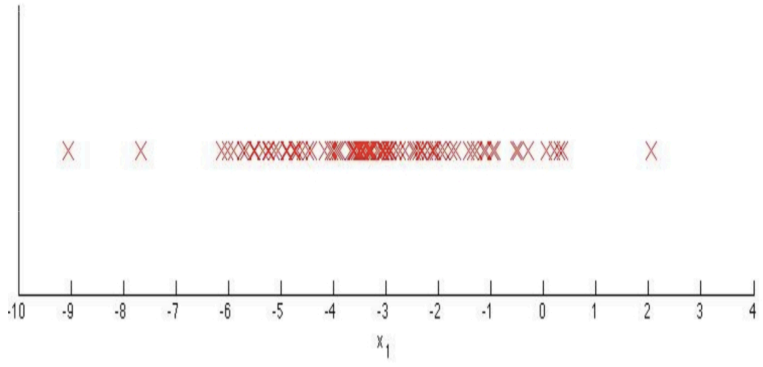
1. The minhash value for D3 is 1
2. The minhash value for D1 is 1
3. The minhash value for D4 is 2
4. The minhash value for D2 is NOT 3
5. The minhash value for D4 is NOT 1

Solution:

1. The minhash value for D3 is 1 - False
2. The minhash value for D1 is 1 - True
3. The minhash value for D4 is 2 - False
4. The minhash value for D2 is NOT 3 - True
5. The minhash value for D4 is NOT 1 - False

Question 5: Anomaly Detection (6/85)

(a) (2%) *You have 1-D dataset* $\{x^{(1)}, x^{(2)}, x^{(3)}, ...x^{(m)}\}$ *and you want to detect anomalies in this dataset. Suppose you fit the Gaussian distribution parameters* μ_1 *and* ${\sigma_1}^2$*. What values of* μ_1 *and* ${\sigma_1}^2$*, you might get?.*



A. $\mu_1 = -6$ and ${\sigma_1}^2 = 2$.
B. $\mu_1 = -6$ and ${\sigma_1}^2 = 4$.

C. $\mu_1 = -3$ and ${\sigma_1}^2 = 4$.

D. $\mu_1 = -3$ and ${\sigma_1}^2 = 2$.

Warning: This question has negative score of -1 for an incorrect answer. (minimum score -1)

Solution:

A. $\mu_1 = -6$ and ${\sigma_1}^2 = 2$.

B. $\mu_1 = -6$ and ${\sigma_1}^2 = 4$.

C. $\mu_1 = -3$ and ${\sigma_1}^2 = 4$.

D. $\mu_1 = -3$ and${\sigma_1}^2 = 2$.

(b) (4%) *Consider applying anomaly detection using a training set* $\{x^{(1)}, x^{(2)}, x^{(3)}, ...x^{(m)}\}$. *Which of the following statements are true?.*

(1) The original model $p(x_1; \mu_1, {\sigma_1}^2) \times \cdots \times p(x_n; \mu_n, \sigma_1 c^2)$ corresponds to a multivariate Gaussian where the contours of $p(x; \mu, \Sigma)$ are axis-aligned.

A. True.

B. False.

(2) Using the multivariate Gaussian model is advantageous when m (the training set size) is very small ($m < n$).

A. True.

B. False.

(3) The multivariate Gaussian model can automatically capture correlations between different features in x.

A. True.

B. False.

(4) The original model can be more computationally efficient than the multivariate Gaussian model, and thus might scale better to very large values of n (number of features).

A. True.

B. False.

Solution:

(1) The original model $p(x_1; \mu_1, {\sigma_1}^2) \times \cdots \times p(x_n; \mu_n, \sigma_1 c^2)$ corresponds to a multivariate Gaussian where the contours of $p(x; \mu, \Sigma)$ are axis-aligned.

A. True.

B. False.

(2) Using the multivariate Gaussian model is advantageous when m (the training set size) is very small ($m < n$).

A. True.

B. False.

(3) The multivariate Gaussian model can automatically capture correlations between different features in x.

A. True.

B. False.

(4) The original model can be more computationally efficient than the multivariate Gaussian model, and thus might scale better to very large values of n (number of features).

A. True.

B. False.

Question 6: Deep Learning (16/85)

(a) (2%) Which of the following activation functions are most vulnerable to vanishing gradients?

Warning: This question has negative score of -1 for an incorrect answer. (minimum score -1)

Select one or more alternatives

Solution:

A. Tanh

B. ReLU

C. Leaky ReLU

D. Sigmoid

E. Softmax

(b) (4%) You train your deep neural network and visualize the learning curve using tensorboard and you see the pattern below, what is the potential problem? and what would be your solution to resolve this issue?

Solution:

They have to mention Early stopping, Dropout, Data augmentation, Regularization

Explain any or all of them to get full 4 points.

(c) (4%) You are traing a CNN on an image with 128×128 resolution. Consider a Convolution layer with 32 filters of size 9x9. Padding and stride parameters are 0 and 1 respectively.

1. What is the dimensions of the next layer (final value after multiplying dimensions)?

Next you decide to add an additional pooling layer of size 2. Again Padding and stride parameters are 0 and 1 respectively.

2. What is the number of weights after this pooling layer?

Solution:

Using the formula

$$n^{[l]} = \left\lfloor \frac{n^{[l-1]} + 2p^{[l-1]} - f^{[l]}}{s^{[l]}} + 1 \right\rfloor$$

(128 - 0 - 9)/1 + 1 * (128 - 0 - 9)/1 + 1 * 32 = 120 * 120 * 32 = 460800

Then the maxpooling
(120 - 0 - 2)/2 + 1 * (120 - 0 - 2)/2 + 1 * 32 = 60 * 60 * 32 = 115200

(d) (2%) *You are training an RNN, and find that your weights and activations are all taking on the value of NaN ("Not a Number"). Which of these is the most likely cause of this problem?*

A. Vanishing gradient problem.
B. Exploding gradient problem.
C. ReLu activation funstion g(.) used to compute g(z), where z is too large.
D. Sigmoid activation funstion g(.) used to compute g(z), where z is too large.

Warning: This question has negative score of -1 for an incorrect answer. (minimum score -1)

Solution:

A. Vanishing gradient problem.
B. Exploding gradient problem.
C. ReLu activation funstion g(.) used to compute g(z), where z is too large.
D. Sigmoid activation funstion g(.) used to compute g(z), where z is too large.

(e) (2%) *You have a pet dog whose mood is heavily dependent on the current and past few days' weather. You've collected data for the past 365 days on the weather, which you represent as a sequence as* $x^{<1>}, \ldots, x^{<365>}$. *You've also collected data on your dog's mood, which you represent as* $y^{<1>}, \ldots, y^{<365>}$. *You'd like to build a model to map from* $x \rightarrow y$. *Should you use a Unidirectional RNN or Bidirectional RNN for this problem?*

A. Bi-RNN, because this allows the prediction of moon day t to take into account more information.
B. Bi-RNN, because this allows the backprop to compute gradients more accurately.
C. Uni-RNN, because the value of $y^{<t>}$ depends only on$x^{<1>}, \ldots, x^{<t>}$ but not on $x^{<t+1>}, \ldots, x^{<365>}$.
D. Uni-RNN, because the value of $y^{<t>}$ depends only on $x^{<t>}$ and not the other day's weather.

Warning: This question has negative score of -1 for an incorrect answer. (minimum score -1)

Solution:

A. Bi-RNN, because this allows the prediction of moon day t to take into account more information.
B. Bi-RNN, because this allows the backprop to compute gradients more accurately.
C. Uni-RNN, because the value of $y^{<t>}$ depends only on$x^{<1>}, \ldots, x^{<t>}$ but not on $x^{<t+1>}, \ldots, x^{<365>}$.

D. Uni-RNN, because the value of $y^{<t>}$ depends only on $x^{<t>}$ and not the other day's weather.

(f) (2%) *Suppose you are training an LSTM. You have a 10000 word vocabulary, and are using an LSTM with 100-dimensional activations a. What is the dimension of u at each time step?*

A. 1
B. 100
C. 300
D. 10000

Warning: This question has negative score of -1 for an incorrect answer. (minimum score -1)

Solution:

A. 1
B. 100
C. 300
D. 10000