CS 361 Machine Learning Exam 3 (January-May Session, 2021)

Total No.: 20 Attempt all questions Time: 30 Minutes

Points: 11/20

1

Which of these are true about measurements in agglomerative clustering? (0/1 Point)

- [x] In an attempt to detect arbitrarily-shaped clusters using single-link measurement criteria, the centroids of the current clusters may cause a chain-effect
- [x] Complete linkage avoids elongated clusters ✓
- [x] Complete linkage avoids outliers
- [ ] None of the above

2

Which of these are true about hierarchical agglomerative clustering? (1/1 Point)

- [ ] The number of clusters need to be pre-specified.

- [x] The output of the clustering algorithm depends on the choice of the similarity metric. ✓
- [ ] The number of merge operations does not always depend on the number of clusters desired
- [ ] The number of merge operations depends on the characteristics of the data set.

3

Which of the following could act as a possible termination condition(s) in K-means?
(1/1 Point)

- [x] No or negligible change of centroids ✓
- [x] Negligible fall in SSE ✓
- [x] Assignment of data points to different clusters does not change between iterations. ✓
- [ ] Only when a global optimum is confirmed using SSE

4

Comments which of the following statement(s) is/are TRUE?
(1/1 Point)

- [x] VC dimension measures the power of the learner ✓
- [ ] A classifier with lots of parameters always be a powerful classifier
- [x] The power of classifier does not necessarily directly proportional to the number of parameters. ✓
- [ ] None of the above

5

For which of the following does normalizing your input features influence the predictions?
(0/1 Point)

- [x] Decision Tree
- [ ] Neural Network ✓
- [x] Soft-margin SVM ✓
- [ ] None of the above

6

Which of the following techniques usually speeds up the training of a sigmoid-based neural network on a classification task?
(1/1 Point)

- [ ] Use full batch descent instead of stochastic
- [x] Good initialization of the weights ✓
- [ ] Increase the learning rate with every epoch
- [x] Use the cross-entropy loss instead of MSE ✓

7

Which of the following points should be generally kept in mind while choosing an algorithm for clustering?
(1/1 Point)

- [ ] Mahalanobis distance is the best possible choice as a distance function because it takes into account the weights to be assigned to different components of the data vectors via a covariance matrix
- [ ] In case k-means is used, the optimal value of k should be found out using the Manhattan method.
- [x] Data may not fully follow any distribution that's ideal to a particular algorithm, so several algorithms must be explored with parameter tuning ✓

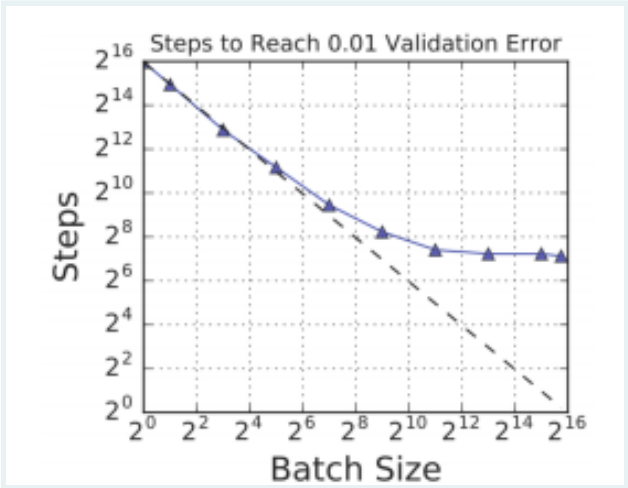
None of the above

8

Let us assume a hypothetical DNN consisting of L fully-connected layers, wherein, during backpropagation, weights of each layer are updated using the average gradient of L layers. What can be inferred about the learning process of such a model?
(0/1 Point)

- [x] Model may converge faster.
- [x] Model may not converge. ✓
- [ ] The information about the loss function is required to comment on the convergence of the model.
- [x] The model may converge faster. However, the final weights will be a linear function of initial weights before training.

9



The figure below shows the number of SGD iterations to reach a certain loss as a function of batch size. Which of the following are true (either judging from the figure or even in general, for SGD)?
(0/1 Point)

- [x] A larger batch converges faster because it presents an average view of more data

samples to the algorithm, making gradient calculation accurate and fast.

- [ ] For too large batch sizes, number of iterations do not change much as SGD tends to approach full batch gradient descent assuming that the underlined data distribution has minimal noise. ✓
- [x] Random picks in SGD makes frequent updates and causes objective function to fluctuate heavily. ✓
- [ ] SGD is especially useful when we have redundancies in the data ✓

✗

10

One (more) of the main reasons why deep learning outperforms conventional machine learning in different applications of the computer vision
(0/1 Point)

- [x] availability of high performance computing resource (GPU) ✓
- [x] conventional machine learning models are incapable of handling computer vision problems efficiently
- [ ] lacks of labelled data
- [ ] None of the above

✗

11

Which is true regarding backpropagation in a neural network?
(0/1 Point)

- [x] It is also called generalized delta rule ✓
- [x] During backpropagation, we should start with a small learning parameter and slowly increase it during the learning process.
- [ ] For an activation function f(x) = x^(1/3), backpropagation would cause trouble ✓
- [x] Error in output is propagated backwards only to update learnable parameters ✓

12

In the case of single-link and complete-link hierarchical clustering, is it possible for a point to be closer to points in other clusters than to points in its own cluster?
(1/1 Point)

- [x] Yes, for single-link clustering ✓
- [x] Yes, for complete-link clustering ✓
- [ ] No
- [ ] Can't say

13

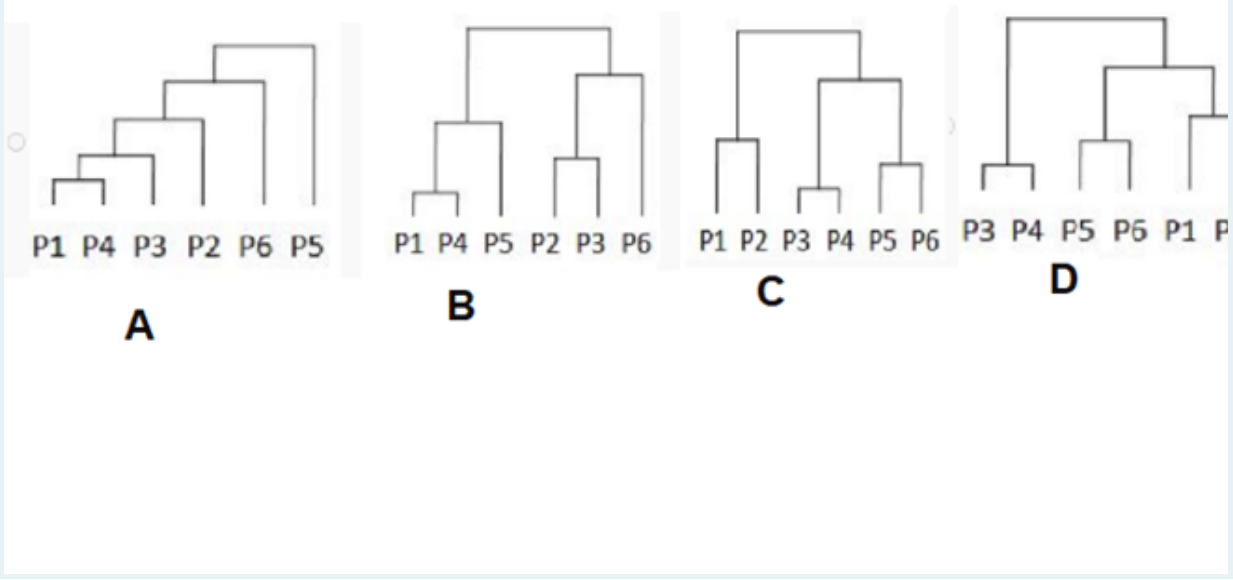
Let's say, you are using activation function X in hidden layers of the neural network. At a particular neuron for any given input, you get the output as "-0.0001". Which of the following activation function(s) could X represent?
(1/1 Point)

- [ ] ReLU
- [x] TanH ✓
- [ ] Leaky ReLU with α = -0.01
- [ ] Sigmoid

14

Given the following similarity matrix, which of the options shows the hierarchy of clusters created by single-link clustering algorithm?
(0/1 Point)

	P1	P2	P3	P4	P5	P6
P1	1.0000	0.7895	0.1579	0.0100	0.5292	0.3542
P2	0.7895	1.0000	0.3684	0.2105	0.7023	0.5480
P3	0.1579	0.3684	1.0000	0.8421	0.5292	0.6870
P4	0.0100	0.2105	0.8421	1.0000	0.3840	0.5573
P5	0.5292	0.7023	0.5292	0.3840	1.0000	0.8105
P6	0.3542	0.5480	0.6870	0.5573	0.8105	1.0000



- [x] A
- [ ] B
- [ ] C
- [ ] D ✓

15

In a soft-margin support vector machine, decreasing the slack penalty term C causes
(1/1 Point)

- [ ] More overfitting
- [ ] Smaller margin
- [x] Less overfitting ✓
- [x] Lower sensitivity to outliers ✓

16

The gradient of a continuous and differentiable function
(1/1 Point)

- [x] is zero at a minimum ✓
- [ ] is non-zero at a maximum
- [x] is zero at a saddle point ✓
- [x] decreases as you get closer to the minimum with a small learning rate. ✓

17

Choose the best applicable(s) inductive bias(es) in relevance to the neural network that outputs Y = F(X).
(1/1 Point)

- [x] Y is some non-linear function of X, where non-linearity depends on choice of activation function. ✓
- [x] Y is some non-linear function of X, where non-linearity depends on the topology of the network. ✓
- [ ] Y is some linear function of X, where the linearity depends on the underlying distribution of X.
- [ ] Y is independent of X.

✗

18

In which of these cases would K-means fail to provide good results?
(0/1 Point)

- [x] Presence of outliers in the data ✓
- [x] inefficient seed (centroid) initialization. ✓

- [ ] for higher dimensional data distribution ✓
- [ ] None of the above

19

In neural networks, nonlinear activation functions
(0/1 Point)

- [x] Speed up the gradient calculation in backpropagation compared to linear functions.
- [x] Aid in learning the nonlinear decision boundaries ✓
- [ ] Applied only to the final layer neurons.
- [ ] Always output values between 0 and 1

20

In case of shattering, which of the following is TRUE?
(1/1 Point)

- [ ] Three co-linear points (2D) can be shattered by straight line
- [x] Any two points (2D) can be shattered by straight line ✓
- [x] Three non-co-linear points (2D) can be shattered by straight line ✓
- [ ] None of the above