1) Suppose you start training a Neural Network and you observe that after several epochs, the training loss does not decrease. Select all the items that may help to solve the issue and explain why they may help (More than one may be correct).

a) Changing the learning rate
b) Changing the network architecture
c) Adding Dropout
d) Adding a regularization term to the loss function

2) For mini-batch stochastic gradient descent (SGD), by increasing the batch size, the number of epochs to reach a target loss

a. decreases
b. increases
c. does not affect the number of epochs

Explain your answer.

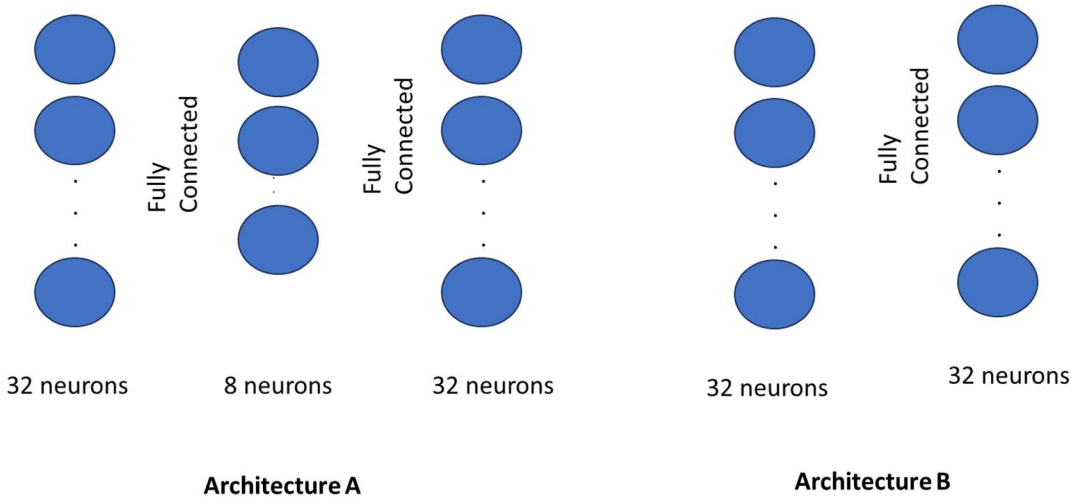
3) Assume that Sigmoid (logistic) activation function is used in our neural networks, but we observe vanishing gradients due to saturated units. Briefly explain what it means and why it happened and suggest a solution.

4) It was discussed that one of the techniques used to meet the data requirements of deep learning is Data Augmentation. Explain what it means, and which DL requirement is satisfied by Data Augmentation. Provide an example where data augmentation works and one example where it does not, or it cannot be even applied.

5) Explain the impact of choosing a very small or a very large leaning rate when SGD is being used to train a convolutional neural network (CNN). Are the issues present for both full gradient descent and stochastic gradient descent with a small batch size?

6) Assume we want to train a multi-layer perceptron neural network on a training set with 4096 samples. The learning rate is 0.01, the mini batch size is selected to be 32. We decided to have 10 epochs. Based on how many samples is the loss calculated (Loss would be a Sigma over m individual sample loss. What is the value of m?)

7) Compare the following two multi-layer perceptron networks where <u>all the neurons</u> use the <u>same linear activation</u> function.



8) Now Suppose that in Architecture A above, the hidden layer (with 8 neurons) uses a threshold (binary) activation function. The output layer uses the softmax activation function, and the loss is cross-entropy loss. What will go wrong if you train the network using gradient descent? (In terms of gradient and backpropagation).

9) Sketch a typical learning curve (e.g., amount of loss) containing both the training and validation sets, when overfitting has happened at some point. You can assume that the training and validation sets have the same size. Provide all the details of your plot.

10) Is an autoencoder used for supervised or unsupervised learning? Explain.

11) When may we use cross validation (e.g, 5-fold CV) instead of having a validation set? Why do we divide the data into training, validation, and test set?

12) Why is it preferred to have an adaptive learning rate in the optimizer which changes during the training process? Do we have an optimizer with such a characteristic?

13) Assume you have a deep multi-layer perceptron with sigmoid activation function. Can the Vanishing Gradient issue occur in this network? What about the Gradient Explosion issue? Briefly justify your answer.

14) How may splitting a dataset into train, dev and test help identify overfitting? What requirements should be met for splitting?

15) You build a neural network classifier to detect cancerous moles, so it is very crucial that the model detects cancer so the patient sees a doctor asap. Which of the following is the most appropriate evaluation metric: Accuracy, Precision, Recall, Loss Value. Explain your choice.

16) You train a neural network classifier on 200 samples. Training converges, but the training loss is very high. You therefore decide to train the network with 10,000 examples. Does your approach to fix the problem work? If so, explain how it will help and what would potentially be the most likely results of training with 10,000 examples. If not, explain why and suggest another solution.

17) You want to build a decision tree classifier. For one of the features, 20% of the values are missing (missing value). How can you handle these missing values? Would you handle the missing values differently if there were 0.2% missing values?

18) How does the presence of outliers in the dataset impact the construction and performance of a decision tree model, and what techniques can be used to address this issue?

19) How does having a shorter decision tree (versus a deeper tree for the same dataset) can improve overfitting? What are the potential drawbacks of a very short tree?

20) In the following confusion matrix for a binary classifier, what is the range of *tn* when we know that the classifier is performing better than a random classifier?

Find the recall, precision, and f1-score as a function of *tn* (if required)

Model predicts ⇩	In reality the class is ⇨	Class 1	Class 0
Class 1		27	5
Class 0		10	tn

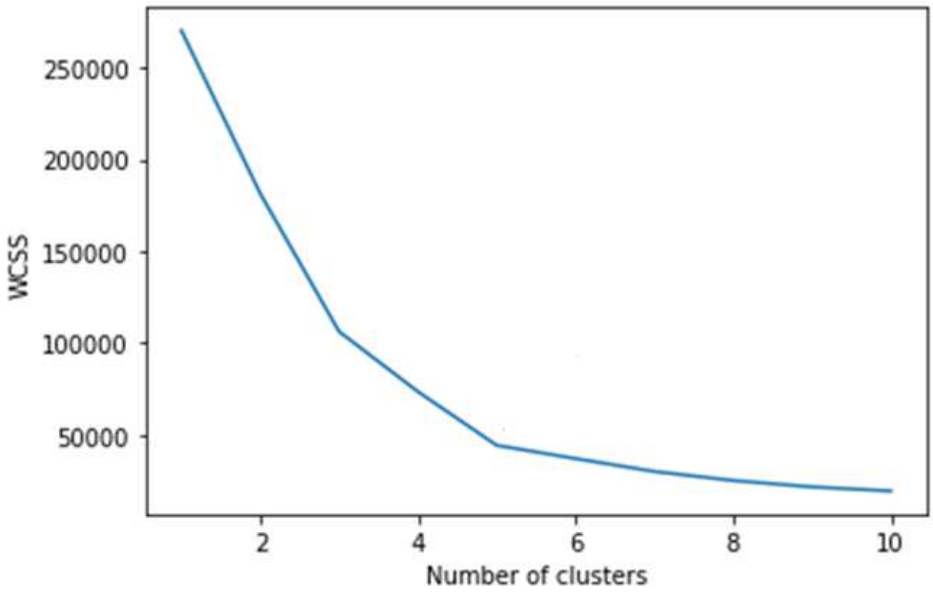
21) This is the confusion matrix of a CNN image classifier which predicts the image class among three classes: {cat, dog, fish}

Model predicts ⇩	In reality the class is ⇨	Cat	Dog	Fish
Cat		27	5	2
Dog		10	7	0
Fish		2	0	31

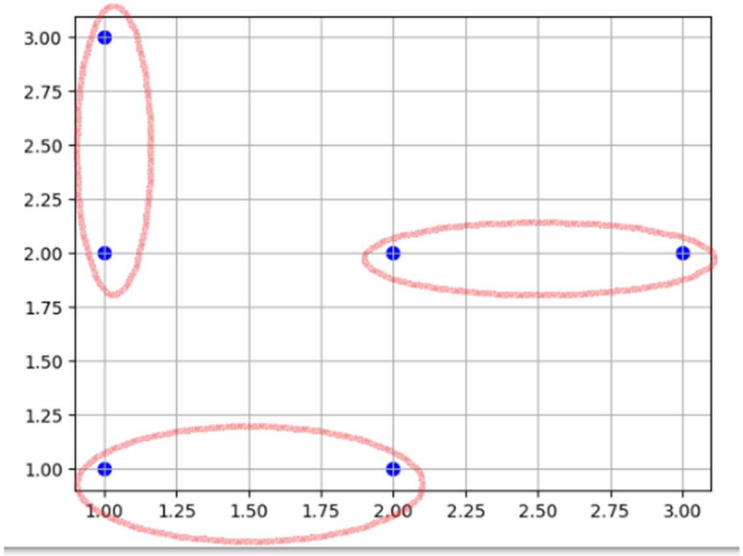
What is the accuracy of the class Cat?

What is the Recall of the class Cat?

22) Considering the following plot which reports the Within-Cluster Sum of Squares (WSS) versus the number of clusters (K), what would be the "optimal" number of clusters to take? Explain your answer.



23) After few iterations in a hierarchical clustering setup, three clusters have been formed, as can be seen in the image. In the next iteration, which two clusters will be merged to form one cluster? Assume the linkage method is "Complete Linkage" and the distance criterion is Euclidean (I know you do not like Manhattan distance, so let's assume Euclidean distance).



24) What is the adjustment impact of these parameters in a Word2Vec mode (if they increase or decrease, how is the model impacted?)

- Vector space dimensions
- Sliding window size

25) Assume you have a 10-dimensional dataset (10 features). How can you represent the samples on a 2-dim space?

26) Using a 2-gram model, how can we find the similarity of two documents using TF-IDF? Give un example. How will the vocabulary size change in comparison to the 1-gram model?

27) What is “Parameter Sharing” in the context of convolutional neural networks? Why is using “Parameter Sharing” reasonable and justified?