Name:

Final Exam Practice

CSE 142 – Machine Learning

December 9, 2021

4:00-6:00pm

With camera on all the time. All the teaching staffs will be watching and check the roster from times to times.

No Talking. No phones. No earphones. No internet search. No keyboard typing.

Write answers on a white paper using your pen.

Be sure to read each question carefully and provide all the information requested. **If the question asks you to explain, do so!**

You will be given extra 5 minutes to upload your answers. Exams must be turned in by 6:05pm sharp.

Good luck!

[Also see the posted several pages of information that will be provided to you with the exam.]

Final Exam Practice Questions

Note: These are just some sample questions for practice. This is not necessarily representative of the exam's length or of the exact topics/questions!

1. A machine learning test results in 50 true positives, 10 false positives, 38 true negatives, and 2 false negatives. What is the error rate? The precision? The F1 score? Would this outcome be better suited for a spam detection problem or a heart disease screening system? Explain.

True positives TP = 50
False positives FP = 10
True negatives TN = 38
False negatives FN = 2
Positives P = TP + FN = 50 + 2 = 52
Negatives N = FP + TN = 10 + 38 = 48
Estimated positives P^ = TP + FP = 50 + 10 = 60
Error rate = (FP + FN) / (P + N) = (10 + 2) / (52 + 48) = 12/100 = **.12**
Precision = TP / P^ = 50/60 = **.833**
True positive rate TPR = TP / P = 50/52 = .962
True negative rate TNR = TN / N = 38/48 = .792
Average recall = (TRP + TNR) / 2 = **.877**

If the task was to predict heart diseases as a binary classification problem, the above outcome would incorrectly predict two cases out of 50 as no heart disease which would be in 4% of the cases. Given the importance of such an error, it would unacceptable. Therefore, the outcome would be better suited for a spam detection problem.

2. For 100 4-dimensional training data points, what are the dimensions of the Gram matrix and the Scatter matrix?

Gram Matrix => k x k => 100x100
Scatter Matrix => N x N => 4 x4

3. What is the scatter matrix for the following set of 2D data points:
(3, 1), (2, 0), (-1, 0), (0, 4), (-3, -2)? (Don't forget the zero-mean step.)

Mean = (0.2, 0.6)
X = [2.8 1.8 -1.2 -0.2 -3.2; 0.4 -0.6 -0.6 3.4 -2.6]
S = XX^T = [22.8 8.4; 8.4 19.2]

4. For a k-means algorithm with k=5, when does the algorithm terminate? If you run it twice, is it guaranteed to return the same results both times?

The algorithm terminates after it reaches a stationary point, where no data points change clusters after recalculating the centroids. If it's rerun with different initial parameters (randomly chosen centroid points) it is not guaranteed to reach the same result as there are commonly many stationary points possible.

5. In a soft-margin SVM, does increasing the tuning constant C increase or decrease the chance of having misclassifications in your training set?

It decreases the chance of misclassified training points, since it weighs the slack variables heavily and thus significantly penalizes these points.

6. The basic perceptron algorithm iterates through the training data until what?

It iterates until every training point is classified correctly. Once all points are rightly classified, perceptron algorithm sets converged = True. If the data is not linearly separable, it will never converge.

7. For a training data point from the negative class, if it is on the wrong (misclassified) side of the classifier by a distance of 3.4 and homogeneous $\boldsymbol{w}$ is (3, 2, 1, -5.0), what is the margin for this point?

-3.4 (A tricky question: The margin is the distance from the line, and you're given the distance, so you don't need $\boldsymbol{w}$ and t. The margin is negative because the point is misclassified.)

8. Should unlabeled points (0.3, 1.1, -2.3) and (0.4, 1.2, -2.0) be clustered together in the same cluster or different clusters? Explain.

Clustering must always take into account the context, which means that we can't really answer this question without knowledge of the clusters and/or the rest of the datapoints. There could be a case where a cluster should contain both these two datapoints and there could be another case where the two datapoints are each located close to dense and disjoint clusters which should mean that they belong to different clusters.

9. Where D is data and H is a hypothesis:

   P(H) = 0.2 P(D | H) = 0.7 P(D) = 0.5

   what is P(H | D)? Has observation of the data increased or decreased the likelihood of the hypothesis?

P(H|D) = P(D|H) P(H)/P(D)
= 0.7*0.2/0.5
= 0.28
The likelihood has increased from 0.2 to 0.28

10. Using the Chebyshev distance metric, give the following distances:
    a. Between (2.3, -1.4) and (3.3, 4.0)
    b. Between (0, 0, 0) and (17, 14, 9)
    c. Between (1, 1, 1, 1, 1) and (0, 0, 1, 0, 0)

a) 5.4
b) 17
c) 1

11. A 1-dimensional distribution has a mean of 12.0 and a standard deviation of 2.0. Using Mahalanobis distance as the measure, what is the distance between 10.0 and 16.0?

Sqrt[(16.0-10.0)/(2.0*2.0)(16.0-10.0)] = 3.0

12. In order to build a kernel perceptron, do we modify the original perceptron algorithm or the dual perceptron algorithm? Explain why.

The dual perceptron algorithm, because it includes the dot project of training data points in its update computation. This dot product is replaced by the kernel function.

13. What is the purpose of a soft-margin perceptron or SVM, compared to the basic perceptron or SVM?

To allow for data points that creep into the margin or even across the separating function; i.e., to work for noisy and/or non-separable data sets.

14. In neural network learning, gradient descent seeks to find good values of what? Is gradient descent guaranteed to find a global optimum solution?

Gradient descent in the context of neural networks is used to calculate the best values for the weights of the edges of the neural network. Unfortunately, gradient descent is not guaranteed to find the global optimum because the error function is not always convex.

15. The sigmoid and tanh functions are often used in neural networks as activation functions. These are functions of a single variable x. What in the neural network does x represent?

The input of the sigmoid or tanh function, x, at a node is the combination (dot product) of the inputs from the previous layer and the weights of the edges connecting the nodes of that input layer to the current node.

16. Give an example of a reinforcement learning problem.

Reinforcement learning is classically employed on problems which are large and complicated, often ones that are modeled as markov decision processes. Some examples include learning to drive a car by listening to feedback, learning how to play a game with unknown probabilities by listening to feedback, and in general any problem where assessments of solution quality are easy to compute but hard to characterize (i.e we can evaluate a solution, but we have difficult taking the derivative of the problem and applying gradient descent)

17. Give two disadvantages of deep learning (compared to other machine learning methods).

1) The models are very complex and with a very large number of parameters to choose and optimize.
2) Very slow to train.
More extensive list of disadvantages in the second-to-last slide of lecture 3/9.

18. In leave-one-out cross-validation with a data set of N points, how many different models are trained?

N models. One to test every data point.

19. A five-layer neural network has how many hidden layers?

Four

20. In a fully-connected neural network with 20 input layer nodes, two hidden layers with four nodes each, and a two-node output layer, how many weights much be learned when training the network?

    20x4 + 4*4 + 4*2 = 104

21. Describe backpropagation learning in a neural net.

    (See 12/1 lecture notes)