Final Exam

COSC 6335 *Data Mining*

December 9, 2022

Your Name:

Your student id:

Problem 1 --- Decision Trees and Supervised Learning [12]

Problem 2 --- Support Vector Machines [6]

Problem 3 --- Clustering [15]

Problem 4 --- Neural Networks [10]

Problem 5 --- Autoencoders and CNN [12]

Problem 6 --- Questions about the ICDM 2021 “best” Paper [10]

Problem 7 --- APRIORI [5]

Problem 8 --- Non-Parametric Density Estimation [6]

76:

**Grade:**



The exam is “open books” and you have 100 minutes to complete the exam. The exam will count approx. 26% towards the course grade. The use of computers and cell phones is strictly prohibited; using calculators is fine.

Write you answers on the exam paper; if you need more space use back of exam paper!

1. **Decision Trees and Supervised Learning [12]**
2. Compute the information-gain[[1]](#footnote-1) for the following decision tree split[[2]](#footnote-2) (give the formula and compute the actual value as well)! [5]

 (10,5,5) (0,5,5)

(5,0,0)

 (5,0,0)

H(1/2,1/4,1/4) – ( H(0,1/2,1/2) + 0 + 0)= ½+1/2+1/2 -1=0.5

Formula: 2.5 points; computation 2.5 points; if error in computations: at most 1 point;

b) Assume you have a classification problem with 16 classes; what is the highest value the entropy function H can take? [1]

log\_2(160=4

c) What are the characteristics of overfitting when learning decision trees? What can be done to deal with overfitting when learning decision tree models? [3]

The model is quite complex with low training error and not optimal testing error [1]

Reduce model complexity/prune tree [1]

Increase number of training examples [1]

d) A confusion Matrix of a classification model for distinguishing apples from oranges and mangos is given below:

What is the accuracy of the classification model; what is its precision for class orange; what is its recall for the class orange? It is okay to represent you answers as fractions; e.g. 17/36! [3]



Accuracy=7+2+1/36=10/36[1]

Precision Orange=2/(1+2+3)=2/6 [1]

Recall Orange=2/12 [1]

*No partial credit*

**2) Support Vector Machines [6]**

The soft margin support vector machine solves the following optimization problem:



What does the first term minimize (be precise!)? What is the purpose of C? Next, add arrows to all examples in the figure below, whose ξi values are positive---the length of the arrow should correspond with the value to the respective ξi !Are examples with a positive ξi value always misclassified; give a reason for your answer?

The inverse margin [1; if they say ‘magin’ only 0.5]; C is a hyper parameter which measures the importance of the low error objective with respect to the low inverse margin objective[1]; draw errors[1.5]. No[1]; example between the dotted line of the example’s class and hyperplane line have positive ξi values, but are classifier correctly [1.5]



**3) Clustering [15]**

a) What cluster models does EM use? How are EM’s cluster model different from K-means’ cluster models? [3]

Cluster model cluster i: mean µi covariance matrix Σi, weight/prior πi [2]

K-means just uses the cluster mean is the cluster model.

b) How does EM determine p(Cj|xi)---the probability that a point xi belongs to a cluster j? Give a verbal description how this probability is computed, outlining how cluster j’s cluster model parameters are used in this computation! [4]

It computes p(xi|Cj) by plugin xi, µj Σj into the multi-variate cluster density function and multiplies the value with the clusters prior πj (p(Cj)) to obtain p(Cj|xi).

If they say: p(xi|Cj)\*p(Cj)) without giving any details how model parameters are used: only 1.5 points.

c. Let us assume we run Fuzzy C-means (FCM) for K=2 and the centroid for cluster 1 is (1,1) and the centroid of cluster 2 is (2,3) and hyper parameter p is 2 and we use Manhattan distance; furthermore, point i is: (2,1). Compute the wi1 and wi2 for point i! [4]

wi1 = 1/(1+1/4)=0.8

wi2 =(1/4)/1.25=0.2

If any errors at most one point!

d. What is a density attractor? What role do density attractors play in DENCLUE’s clustering process? [4]

A local maximum of a non parametric density function [1]. Points in the dataset are “hill climbed” and associated with a density attractor; points in the dataset that are associated with the same density attractor form a cluster [1.5] Other answers might deserve credit!

**4) Neural Networks [10]**

a) Let us assume we have an edge in a neural network which connect node A to node B with weight wA,B. In general, what factors influence the size of the weight increase/decrease of weight wA,B? [3]

Let us assume a weight w of an edge that connect node A to node B is updated: A🡪B

The steps size of the weight update depends on

* 1. The learning rate [1]
	2. The activation on node A [1]
	3. The derivative of the activation function of node B for its linear input [0.5]
	4. The error in node [0.5]

If they say instead of c. and d. the associated error for node B this is also correct.

b) Take a look at the sub neural network consisting of nodes A, B, C, and D in the figure below; give a formula that computes the associated error ΔA for a node A. Assume the used activation function is g and its derivative is denoted by g’, and the activation of a node X is denoted by aX and the linear input of a node X is denoted by zX. First provide a general formula; then, replace general variables in the formula by their actual known values, for those which are known! [4]

 wA,B=0.5

 ΔB=0.2

 A B

 wC**,A**=1 wD,A=0.5

 C D

Formula: ΔA=g’(za)\*wAB\*ΔB=g’(zA)\*0.5\*0.2=g’(0.2+0.1)\*0.5\*0.2=g’(0.4)\*0.5\*0.2

If the formula is not expanded just 2 points. No partial credit for false expansions.

c) What role does backpropagation play in neural network learning. Limit your answer to 3-5 sentences! [3]

Backpropagation is an algorithm that back propagates the errors from output nodes to the input nodes.

The backpropagation algorithm works by computing the gradient of the loss function with respect to each weight via the chain rule, computing the gradient layer by layer, and iterating backward from the last layer.

Other answers might deserve full or partial credit, but if the answer does not mention “computing errors for nodes in the intermediate layer” give them at most 1 point!

5) Autoencoders and CNN [12]

a) Discuss the purpose and role convolutional layers and pooling layers play in the architecture on a Convolutional Neural Network (CNN)! Limit your answer to 4-6 sentences![4]

b) What can autoencoders be used for? [2]

c) What are the differences between a normal auto-encoder and variational auto-encoder (VAE) in terms of architecture, training procedure and loss function? [6]

6) Questions about the ICDM 2021 Best Paper Award Paper [10]

a) What problem do the authors of the paper solve? What are the major applications of the methods that were discussed in the paper? Limit your answer to 2-3 sentences! [3]

b) What is spatial heterogeneity? Why does learning spatial heterogeneity improve the accuracy of spatial deep learning models? Describe how the approach, described in the paper, learns spatial heterogeneity? Limit your answer to 6-10 sentences! [7]

7) APRIORI [5]

Assume the APRIORI algorithm identified the following 8 4-item sets that satisfy a user given support threshold: **abcd, acdf, acdg, adfg, bcde, bcdf, cdef, defg;** what initial candidate 5-itemsets are created by the APRIORI algorithm; which of those survive subset pruning?

acdfg (gets pruned; misses cdfg)

bcdef (gets pruned; misses bdef)

Correct candidates generated but wrong pruning answer: 3 points

If they generate too many candidates or forget candidates: at most 1.5 points

8) Non-Parametric Density Estimation [6]

a) We use Gaussian Kernel Density estimation for the following dataset

O={(0,1), (2,2)}.

We additionally, assume the bandwidth σ=1 and moreover assume Manhattan distance[[3]](#footnote-3) is used as the distance function and assume that (1,2) is the query point whose density needs to be computed. Compute fGauss ((1,2))!

**fGauss((1,2))=**

Just giving an expanded formula as your answer is sufficient; you do not need to compute the exact value of the density; just having an expression such as e-12 + e-2.5 is fine! [4]

If a single error at most 1.5 points, if patially correct; otherwise 0.

b) Non-parametric density estimation methods use a bandwidth parameter σ. What does this parameter control? [2]

It determines how quickly the influence of points on the query point decreases with distance!

1. (entropy before the split) minus (entropy after the split) [↑](#footnote-ref-1)
2. There are 3 classes, and 20 examples are associated with that node, 10 of which belong to class1, 5 belong to class2 and 5 belong to class3; after the 3-way split the first node contains 10 examples 5 of which belong to class2 and 5 of which belong to class3... [↑](#footnote-ref-2)
3. d((x1,y1),(x2,y2))=|x1-x2| + |y1-y2| [↑](#footnote-ref-3)