«УТВЕРЖДАЮ»

Проректор по науке и инновациям

Габдуллин М.Т.

Центр послевузовского образования КБТУ Post Graduate Centre

Вопросы для проведения вступительного экзамена по образовательной программе докторантуры D094 «Информационные технологии» по направлению 8D06101 «Информатика, вычислительная техника и управление» на 2019-2020 учебный год

1. K-mean algorithm. Working principles.
2. Give the definition of computer system performance. How are they classified?
3. In our objective functions, we usually regularize the weight parameters of the softmax but not its biases. Explain why.
4. Discuss whether or not each of the following activities is a data mining task.

(a) Dividing the customers of a company according to their gender.

(b) Dividing the customers of a company according to their profitability.

(c) Computing the total sales of a company.

(d) Sorting a student database based on student identification numbers.

(e) Predicting the outcomes of tossing a (fair) pair of dice.

(f) Predicting the future stock price of a company using historical records.

(g) Monitoring the heart rate of a patient for abnormalities.

(h) Monitoring seismic waves for earthquake activities.

(i) Extracting the frequencies of a sound wave.

4.1.Explain recognition systems and their classification.

5.In 2-3 short sentences, describe why we initialize the weights in our model to be i)small and ii)random numbers. Hint: Think of the softmax and tanh nonlinearities.

6.Classify machine learning types. Explain supervised learning.

7.Suppose that you are employed as a data mining consultant for an Internet search engine company. Describe how data mining can help the company by giving specific examples of how techniques, such as clustering, classification, association rule mining, and anomaly detection can be applied. For each of the following data sets, explain whether or not data privacy is an important issue.

l(a) Census data collected from 1900-1950.

l(b) IP addresses and visit times of Web users who visit your Website.

l(c) Images from Earth-orbiting satellites.

l(d) Names and addresses of people from the telephone book.

l(e) Names and email addresses collected from the Web.

8.Feature space. Feature selection.

9.Your training error and cost are high and your validation cost and error are almost equal to it. Answer the following questions:

a) What does this mean for your model?
b) What actions would you take?

10.What is parallization on hardware level? Give examples.

11.Write a pseudocode for K-means and its variants and give an example of clusters.
11.1.Max-min distance algorithm.

12.History of AI, modern research fields and main problems.

13.Artificial neural network su□er(s) from the vanishing gradient problem. Circle all that apply and JUSTIFY YOUR ANSWER.

a) 1-Layer Feed-forward NN (i.e., the NN from problem set 1)
b) Very Deep Feed-forward NN
c) Recurrent NN
d) Recursive NN
e) Word2vec CBOW
f) Word2vec Skip-Gram.

14.Describe Nearest Neighbor Classifiers.

15. Class diagram in Design Software.

16.Is True or False that adding more hidden layers will solve the vanishing gradient problem for a 2 layer neural network.

17.Describe Rule based classifiers.
18.Explain MapReduce programming paradigm.

19.Cluster analysis. Ddistance formula.

20.Is True or False that adding L2-regularization will help with vanishing gradients.

21.Describe areas of application for Neural Networks (NN) and give a basic concept of NN.
22.What is Anomaly Detection in data mining. Give an approach to deal with Anomaly in data.
23. Estimation algorithm.

24.What algorithms exist to avoid False Discoveries and give an example.

25.Supervised and unsupervised machine learning.

26.Is True or False that clipping the gradient (cutting o□ at a threshold) will solve the exploding gradients problem.

27.Structural approach of software design.

28.Association Analysis . What is Apriori Principle. Illustrate on example.
29.You are consulting for a healthcare company. They provide you with clinical notes of the first encounter that each patient had with their doctor regarding a particular medical episode. There are a total of 12 million patients and clinical notes. Figure 2 shows a sample clinical note. At the time that each clinical note was written, the underlying illnesses associated with the medical episode were unknown to the doctor. The company provides you with the true set of illnesses associated with each medical episode and asks you to build a model that can infer these underlying illnesses using only the current clinical note and all previous clinical notes belonging

to the patient. The set of notes provided to you span 10 years; each patient therefore can have multiple clinical notes (medical episodes) in that period. Each note can contain any number of tokens (see Figure). Some tokens (e.g. "Meds") occur more frequently than others in the collection of notes provided to you.

History
ROS: no change in bowel/urinary habits
Meds: no Rx or OTC.
All: NKDA.
FH: mother - schizophrenia
PMH: asthma, good control. No surgeries, traumas or hospital.
SxH: sex. active with multiple F and M partners, inconsistent use of condoms, no h/o STDs
SH: NO cig/eoth. Uses PCP and ecstasy x 1y, once/week, last intake yesterday. College student.

Physical Examination
Pt is in NAD. Speech fluent, talkative, mood euphoric, affect c/w mood, behavior inappropriate. Cooperative. Appearance disheveled.
VS: WNL
HEENT: EOMI, PERRLA.
Neck: NL Thyroid Gland

You need to create a distributed representation of each patient note by combining the distributed representations of the words contained in the note

a) Given the sample note provided in Figure, how would you map the various tokens into a distributed vector representation?
b) You have the option of representing each note as the summation of its constituent word vectors or as the average of its word vectors. Both seem reasonable. What's your best course of action? Briefly justify your selection.
c) You must normalize (magnitude-wise) your wordvectors before you perform the operation you decided to do in b). Assuming you might try a standard neural network model, for which nonlinearities might that matter more?

30.Structural approach of software design. Examples.

31.Association Analysis. Explain Support Counting Using a Hash Tree.
32.Note the reasoning that is not inductive?

□ Many birds can fly. This is because they have wings.

□ Argentina is a republic; Brazil is a republic; Venezuela is a republic; Ecuador is a republic. Argentina, Brazil, Venezuela, Ecuador - Latin American states; there are no other Latin American states. Consequently, all Latin American states are republics.

□ Aluminum - solid and metal; iron, copper, zinc are also solids and metals; platinum - metal. Therefore, platinum is a solid.

□ If helium is metal, it is electrically conductive. Helium is not electrically conductive. Hence helium is not a metal.

33.Object oriented approach of software design.

34.What is a combinatorial explosion? *

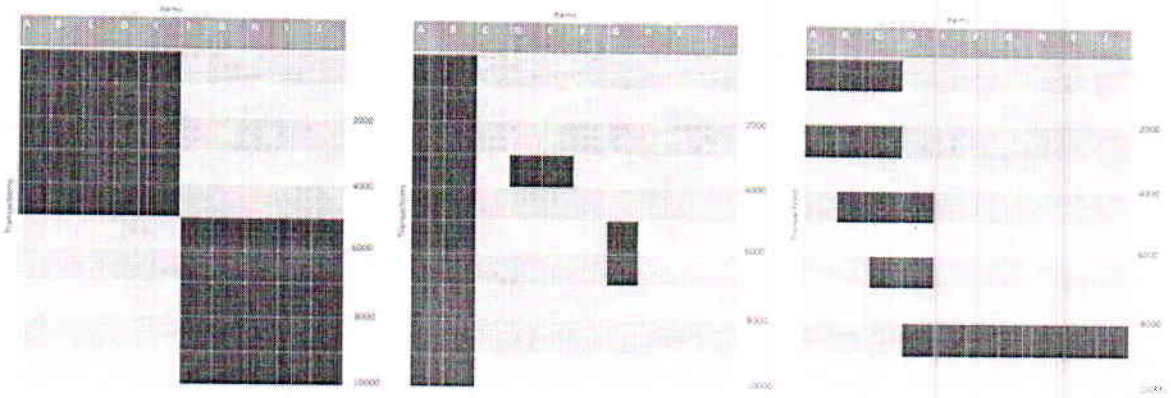
Mark only one oval.

□ A sharp increase in the time of the algorithm with an increase in the number of options

□ Machine proof of the theorem

□ Decreased number of options

1. Association Analysis. Given the following transaction data sets (dark cells indicate presence of an item in a transaction) and a support threshold of 20%, answer the following questions



a. What is the number of frequent itemsets for each dataset? Which dataset will produce the most number of frequent itemsets?
b. Which dataset will produce the longest frequent itemset?
c. Which dataset will produce frequent itemsets with highest maximum support?
d. Which dataset will produce frequent itemsets containing items with widely varying support levels (i.e., itemsets containing items with mixed support, ranging from 20% to more than 70%)?
e. What is the number of maximal frequent itemsets for each dataset? Which dataset will produce the most number of maximal frequent itemsets?
f. What is the number of closed frequent itemsets for each dataset? Which dataset will produce the most number of closed frequent itemsets?

35. Describe Hierarchical clustering and give an example.
36. Which of these is an analogy?

□ Clive Lewis was British, Christian, literary critic, professor at Oxford University and the author of scholarly treatises. John Tolkien was British, Christian, literary critic, professor at Oxford University and the author of scholarly treatises. Clive Lewis wrote wonderful fairy tales. Probably, John Tolkien also wrote wonderful fairy tales.

□ Tolkien's works contain dragons and this is fantasy. In Martin's works there are dragons and this is fantasy. Therefore, the book belongs to the fantasy genre because it has dragons.

□ All cats have a fluffy tail. Everyone who has a fluffy tail is a predator. Consequently, all cats are predators.

□ Apples grow on trees, give juice and are fruits. Pears grow on the trees, give juice and are fruits. Therefore, everything that grows on trees and gives juice is fruit.

3. K-mean algorithm. Working principles.

4. Give the definition of computer system performance. How are they classified?

37. Structural approach of software design. Examples.

38. What formula corresponds to the calculation of the error at the output layer of the neural network?

Check all that apply.

$$D_k(y_1, y_2) = (y_1 - a_1)^2 + (y_2 - a_2)^2$$

$$H(p, q) = -\sum_x p(x) \log q(x).$$

$$[f * g](t) \equiv \int_0^t f(\tau)\, g(t - \tau)\, d\tau,$$

Руководитель ЦПО_____**Естекова Г.Б.**