A Few Solution Sketches

Final Exam Version B

COSC 6335 *Data Mining*

December 7, 2022

Your Name:

Your student id:

Problem 1 --- Supervised Learning [6]

Problem 2 --- Support Vector Machines [6]

Problem 3 --- Clustering [12]

Problem 4 --- Neural Networks [10]

Problem 5 --- Autoencoders and CNN [12]

Problem 6 --- Miscellaneous Questions [11]

Problem 7 --- Similarity Assessment [8]

Problem 8 --- Parametric Density Estimation and EM [11]

76:

**Grade:**



The exam is “open books” and you have 105 minutes to complete the exam. The exam will count approx. 27% towards the course grade. The use of computers and cell phones is strictly prohibited; using calculators is fine.

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

1. **Supervised Learning [6]**

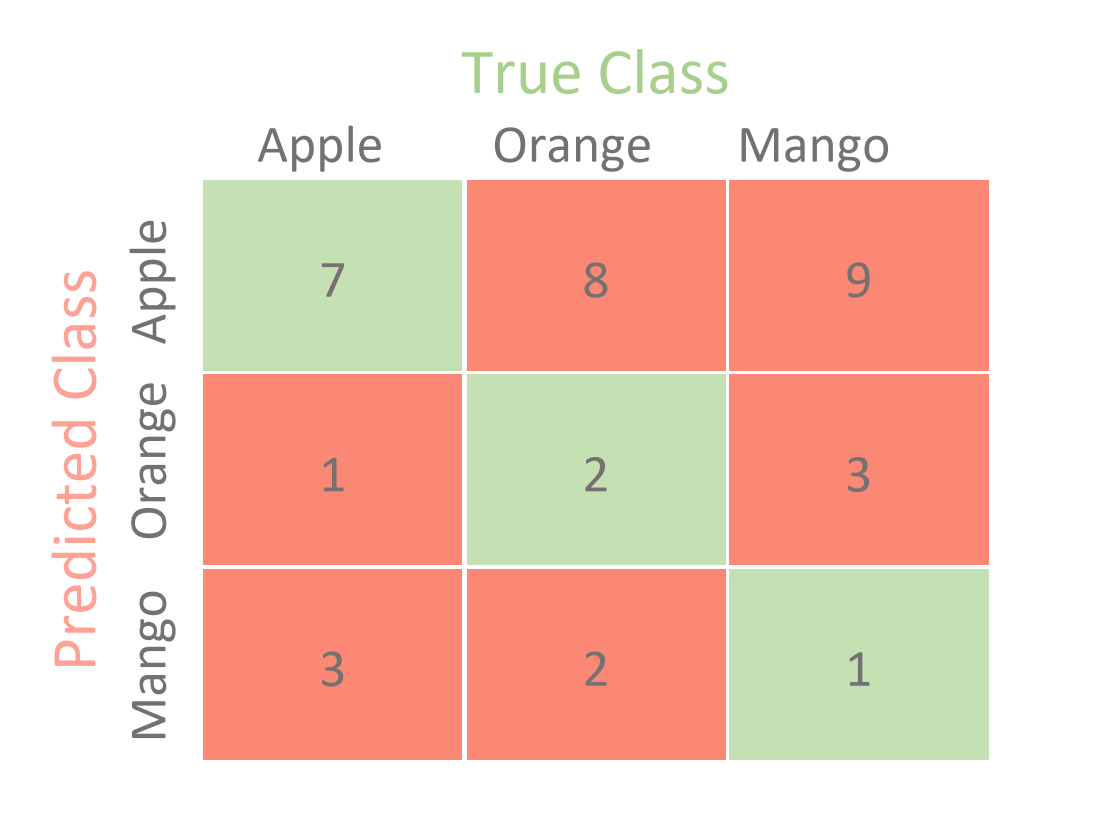
a) Assume you learn a decision tree for a dataset D and you observe overfitting. What things could done to reduce overfitting? [3]

a. create more examples for D [1.5]

b. reduce the number of nodes of the decision tree [1.5]

b) 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 Apple; what is its recall for the class Apple? It is okay to represent your answers as fractions; e.g. 17/36! [3]



Accuracy: 7+2+1/7+8+9+1+2+3+3+2+1 [1]

Precision Apple: 7/7+8+9 [1]

Recall Apple: 7/7+3+1 [1]

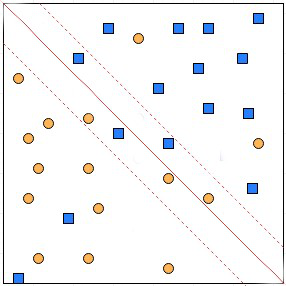
*No partial credit!*

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

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

svn-equation

What does the second term minimize (be precise!)? [1]What is the purpose of C? [1]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 ![2]Are examples with a positive ξi value always misclassified; give a reason for your answer! [2]



No answer given at the moment.

**3) Clustering [12]**

a) 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: (2,2). Compute the wi1 and wi2 for point i! [4]

Wi1= 1/1\*\*2/(1+1/4)=0.8

W*i*2=1/2\*\*2)/(1+1/4)=0.2

b) Assume we use FCM for 4 points and k=2 and the points and their weights are as follows:

Point 1: (0,0) with w11=1 and w12=0

Point 2: (3,3) with w21=0.7 and w22=0.3

Point 3: (8,9) with w31=0.1 and w32=0.9

Point 4: (12,13) with w41=0 and w42=1

Using the methods FCM uses, compute the centroid of cluster 2; give the formula and its vector. [4]

Centroid2= (1\*(12,13)+0.9\*(8,9)+0.3\*(0.3,0.3)) / 2.2

((12+7.2+0.9/2.2, (13+8.1+0.9)/2.2)=

(20.1/2.2, 22 /2.2)=(9.14,10)

Can give them 2.5 points if they use the correct formula and have a major calculation error and 3 points if they had a minor calculation error; 3.5 points if their answer is (20.1/2.2,22/2/2). At most 1 point, if they use the wrong formula

c) Assume we apply the CLIQUE algorithm to a numerical dataset with attributes A, B, C, D and E. How is CLIQUE different from more traditional clustering algorithms, such as K-means? How does CLIQUE form clusters? [4]

*Finds clusters in the subspace rather in the complete space A-B-C-D-E-F [2]*

*Clusters in subspaces are formed by a growing algorithm which starts with a seed grid-cells which are dense and adds neighboring grid-cells which are dense [2]*

Other answers might deserve full or partial credit!

**4) Neural Networks [8]**

a) 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=0.2

C D

Formula: ΔA=g’(za)\*wAB\*ΔB=g’(zA)\*0.2\*0.4=g’(0.24)\*0.5\*0.2

If not correct at most 1 point partial credit

b) The step width in NN gradient descent search varies, depending on various factors including depending on the derivative of the activation function and the activation of the input node. Explain why each of those two variations is desirable? [4]

If the gradient of the activation function is low small steps are used to not overshoot the local minimum which is characterized by a gradient of zero, [2]

If the activation of the input is low the link has very little influence on the activation of the output node; consequently, the weight of the connection is only changed a little or not at all. [2]

5) Autoencoders and CNN [12]

a) Discuss the purpose and role convolutional layers, pooling layers and fully connected players play in the architecture on a Convolutional Neural Network (CNN)! Limit your answer to 5-7 sentences![5]

b) What is the reconstruction loss of a model?[2]

||x-x\_hat|| or verbal explanation saying it measures how much the output deviates from the input for the examples in the training dataset.

c) How are variational autoencoders different from traditional auto-encoders (VAEs)? [2]

d) Kullback–Leibler (KL) divergences are used in loss functions Variational Autoencoders (VAEs) employ? What is their impact on the learnt VAE models? [3]

*The goal is to penalize latent space representations which deviate from from a simple prior distribution; e.g N(0,1) or a n-dimensional distribution with 1’s in the diagonal and 0’s off the diagonal of the covariance).*

If they say that this approach is a form of regularization to obtain simple models, without mentioning priors you can give them up to 3 points for their answer.

6) Miscellaneous Questions [11]

a. Describe the role and interactions of narrative, visuals and data in data storytelling! [4]

Should say what is on the lecture slide which discusses this relationship

b. What is the goal of association analysis? [2]

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

Candidate 5-items that are created are [3]; if they give a different answer at most 1 point partial credit.

acdfg

bcdef

cdefg

As cdfg and bdef defg are not 4-items none of the 5-item sets survives subset pruning [2]

**7) Similarity Assessment [8]**

Design a distance function to assess the similarity of customers of a supermarket; each customer in a supermarket is characterized by the following attributes[[1]](#footnote-1):

1. Ssn
2. Items\_Bought (The set of items the bought last month)
3. Amount\_spend (Average amount spent per purchase in dollars and cents; it has a mean of 60.00 a standard deviation of 40, the minimum is 0.05 and the maximum is 600)
4. Age (is an ordinal attribute taking 6 values: child, young, teenager, medium, old, very\_old)

Assume that Items\_Bought and Amount\_Spend are of major importance and Age is of a minor importance when assessing the similarity of the customers. [8]

Let φ be a function which maps child, young, teenager, medium, old, very\_old to 0:5 and ‘|’ is assumed to be the set cardinality operator; e.g. |{1,3}|=2. Use z-scores to normalize the Amount\_Spent attribute.

Solution:

d(u,v)= (1\*(1- (|u.Items\_Bought ∩ v.Items\_Bought|/(|u.Items\_Bought ∪ v.Items\_Bought)|)) + 1\*|u.Amount\_Spend-v.Amount\_Spend|/40 + 0.2\*|φ(u.Age)-φ(v.Age|/5)/2.2

Remark: The fact that Age is less important is captured in d by using a weight of 0.2 age distances.

If they do not define a mathematical sound distance function at most 2.5 points. You can still give partial credit for correct pieces of the density function. 3 points for the first part and 2 points for part 2 and 3 points for part 3.

8) Parametric Density Estimation and EM [11]

a. How do parametric density estimation techniques find the parameters of models they try to fit to a dataset? For example, if we fit a Gaussian Model to a 1D dataset how does this approach choose the mean value μ and the standard deviation σ? [4]

The idea is to select parameters; e.g. value μ and the standard deviation σ in the case of a Gaussian 1D- distribution---which maximizes the probability of the examples in D: Maximize the sample that is: ∏d∈D P(d|μ,σ)

where P is the density function of the distribution whose parameter need to be selected.

Other solutions might deserve credit.

b. What does the E-step of the EM algorithm compute and what formula does it use in its computation? [4]

It computes the probability that the i-th objects belongs to the j-th clusters (for each object/cluster pair) [2]

Formula used[2]:

c. How does EM compute the value of [3]

By plugin xi into the density functions of the j-th Gaussian model with parameters μj, and Σj.

p(xi)= *Ν*(xi|μj,Σj)

with

\mathcal{N}(\mathbf{x}|\mathbf{\boldsymbol\mu}, \mathbf{\Sigma}) =
\frac{1}{(2 \pi)^{D/2}}
\frac{1}{ | \mathbf{\Sigma} |^{1/2} }
\exp \{ -\frac{1}{2} (\mathbf{\mathbf{x}}-\mathbf{\boldsymbol\mu)}^T \mathbf{\Sigma}^{-1} (\mathbf{\mathbf{x}}-\mathbf{\boldsymbol\mu)} \}

1. E.g. (111234232, {Coke, 2%-milk, apple}, 42.42, ‘medium’) is an example of a customer description. [↑](#footnote-ref-1)