COSC 3337

Review on Nov. 1, 2022

For the Nov. 3 Midterm2 Exam

**Remark: This review does not review much DBSCAN and does not review at all non-parametric density estimation; however, carefully study the respective slides and also the GHC task slides which centered on non-parametric density estimation for Midterm2.**

**1) Neural Networks**

a) Describe how multi-layer neural networks, consisting of 3+ layers learn a model for a training set! Limit you answer to at most 9 sentences! [7]

Neural network learning tries to find weights that minimize the error in the neural network prediction for a training set [1]. Neural networks employ gradient decent hill climbing to find the “best” weights. [1]. In particular, Neural network learning adjust weights using the gradient of the error function of the training set [1]; the search starts with a random initial weight vector and weights are adjusted in the direction of the steepest negative gradient of this error function---that is weights are updated accordingly moving in the direction that reduces the error the most [2]; The step width of this weight update depends on the learning rate and other factors [1]. In order to apply this procedure, the error for each none-input node has to be known. As this error is not initially given intermediate for intermediate layer nodes, it is computed using the back-propagation algorithm [2].

Other observation might deserve credit. At most 7 points!

b) Assume the following subset of a neural network is given:

wA,B=0.2

ΔB=0.4

A B

Assume that B is an intermediate node of a neural network, the forward propagation activation values of nodes aA and aB are 0.5 and 0.8 and the current value of WA,B is 0.2; the associated error ΔB of node B that was computed by the back propagation algorithm is 0.4, the learning rate γ is assumed to be 0.5. First give the general weight update formula and then compute the new value of weight wA,B! [5]

Formula: wAB= wAB + γ\*aA\*ΔB [1]

where aA denotes the activation of node A; we receive as the new weight of WAB:

WAB

=0.2 + 0.5\*0.5\*0.4

=0.2+0.1

=0.3

No Partial Credit.

c) Looking at the sub neural network in the figure below; what does the back-propagated associated error ΔA for a node A depend on? Give a formula to compute its value, assuming g is the activation function of the neural network.

wA,B=0.5

ΔB=0.4

A B

wC**,A**=0.6 wD,A=0.2

C D

ΔA depends on: the associated error ΔB in the node B, the weight of the connection between A and B and the derivative g’ of the activation function g for the linear input of node A zA.

Formula:

ΔA=g’(za)\*wAB\*ΔB

=g’(zA)\*0.5\*0.4

=g’(0.12+0.08)\*0.5\*0.4

=g’(0.2)\*0.5\*0.4

We assume that g(0.2)=0.2

b) In general, what factors influence the size of the weight increase/decrease of weight wA,B—the step size of the weight update? [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.

d)You use a neural network tools and the tool indicated that the training took 500 epochs. What does that mean? [2]

During the training the ML algorithm went through the examples of the training set 500 times.

**2. K-Means and K-Medoids/PAM and Clustering in General [26]**

1. Assume we apply K-medoids for k=2 to a dataset consisting of 4 objects numbered 1,..,4 with the following distance matrix:

0 6 5 2 🡨object1

0 4 3

0 1

0 (e.g. the distance between object 2 and 4 is 3)

The current set of representatives is {3,4} (objects 3 and 4); indicate all computations k-medoids (PAM) performs in its next iteration! Does k-medoids get a new set of representatives or does it terminate in the next iteration? [6]

RS={3,4} clusters: {3} (1,2,4} SEE=2\*\*2+3\*\*2

New Represnetative sets are created

{1,4} …. SSE=3\*\*2+1\*\*2

{2,4} {2} {1,3,4} SSE=2\*\*2+1\*\*2

{1,3} …SSE=4\*\*2+1\*\*2

{2,3} … SSE=5\*\*2+1\*\*2

The SSE decreased and therefore PAN will run for another generation for the “new” representative set {2,4}

One error: at most 3.5 points; 2 errors at most 1 point.

b) Assume the following dataset is given: (1,1), (2,2) (4,4), (5,5), (4,6), (6,4) . K-Means is used with k=2 to cluster the dataset. Moreover, Manhattan distance is used as the distance function (formula below) to compute distances between centroids and objects in the dataset. Moreover, K-Means’s initial clusters C1 and C2 as follows:

C1: {(1,1), (3,3), (4,4), (6,6)}

C2: {(6,4), (4,6)}

Now K-means is run for a single iteration; what are the new clusters you obtain[[1]](#footnote-1) [4]

**d((x1,x2),(x1’,x2’))= |x1-x1’| + |x2-x2’| Manhattan Distance**

centroid C1= (3.5,3.5}

centroid C2= (5,5)

New Clusters

C1={(1,1), (3,3), (4,4)}

C2={(6,6},(4,6), (6,4)}

One error at most 1.5 points; 2 errors: 0 points

Problem 2 continued

c) Compare k-means with Hierarchical clustering; what are the main differences in the way they are forming clusters and in general? [4]

K-Means creates a single clustering and HC creates a hierarchy of object sets; that is, multiple clusterings [2]

HC creates a dendrogram by merging the closest clusters[1]; K-means creates clusters by assigning the objects in a dataset to the closest centroid [1]

**Graphical user interface, text

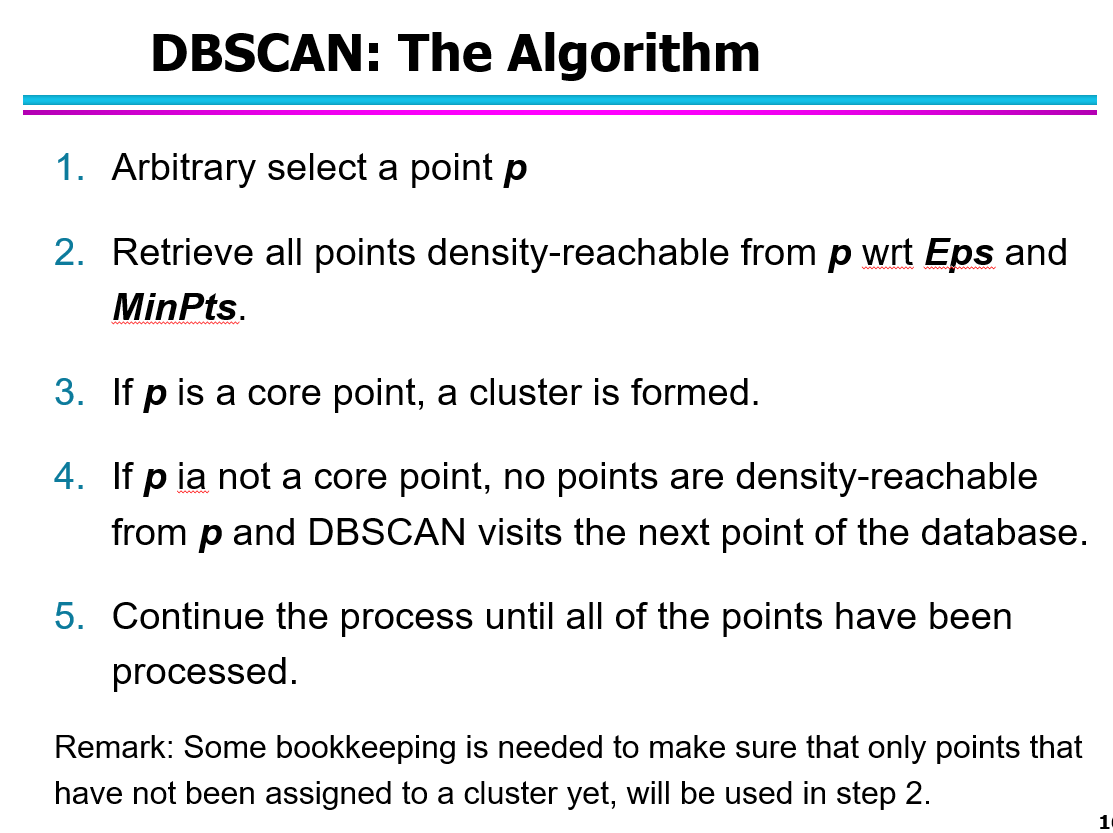
Description automatically generated  
Text

Description automatically generated  
Graphical user interface, text

Description automatically generated** **Graphical user interface, text, application, email

Description automatically generated** Graphical user interface, text, application

Description automatically generated Graphical user interface, text, application

Description automatically generated  Table

Description automatically generated with medium confidence Graphical user interface, text, application

Description automatically generated Graphical user interface, text, application, Word

Description automatically generated Graphical user interface, text

Description automatically generated Graphical user interface, text, application, email

Description automatically generated **3. Similarity Assessment [10]**

Design a distance function to assess the similarity of graduate students; each graduate student is characterized by the following attributes[7.5]:

1. Ssn
2. qud (“*quality of undergraduate degree*”) which is ordinal attribute with values ‘excellent’, ‘very good’, ‘good’, ‘fair’, ‘poor’.
3. gpa (which is a real number with mean 2.7 standard deviation is 0.5, and maximum 4.0 and minimum 1.8)
4. gender is an nominal attribute taking values in {male, female}.

Assume that the attributes qud and gpa are of major importance and the attribute gender is of a minor importance when assessing the similarity between students. Using your distance function compute the distance between the following 2 students: c1=(111111111, ‘good’, 2.7, male) and c2=(222222222, ‘poor’, 3.7, female)[2.5]!

We convert the qud values ‘excellent’, ‘very good’, ‘good, ‘fair’, and ‘poor’ using a mapping function **φ** to 4:0 and then divide number of values the ordinal attribute takes minus 1; that is, by 4 in this case**;**

finally we compute the distancesby applying the L-1 norm to the mapped values of the attribute qud.

Normalize gpa using Z-score and find distance by using L-1 norm

dgender(a,b):= if a=b then 0 else 1

Assign weights 1 to qud, 1 to gpa and 0.2 to gender attributes. We obtain:

**d(u,v) = (1\*|(u.gpa – v.gpa)/0.5| + 1\*|φ(u.qud) – φ(v.dud)|/4 + 0.2\*dgender(u.gender, v.gender))/2.2**

**No normalization (no division by the sum of the attribute weights): 1 point penalty for an otherwise correct solution; otherwise, 0.5.**

**One Error: at most 4.5 of 7.5 points**

**Two or more errors: 0-1.5 points**

For 2 students c1=(111111111, ‘good’, 2.7, male)

and c2=(222222222, ‘poor’, 3.7, female) we obtain for their distance.

d(c1,c2)= (2+2/4+0.2)/2.2=2.7/2.2≈1.2

**One Error in the computations for the example: at most 1 point  
  
  
Graphical user interface, text, application

Description automatically generated**

**4. Miscellaneous Questions**

a. What is an outlier in DBSCAN?

An outlier a non-core point that does not lie within an ε-radius of a core point[[2]](#footnote-2).

b. What is the purpose of validation sets when learning classification and prediction models? What is the purpose of test sets when learning classification and prediction models? [2]

validation set: used to select hyper parameter during training

test set: used to assess the accuracy for classifies unseen examples in straining or used to estimate the generalization error.

c) An unnamed scientist uses training and testsets for supervised learning that share 25% of the examples—25% of the examples in the training set belong to the corresponding test set. So you believe this is a good idea? Give reasons for your answer! [3]

No [1] The purpose of testsets is to estimate the generalization error/to assess how well the classifier work of unseen examples. As the approach uses examples used in training in the testset, the obtained test set accuracies are no longer a good estimate for the generalization error, as examples used in training are used to assess the testset accuracy. Comment: Using training set examples in test sets is even considered as cheating by the ML community.

5. Recurrent NN Review Question(s)

What is RNN? Explain the main difference between an RNN network and a fully connected Neural network. Why has LSTM been created as a variant of RNN?

RNNs are a family of neural networks that are used for processing sequential data. An RNN network takes the order of occurrence into account for processing the input while a FCNN throws away the relation between different time steps.

Regular RNN networks is not capable of capturing long-term dependance in a long input time series. Therefore, LSTM network has been designed to circumvent this issue. The LSTM network uses three gates to control the flow of information and hence can capture long-term dependances in input. These gates are forget gate, input gate and output gate.

Forget gate: decides what portion of the information from previous time step should be forgotten. This can be done by computing forget weights which can take a value between zero and one. A value 0 means all previous information should be thrown away and a value one means we need all previous information for the next step.

Input gate: decides what portion of new information must be added to previous information.

Output gate: converts the cell state of an LSTM network to desired output.

6. Hierarchical Clustering [8]

A dataset consisting of object A, B, C, D, E with the following distance matrix is given:

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| distance | A | B | C | D | E |
| A | 0 | 9 | 2 | 3 | 1 |
| B |  | 0 | 8 | 6 | 5 |
| C |  |  | 0 | 7 | 10 |
| D |  |  |  | 0 | 4 |
| E |  |  |  |  | 0 |

a) Assume single[[3]](#footnote-3) link hierarchical clustering is applied to the dataset? What dendrogram will be returned? [4]

A-E, A-E-C, A-E-C-D. A-E-C-D, A-E-C-D-B

One error: at most 1 point

b) Hierarchical clustering computes dendrograms; what is the dendogram? What is the value of creating dendrograms—what can they be used for? [4]

*A* ***dendrogram*** *is a* [*tree*](https://en.wikipedia.org/wiki/Tree_(graph_theory)) *diagram frequently used to illustrate the arrangement of the clusters produced by* [*hierarchical clustering*](https://en.wikipedia.org/wiki/Hierarchical_clustering)*. Edges of the dendrogram represent split/merge relationships between the nodes of the tree which represent clusters[2]*

Dendograms organize dataset hierarchically---identifying homogeneous groups at differerent levels of granularity---which is important in bio-informatics and social sciences as it allows for discovering meaningful categories/classes in data.[2]

**2. Hierarchical Clustering [8]**

A dataset consisting of object A, B, C, D, E and F with the following distance matrix is given:

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| distance | A | B | C | D | E | F |
| A | 0 | 9 | 8 | 1 | 3 | 11 |
| B |  | 0 | 2 | 6 | 5 | 12 |
| C |  |  | 0 | 7 | 10 | 4 |
| D |  |  |  | 0 | 15 | 13 |
| E |  |  |  |  | 0 | 14 |
| F |  |  |  |  |  | 0 |

a) Assume single[[4]](#footnote-4) link hierarchical clustering is applied to the dataset! What dendrogram will be returned? [7]

A D E B C F

Partially correct (only 1 error): -4

Not drawn dendrogram: -3

b) How does hierarchical clustering differ from more classical clustering algorithms, such as K-Means and DBSCAN? [4]

Hierarchical clustering computes multiple clusterings that form a hierarchy whereas K-means and DBSCAN form a single set of clusters—a single clustering. [4]; it also organizes the objects in the dataset in a hierarchy with groups and subgroups [2]

**1) DBSCAN**

**a)** Assume I run DBSCAN with MinPoints=6 and epsilon=0.1 for a dataset and I obtain 4 clusters and 5% of the objects in the dataset are classified as outliers. Now I run DBSCAN with MinPoints=5 and epsilon=0.1. How do expect the clustering results to change?

**There number of core points will usually increase and therefore the number of outliers will t usually decrease (under) 5%. New clusters might occur and multiple clusters might be merged into a single cluster; consequently, it is not clear if the number of clusters will increase, decrease, or remain the same.**

**b)** Assume we have two core points o and v that are within each other’s radius—will o and v belong to the same cluster? Now assume, that we have a border point b within a radius of a core point u—will b and u always belong to the same cluster? Give reasons for you answer! [3]

Yes, o is density reachable from v and v is density reachable from o.

No if b is also in the radius of another corepoint w, and w is not density reachable from u; in this case b might end up in the cluster formed by growing around w, if we is processed before u.

**3) Classification [16]**

a) Assume you use an ensemble approach. What properties should base classifiers have to obtain a highly accurate ensemble classifier? [2]

The classifiers should be diverse and make different kind of errors and have an “okay” but not necessarily high accuracy (above 50%).

b) Why does the AdaBoost increase the weight of an example that has been misclassified? How does AdaBoost change the weight of an example that has been misclassified by the previous classifier? Be precise! Give a verbal answer not a formula! [5]

To enhance the chance that they are classified correctly next time, leading to a classifier that classifies those examples correctly and which is therefore makes different errors in comparison to the previously learnt classifier, leading to a more diverse ensemble. [2]

The weight of this example is increased [1.5]; the weight increases in proportional to the accuracy of the previously learnt classifier---that is, if the accuracy of the previous classifier is low the weight increase will be less! [1.5]

c) 2-Layer Neural Networks do not use the Backpropagation Algorithm—why is the case? What does the back-propagated error depend on? [5]

Back propagation is used to calculate the error in the intermediate layers in multilayer neural network. Since, 2-layer neural network don’t have any intermediate layer, it does not use back-propagation. [2]

Back-propagated error depends on the derivative of the activation function for the input value that is fed into the node, the weight of the connection and the error that is back-propagated. [3]

d) Decision trees use tests <attribute> ≥ <value> (e.g. A≥33) for numerical attributes. How does the decision tree induction algorithm determine the value (e.g. 33) of the attribute A that is used in the decision tree node test? [4]

The algorithm exhaustively evaluates the merit of each possible split using purity gain, information gain, GINI…, and then selects the value with the maximum purity,…

among all the possible splits,

1. If there are any ties, break them whatever way you want! [↑](#footnote-ref-1)
2. If a non-core point lies within the ε-radius of a core point it is a border point! [↑](#footnote-ref-2)
3. When assessing the distance between clusters the minimum distance is used. [↑](#footnote-ref-3)
4. When assessing the distance between clusters the minimum distance is used. [↑](#footnote-ref-4)