

i Information

Exam: DAT550 Data Mining

Date: May 12, 2020

Duration: 2 hours

Supporting material: All technical support material is permitted. You are not allowed to get help from other people when working on your exam assignment. We are also reminding you that you, when registering for the semester, you signed that you have read and understood the rules for cheating and plagiarism in the Exam Rules and Regulations at the University of Stavanger. Plagiarism control will be carried out.

The exam consists of 5 sections with 13 problems on a total of 13 pages, amounting to 100 points. The 6th section is intentionally left blank so that you can use it to show any calculations.

Important contact information during the exam: If something should happen during the exam and you need help, you can call one of the phone numbers below, both if you need technical support, administrative support or have questions for.

1. course responsible. Vinay Setty, tlf. 518 32760, If you spot any errors in the exam or have confusion you can call me and if the error needs to be fixed by everyone I will send an email to everyone.

2. Administrative support: 51 83 17 15 / 51 83 31 33 / 92 81 65 97 / 91 78 67 16 Technical support: 51 83 20 14/51 83 20 30

Your responses should be entered into Inspira Assessment.

There is no negative scoring on any of the questions! Hope you appreciate that :)

For some multiple choice questions there may be multiple correct answers so pay attention.

Important Information about Exams in Inspira

If you have been given extra time on the exam, it has been added to your used. You can see the count down on the top of the page. Here you will also find your candidate number. The candidate number should be written on all pages. Remember to write page number on all pages and total page number on the first page. NB! Do not write your student number, name or anything else that can identify you in you file or on your answer sheets, only your candidate number. Inspira will secure your identity and will ensure that the evaluation is anonymous.

You can write your answers using blank, ruled or squared paper. You can download and print a squared sheet with weak squares that are appropriate for scanning from Canvas. You can also write your answers in a program such as Latex if you prefer to. The answers are to be handed in as a pdf-file.

Handing in

The exam will automatically close for uploading when the time is up. Remember that the time given includes the time it will take you to scan and upload your documents (see attachment for tips concerning scanning and uploading). We recommend that you get yourself informed of how to best carry these steps out in due time and before you start working on the assignments.

Best of luck!

1.1 **Reservoir sampling**

Consider a reservoir sampling process with r units of memory for a stream of data elements. Without loss of generality, assume that the data elements are natural numbers.

(a) Is the set of samples monotonic in the following sense: if an element is in the sample set with r units of memory, will it be guaranteed to be in the sample set also with $r + 1$ units of memory – at each possible time point?

Select an alternative

- ☐ No
- ☐ Yes

(b) Assume a special case that the data elements are binary values, i.e, either 0 or 1; no other value ever occurs in the stream. Consider a long but finite sequence of observations, say N data elements. Suppose we use reservoir sampling to estimate the fraction of 0's, and we care only about the result at the end of the sequence, that is, after having seen all n elements. Does the order in which 0's and 1's appear in the sequence matter? I.,e all 0s in the beginning and all 1s in the end, compared to random order for example.

Select an alternative

- ☐ No
- ☐ Yes

Maximum marks: 6

1.2 **PCA**

Select True or False for the following statements about PCA (Principal Component Analysis).

PCA Algorithm

- ▶ Step 3:
 - ▶ Compute SVD on the covariance matrix C
 - ▶ $[U, S, V] = \text{svd}(C)$
 - ▶ Could also use Eigen value decomposition
 - ▶ U matrix is also an $[n \times n]$ matrix
 - ▶ to reduce a system from n -dimensions to k -dimensions
 - ▶ Just take the first k -vectors from U (first k columns)

$$U = \begin{bmatrix} | & | & \dots & | \\ u^{(1)} & u^{(2)} & \dots & u^{(n)} \\ | & | & & | \end{bmatrix} \in \mathbb{R}^{n \times n}$$

29

PCA Transformation

- ▶ Next we need to find some way to change X (which is n dimensional) to z (which is k dimensional)
- ▶ (reduce the dimensionality)
- ▶ Take first k columns of the u matrix and stack in columns
- ▶ $n \times k$ matrix - call this U_{reduced}
- ▶ We calculate z as follows
- ▶ $z = (U_{\text{reduced}})^T * X$
- ▶ So $[k \times n] * [n \times 1]$
 - ▶ Generates a matrix which is
 - ▶ $k * 1$

30

(a) PCA formulation from the lecture (see the above pictures) minimises the variance of the projected data (principal components)

Select an alternative

- ☐ True
- ☐ False

(b) Principle components in the above PCA algorithm are the eigenvectors of the co-variance matrix C corresponding to largest eigenvalues.

Select an alternative

- ☐ True
- ☐ False

(c) The PCA of $A - B$, for arbitrary matrices A and B , is the PCA of A minus the PCA of B . You may assume that $A - B$ is symmetric if necessary.

Select an alternative

- ☐ False
- ☐ True

Maximum marks: 9

2.1 Entropy

Let $p = (p_1, p_2, \dots, p_n)$ be a discrete probability distribution (i.e. $p_i \geq 0$ for all i and $\sum_{i=1}^n p_i = 1$). Assume $n = 32$. Recall that the entropy of p is defined as $H(p) = - \sum_{i=1}^n p_i \cdot \log_2(p_i)$ (here we define $0 \cdot \log_2(0) = 0$ without losing generality).

What is the minimum value of the function $H(p)$

What is the maximum value of the function $H(p)$

When is the minimum of the entropy function $H(p)$ is obtained?

Select one alternative

- ☐ When p follows gaussian distribution
- ☐ When $p_i = 0.5$, for at least some i
- ☐ When p is a uniform distribution
- ☐ When $p_i = 1$, for at least some i

When is the maximum of the entropy function $H(p)$ is obtained?

Select one alternative

- ☐ When p is a uniform distribution
- ☐ when p follows a Gaussian distribution
- ☐ When $p_i = 1$ for at least some i
- ☐ When $p_i = 0.5$ for at least some i

Maximum marks: 8

2.2 Probability of an event


A random nice man called Bob feels sick and goes to the doctor, and following the protocol, the doctor tests him for the Covid-19 and the test comes positive. Bob gets worried, but the doctor warns that the coronavirus test can give false positive rests in 1% of the cases. Bob also learns later from the news reports that only 0.1% of the population is likely to get Covid-19.

What is the probability that Bob actually has Covid-19 based on the above information?

Just to be sure, Bob decides to get a second test from a different lab and unfortunately that test also comes positive. What is the probability that Bob actually has the Covid-19 after the second positive test?

Maximum marks: 6

3.1 Deep Learning true/false

(a) You are building a binary classifier for recognizing fake news fake ($y = 1$) vs. not-fake ($y = 0$) documents. Which one of these activation functions would you recommend using for the output layer: 
(Tanh, leaky ReLU, Sigmoid, ReLU)

(b) The relu activation usually works better than tanh activation function for hidden units because it avoids vanishing gradients.

Select an alternative

- ☐ True
- ☐ False

(c) The use of bidirectional RNNs helps to learn dependencies in reverse order.

Select an alternative

- ☐ False
- ☐ True

(d) CNNs are equivariant to translation and scaling.

Select an alternative

- ☐ True
- ☐ False

(e) The gates within LSTM and GRU cells should use ReLU activation to ensure strong gradients.

Select an alternative

- ☐ False
- ☐ True

(f) A 2-dimensional CNN can not process images of arbitrary dimensions.

Select an alternative

- ☐ False
- ☐ True

(g) When using batch norm, scale and shift are trainable parameters.

Select an alternative

- ☐ False
- ☐ True

(h) Batch normalization results in neural networks initialization to be less important

Select an alternative

- ☐ False
- ☐ True

(i) In gradient descent with momentum, the velocity v is an exponentially decaying moving average of the negative gradients.

Select an alternative

- ☐ False
- ☐ True

(j) Full batch gradient descent can be fully parallelized.

Select an alternative

- ☐ False
- ☐ True

Maximum marks: 10

3.2 **Convolutions**

Consider the following matrix representation M of some image and a filter F:

$$M = \begin{pmatrix} 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 0 \\ 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 1 & 1 \end{pmatrix}$$

$$F = \begin{pmatrix} 0 & 1 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}$$

1. Apply F to M. Use a stride of 1 and zero-padding on M such that the resulting matrix has the same dimensions.
2. Apply a 4 × 4 average-pooling layer to the result of (a). Use a stride of 1 again.

Fill in your answer here

Format

B

I

U

x₂

x²

I_x

Words: 0

Maximum marks: 10

3.3 RNNs

Suppose you are given a vanilla RNN (no LSTM or GRU) with some initial state s_0 .

1. (a) Apply the RNN to an input sequence $x = (x_1, x_2, x_3)$ showing how the internal state is updated and how the outputs are computed.
- (b) Suppose you want to train an RNN with a set of sequences which are not guaranteed to have the same lengths.
 - (i) What problem are you likely to face when you want to train the model with mini- batches?
 - (ii) What is the standard way of dealing with this problem?
 - (iii) Can you think of a clever way of making the problem less frequent in the first place?

Fill in your answer here

Maximum marks: 10

3.4 **NN prediction**

Assume we have a neural network with ReLU activation function and want to perform a regression task, the weights and baiases (and therefore the structure) are given by

$$W_1 = \begin{pmatrix} 1 & 2 & 3 \\ 2 & -1 & 2 \end{pmatrix} \quad W_2 = \begin{pmatrix} 1 & 1 \\ 1 & -3 \\ 1 & 2 \end{pmatrix} \quad W_3 = \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix} \quad w4 = \begin{pmatrix} 1 \\ -1 \end{pmatrix}$$

and

$$b_1 = (1 \quad 1 \quad 0) \quad b_2 = (0 \quad 1) \quad b_3 = (-1 \quad 1) \quad b_4 = 1$$

Predict the output for the instance $x = (-1, 1)$

Remember the relu is a simple function $g(x) = \max(0, x)$ if it is a matrix apply the function element-wise

What is the output of the neural network after a forward pass?

You can show your math working here.

Maximum marks: 10

4.1 DNN applications

Select true or false

1. The number of parameters in an RNN language model grows with the number of time steps.

Select an alternative

- ☐ False
- ☐ True

2. In word2vec model each word has a separate weight

Select an alternative

- ☐ False
- ☐ True

3. A CNN can capture long-term dependencies in text sequences

Select an alternative

- ☐ True
- ☐ False

4. Deepwalk uses word2vec to train node embeddings by transforming graph into sequences of nodes using random walks

Select an alternative

- ☐ False
- ☐ True

5. Graph convolutional networks have separate weights for each node in the graph

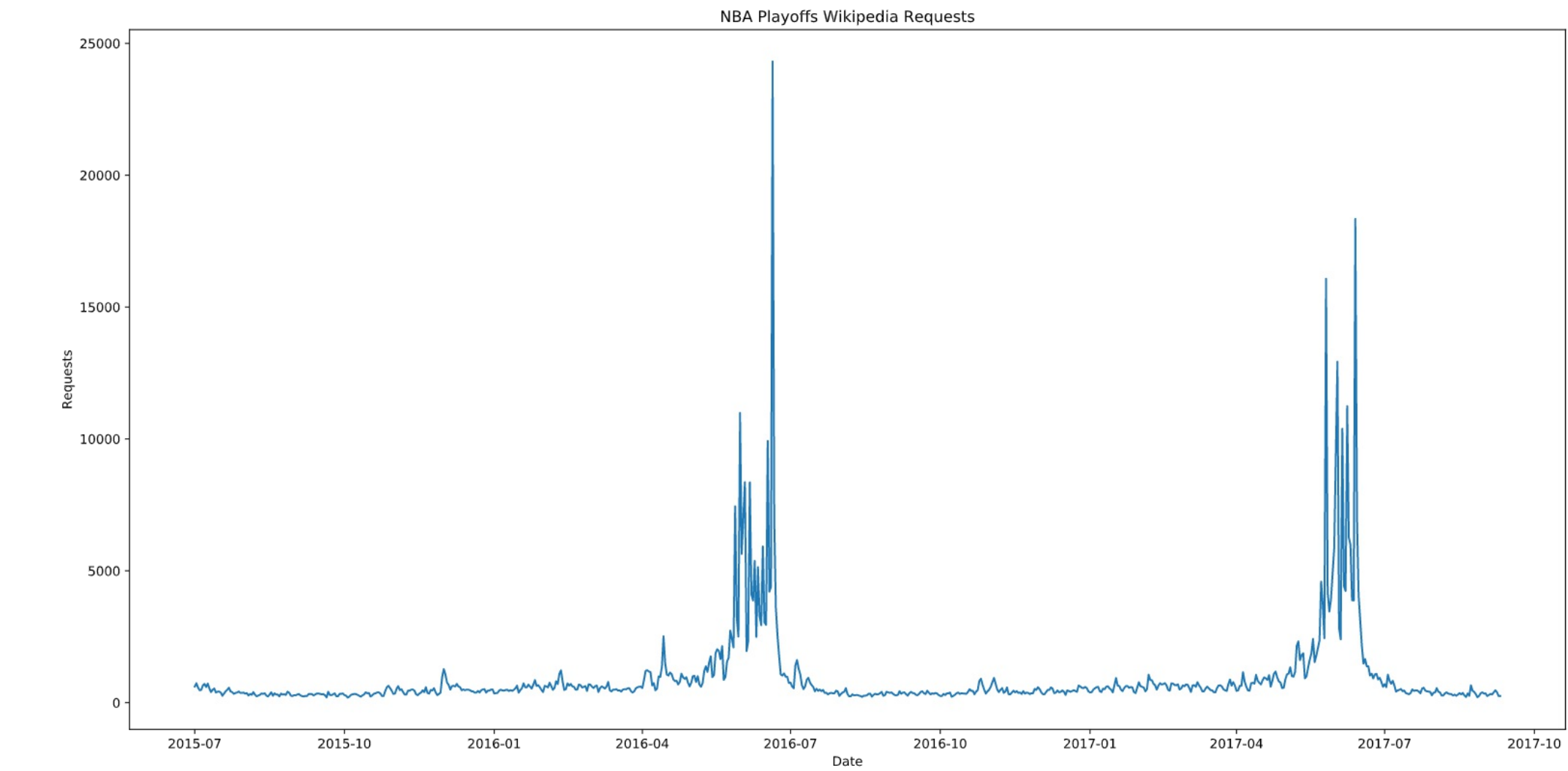
Select an alternative

- ☐ False
- ☐ True

Maximum marks: 5

4.2 Time Series analysis

(a) The time series shown here is the amount of readers of the Wikipedia article for the NBA Playoffs for a 2 year period. Which of the following statements are true or false about the given time series?



1. It has a positive trend

Select an alternative

- ☐ False
- ☐ True

2. It has a negative trend

Select an alternative

- ☐ False
- ☐ True

3. It has a seasonal component with a period of 1 year

Select an alternative

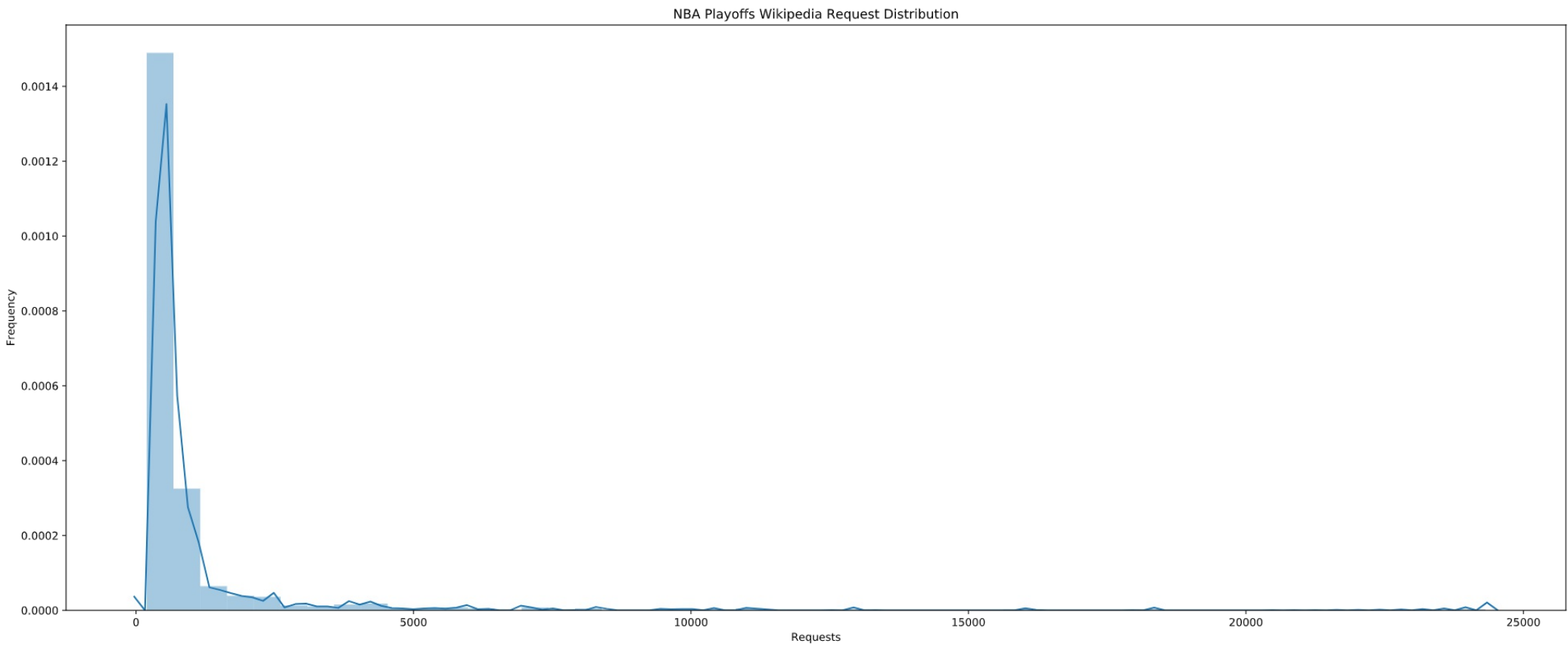
- ☐ True
- ☐ False

4. It has a seasonal component with a period of 2 year

Select an alternative

- ☐ True
- ☐ False

(b) Here, the distribution of the number of requests for the NBA playoffs is shown. Choose true or false for each statements



1. The data is skewed
Select an alternative

- ☐ False
- ☐ True

2. The data is follows normal distribution

Select an alternative

- ☐ False
- ☐ True

3. It likely has outliers

Select an alternative

- ☐ True
- ☐ False

4. The mean of the number of requests would represent the data correctly

Select an alternative

- ☐ False
- ☐ True

Maximum marks: 4

5.1 **Minhashing**

Consider the following three documents d1, d2, and d3, consisting of just one sentence each:

d1 = ⟨Peter Piper picked a peck of pickled peppers⟩
d2 = ⟨Peter picked a big peck of pickled peppers⟩
d3 = ⟨Peter Piper picked a peck of mixed peppers⟩

Step 1: Extract a set of shingles (i.e., word-level 2-grams) for d1, d2 and d3
Step 2: Compute the pairwise Jaccard similarities between the sets of shingles you obtained from step 1.

sim_{Jaccard}(d1,d2) =

sim_{Jaccard}(d1,d3) =

sim_{Jaccard}(d2,d3) =

Step 3: Using the following two hash functions as random permutations, compare the pairwise Jaccard similarities of all the documents using their minhashes.

$h_1(x) = 8x + 9 \bmod 47$
 $h_2(x) = 2x + 4 \bmod 47$

sim_{minhash}(d1, d2) =

sim_{minhash}(d1, d3) =

sim_{minhash}(d2, d3) =

Maximum marks: 12

5.2 **EM and K-means**

Consider the following six 1D data points.

data point	x
x ₀	0
x ₁	0
x ₂	3
x ₃	8
x ₄	8
x ₅	5

Apply the Expectation Maximization (EM) algorithm for the above data with K = 2. Compute the mean and variance after one complete application of the E and the M steps, starting from the M step. Start with uniform distribution i.e, assuming C1 and C2 are the two clusters (remember K = 2) for each point x_j, P(C₁|x_j) = P(C₂|x_j) = 0.5. Also assume Gaussian distributions for the two clusters and use bayes theorem to compute the probabilities.

From lecture, assuming b = P(C|x_j), Remember the mean $\mu_b = \frac{b_1x_1+b_2x_2+\dots+b_nx_n}{b_1+b_2+\dots+b_n}$

Remember the variance is computed as $\sigma_b^2 = \frac{b_1(x_1-\mu_b)^2+\dots+b_n(x_n-\mu_b)^2}{b_1+b_2+\dots+b_n}$

What is the mean of cluster C1

What is the mean of cluster C2

What is the variance of cluster C1

What is the variance of cluster C2

What is the P(C₁|x₀) and what is the P(C₂|x₀)

You may show your math working here

Maximum marks: 6

5.3 **DBSCAN**

Consider the data in the figure below, Answer to the following questions assuming that we are using Euclidean distance, that ε = 2, and minpts = 3.

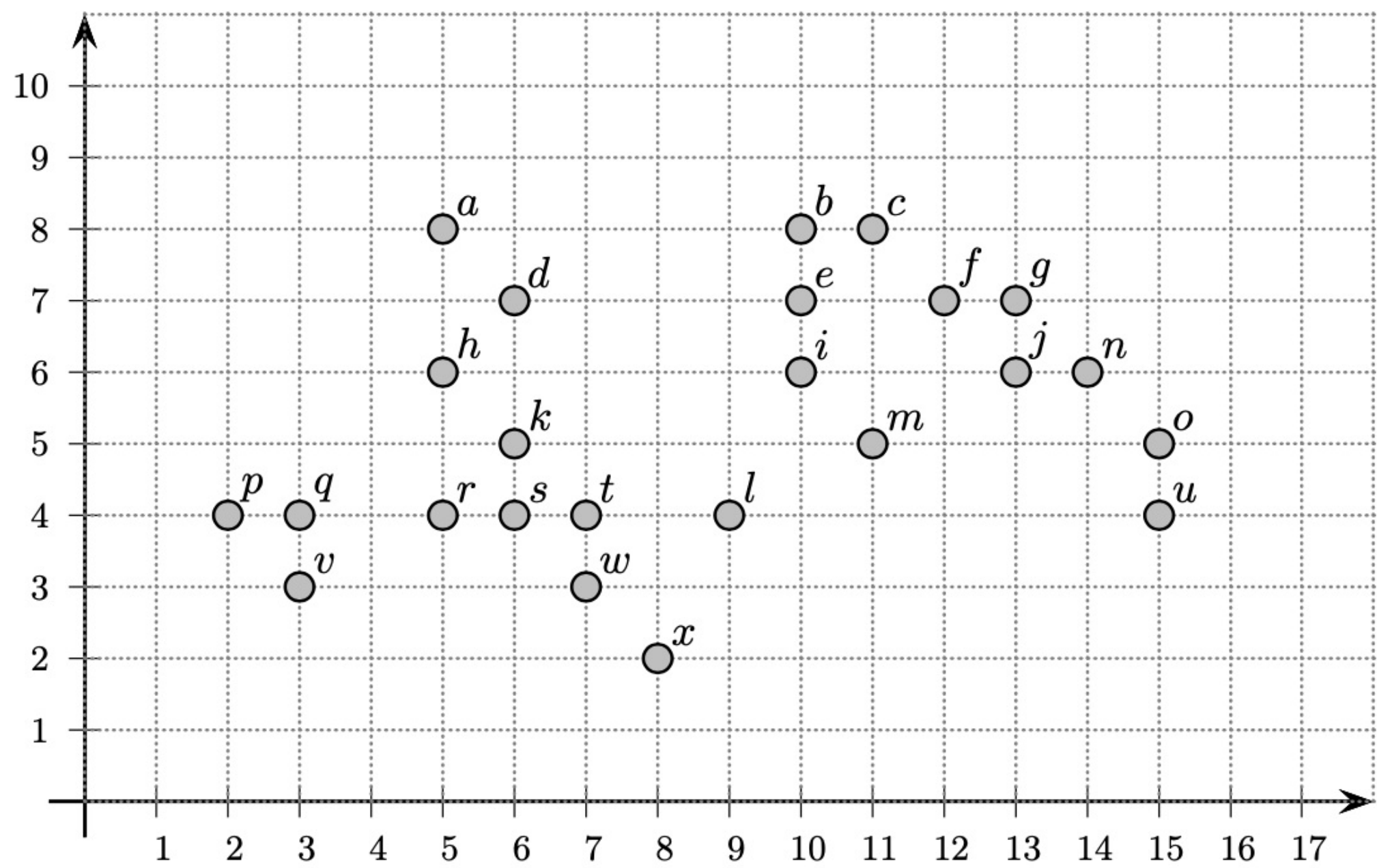


Figure: Data points with unit length grid.

Select True/False

Note you don't need to calculate the distances explicitly, just look at the distance units in grid (but follow Euclidean distance).

(a) point 'p' is a core point

Select an alternative

- ☐ False
- ☐ True

(b) point 'w' is a core point

Select an alternative

- ☐ True
- ☐ False

(c) 'a' is directly density-reachable from 'd'

Select an alternative

- ☐ False
- ☐ True

(d) 'L' (it's in upper case to make it less ambiguous) is density-connected to 'x'

Select an alternative

- ☐ False
- ☐ True

Maximum marks: 4

6.1 **Provide your explanation for numerical computations here!**
(optional)

You can use this space to provide explanation for your answers!
Fill in your answer here

Maximum marks: 0