CS4364/6364 Intro to Machine Learning
Dr. Dobolyi

Final Exam

F2021

Please **DO NOT START** the exam until instructed, out of fairness to all students. 110 minutes.

Score: _____ / 48 pts

Name: _____

1. **Very short** answers (2 pts each == 34 points total):
   a. Explain:
      i. What a support vector machine is:

      ii. What it is used for:

   b. Explain:
      i. What layer of a neural network is a softmax activation function usually found in:

         classification stage

      ii. and why?

         used to add all activation to 1 for nomalization and easier back propagation.

   c. What is a linear layer in a neural network?

   d. What is a way a human might be able to tell that a GAN-generated image of a human face and torso is fake?

   e. What does the discriminator do in a GAN trained to generate synthetic faces?

   f. Why does dropout help when training a neural net?

   g. Why is batch learning, as opposed to single-sample (online) learning, useful when training a neural net?

      1: not sensitive to noise
      2: faster

h. What is:
   i. the formula for L2 regularization?
   ii. What does it achieve?

i. How would you solve the problem of model X underfitting its dataset?

1: NN add more layer
2: DT add more tree.
model not powerful enough

noisy feature
address string

j. Give an example of noisy labels (not features) in your data

weather

Weather
car's color.

k. How would you evaluate a binary model that has a large class imbalance in its dataset?

1: validation loss low
2: test loss high?

l. What is the difference between bagging and boosting when talking about ensembling decision tree models?

1:
1: combine desition tree
2: update weight with wrong prediction

m. Give an example of:
   i. A model that does supervised learning:
   ii. A model that does unsupervised learning:

k-means clustering, nn -> NLP

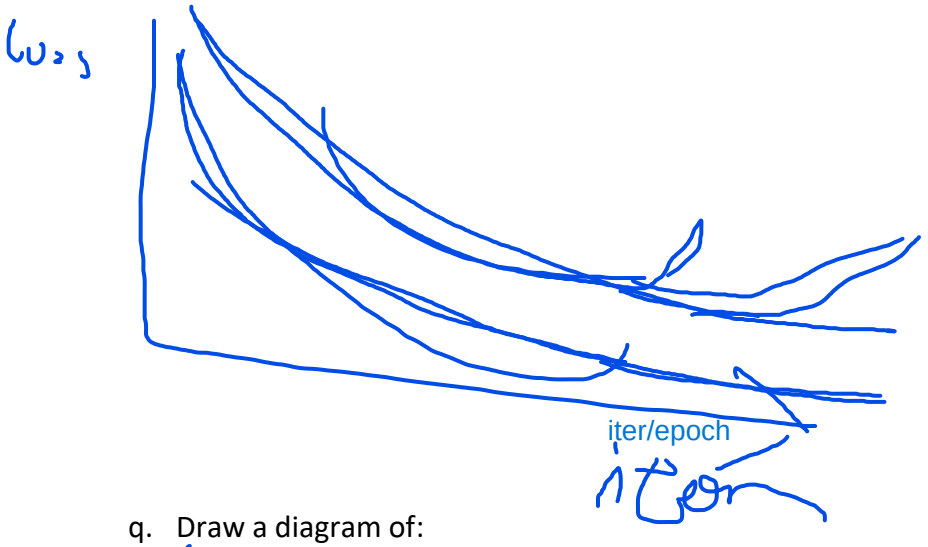
n. Name two ways you can handle missing data in your training dataset:
   i.
   ii.

remove
replace

o. Why are gradients useful when trying to update the weights of a neural network during training?

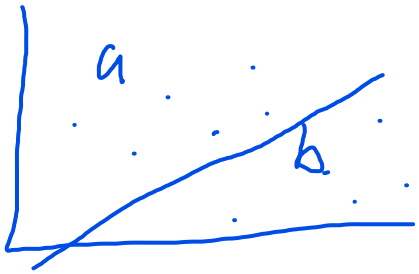
tell you the direction, and predicted distance to update weight

p. Draw a plot of training, validation and loss for a model that successfully learns something, and is not overfit. Label your axes and your lines.

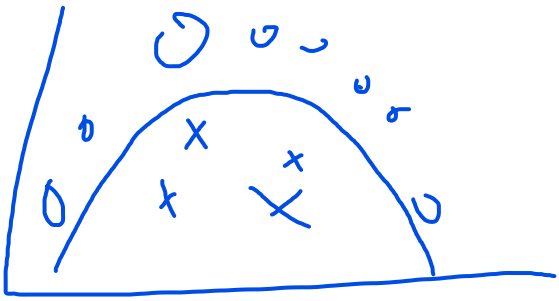


q. Draw a diagram of:

i. A two-dimensional binary dataset (points on a graph) that a linear model could classify well:



ii. A two-dimensional binary dataset (points on a graph) that a linear model could NOT classify well:



Multiple choice answers (1 pts each == 14 points total). **Optional: justify your answers.**

2. Which of the following is not true about the Markov Property in HMMs?
   a. It is possible to deterministically calculate future states knowing only the current state
   b. A state has explicit memory of all previous visited states
   c. The conditional probability of future states only depends on the current state
   d. A and B
   e. B and C
   f. A and C

3. What is true about the Bag of Word (BOW) model for representing natural language?
   a. It takes into account temporal relationships between words in a sentence
   b. All words are given equal weight
   c. It is a vector of 0s and 1s for each sentence/document
   d. A and B
   e. B and C

4. BERT, like Word2Vec and GloVe, is unable to learn temporal relationships between any pair of words in a sentence
   a. True
   b. False

5. What is true about K-means?
   a. It is a regression algorithm that uses only a subset of the features
   b. Its centroids (for each cluster) are samples in the dataset
   c. It learns the best number of clusters to have
   d. A and C
   e. None of the above

6. What is true about the K-Nearest Neighbors model?
   a. The model learns the best number for K through training
   b. K represents the number of samples in a class
   c. The model is a clustering model
   d. A and B
   e. B and C

7. The algorithm for a CNN learns the number of feature maps at each layer.
   a. True
   b. False

1:feature learning stage
1: covolution(relu+xx)+pooling
2:classification stage
1: softmax

8. In this course we saw how dataset augmentation is used to generate identical copies of images for training CNNs, especially for classes with fewer samples.
    a. True
    b. False

9. Using Stochastic Gradient Descent will always be able to find the global minimum for the loss.
    a. True
    b. False

10. Batch normalization is meant to be applied to the outputs of the nodes' activation functions in a layer.
    a. True
    b. False

11. During gradient descent with the log loss function, if all inputs have the same sign, all gradients across all weights will also have the same sign.
    a. True
    b. False

12. Which of the following is true about using the zero-one loss function?
    a. Its values of 0 and 1 for the loss make it more challenging to learn good model weights as they treat all misses equally
    b. The loss function is continuous and differentiable.
    c. A and B
    d. Neither A nor B

13. How does a larger step size in gradient descent change convergence behavior?
    a. It always causes convergence time to increase.
    b. It always causes convergence time to decrease.
    c. It can cause the algorithm the oscillate around the optimal value.
    d. None of the above.
    e. All of the above

14. When training a GAN, first you train the discriminator, and then you train the generator.
    a. True
    b. False

15. Some potential ways to increase generalization of a RandomForest model are:
    a. Limit the number of trees
    b. Increase the number of samples allowed in a leaf node
    c. Artificially limit the number of features considered for splitting at each node
    d. A and B
    e. B and C
    f. A, B, and C

Extra credit: Name two things you learned from the group project presentations:

1
2:

16. We should strive to always reduce the number of features in our models to the minimal amount necessary to make good predictions because:
    a. Such a model will generalize better
    b. Such a model will get the highest-scoring answer on the holdout dataset
    c. Such a model is easier to interpret, in terms of what features it thought were most important
    d. A and B
    e. B and C
    f. A, B, and C