EE 769 Intro to ML

End Semester Examination, 2021-May-06, 30+3 points

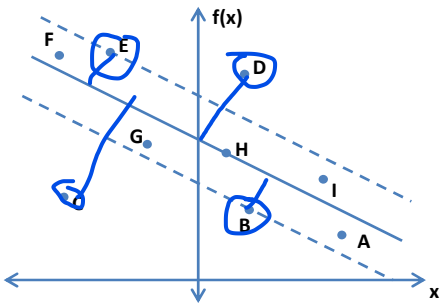

Figure 1

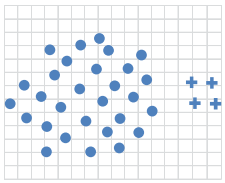

Figure 2

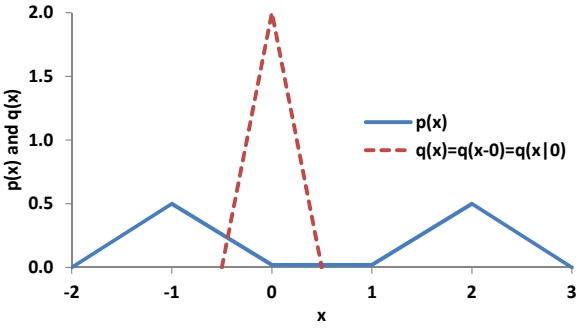

Figure 3

1. In Figure 1 depicting a linear support vector regression, for each point, write down: (a) if it is a support vector, and (b) the value or range of the loss. [1.5+1.5]

2. In Figure 2, suppose that you expect all the points with a circular marker to be in a first cluster, and those with a + mark to be in a second cluster. For each of the following clustering algorithms – (a) k-mean, (b) fuzzy c-means, and (c) DBSCAN – write whether such a result can be achieved or not. If so, then suggest the hyperparameter settings for which the desired result is possible. If not, then give reason. Note: the grid is shown for reference of distances. [1+1+1]

3. In Figure 3, the target distribution $p(x)$ (from which we want to draw the samples) and the proposal distribution $q(x|x') = q(x - x')$ are shown for a Metropolis (or Metropolis-Hastings) algorithm. If a Markov chain is initialized at $x = 2$, will it be able to sample from around the other mode at $x = -1$? If yes, then explain how. If not, then suggest how $q(x|x')$ should be modified to rectify the situation. [1+1]

4. For an n-ary classification neural network, if the L2 loss is used instead of the cross entropy loss, then will it lead to any useful training? Explain. [1.5]

5. Assume that you have a decision tree (cascade classifier), where each internal node is not a threshold classifier, but a neural network. If it is guaranteed that each neural network performs better than chance at classifying the training data, then:
   (a) What is the minimum training error that can be achieved? [1]
   (b) What is the maximum depth of such a tree? [1]
   (c) Give at least two ways to regularize such a classifier. [1]

6. What are the two conditions that individual models in an ensemble should meet in order to guarantee that the result is an improvement over the constituent models? Choose from: (a) each model should perform better than chance, (b) at least half the models should perform better than chance, (c) model outputs should be independent of each other, (d) model outputs should be dependent on each other, (e) models should be paired where one model performs correctly on exactly those samples on which the other model performs incorrectly. [1+1]

7. Given a data matrix X with an arbitrary mean, and an eigen decomposition of its covariance matrix $\Sigma = U\Lambda U^T$, write the steps to project a given test point y from the same domain as the training points x onto the the first two principal components of the training data. [2]

8. In the physical space-time of four dimensions (x, y, z, t), what is the underlying number of dimensions of points scattered uniformly on a sheet of paper, if the sheet of paper is folded into a cylindrical pipe that is moving with time without changing the relative distance of points from each other (cylindrical shape)? Justify your answer. [2]

9. For an adaboost binary classifier, explain how to incorporate arbitrary sample weights (allowing two training samples to differ in importance) for the following types of individual models (weak learners): (a) a neural network, and (b) a decision tree. [1.5+1.5]

10. List two ways of getting different neural networks for ensembling. [2]

11. Use backpropagation to differentiate mean square error loss of the function $f(x) = W_3.g(W_2.g(W_1.x + b_1) + b_2) + b_3$, where g is the ReLU function, with respect to b_2. Show the steps. All capital letters represent matrices, and small letters represent vectors. [2]

12. For gradient descent on the function $f(x_1, x_2) = x_1^2 - x_2^2$, starting at the point $(2, 1)$, with a learning rate of $\eta = 0.2$, compute two iterations with and without momentum with decay factor $\alpha = 0.9$, and comment on which one seems to be descending faster. [3]

13. For a neural network with a convolutional layer whose input feature map is of size 54×54 with 10 input channels, and 30 convolutional kernels of size 5×5, calculate the following, assuming sigmoid activation and no zero padding: (a) size of the output feature map, (b) number of weights, and (c) number of biases. [1+1+1]

14. Write the expression for cross-entropy loss, when the output of a neural network is $[0.2, 0.15, 0.65]^T$, while the target class is the third class. [1.5]