Study Problems for Test 1
CMSC 438/691, Fall 2024

Below is a list of study problems for Test 1 covering all the material for the Test (up to and including Lecture 11). Problems starting with "GRAD:" are for CMSC 691 only, other problems are for both CMSC 438 and CMSC 691.

Test 1 will be on Wednesday, 10/9, during class time. It will be done online, in your browser at home, via Canvas/Gradescope. It will be in a form of a multiple-choice/fill-in blank (e.g. numerical result) test.

The purpose of the study problems is to provide list of topics to focus in preparing for the Test. If something is unclear as you think about answering these questions, there will be opportunity to discuss: we will go over these study problems during our Monday, 10/7, class.

1. In which situations is machine learning not the best strategy?

2. What is supervised machine learning? What is a classification problem?

3. What is a feature vector and how does it relate to examples (samples)?

4. What is feature space? What is the dimensionality of the space?

5. What is a decision boundary in two-class classification problems? If a classification model for two classes (positive and negative) is given by function f(x), which values of the function define the decision boundary?

6. For a dataset with 2 features [x1, x2], for a linear classifier with w=[0.3, -0.1], what is the formula for the decision boundary?

7. For a dataset with 2 features [x1, x2], for a linear classifier with w=[0.3, -0.1], what is the predicted class for x=[0.1, 2.1]?

8. For a one-feature, two-class classification problem with classifier y=wx, simulate the first 3 steps of the perceptron algorithm (i.e., the method from HW1).

9. Describe a two-class classification problem in the language of probabilities. What do P(x,y), P(x|y), and P(y|x) mean in the context of two-class classification? Which one is called joint distribution, and which ones are conditional distributions? What are the formulas relating them? Which one is directly useful for making class predictions?

10. What are the formulas for: a) probability of independent events, b) law of total probability?

11. In which way does Bayes' theorem connect P(x|y) with P(y|x)?

12. In machine learning, is the distribution changing over time? Is it known?

13. GRAD: What is the Optimal Bayes Classifier – how is the class decided for a sample with a given feature vector x? For an example with discrete features like the dwarfs vs hobbits, show what the optimal classifier is, and calculate the expected error of the optimal classifier.

14. What is the "histogram classifier"? Why is it not widely used? In the case of 10 binary features, how many possible feature vectors are there?

15. What parameters describe a Gaussian distribution in 1D? What about n-D? In 1-D and n-D, what are the sizes/shapes of the parameters as a function of n?

16. GRAD: What is maximum likelihood estimation? What is "likelihood" and how is it related to conditional probability?

17. GRAD: What is the "i.i.d." assumption, and how does it help in MLE?

18. GRAD: What is the difference between MLE and Bayesian learning?

19. GRAD: Show proof that the average is the maximum likelihood estimation of the mean for one-dimensional Gaussian data.

20. What is a gradient? For a function from n-D to 1D, what shape does the gradient have (a number, a vector, a matrix, and if not a number, what size of a vector or matrix is it?).

21. GRAD: Do these functions have gradients everywhere (for all inputs): f(x1,x2)=3; f(x1,x2)=x1+x2; f(x1,x2)=x1*x2; f(x1,x2)=|x1+x2|; f(x1,x2)=sign(x1*x2)?

22. What is the gradient of functions: f(x1,x2)=0, f(x1,x2)=-3, f(x1,x2)=x1, f(x1,x2)=x1-x2?

23. Describe gradient descent. Does it involve adding or subtracting the gradient? In machine learning, what function's gradient are we taking, and with respect to what variable(s)? In relation to the "isolines" of a function, in which direction does the gradient point?

24. How is "error rate" defined for a binary classifier? Is it a good objective function for optimizing using gradient descent? If not, what is the key issue?

25. How is Mean Squared Error defined – what is the formula? How is it related to the error rate? Are there issues with using MSE as a metric for classification problems?

26. In a linear model $w^T x+b$, in which way b influences the decision boundary? Does it affect the orientation of the boundary? Does it affect the distance of the boundary from the (0,0,…) coordinates?

27. In a linear model in 4D space, if w = [2,1,0,-1], which feature(s) are irrelevant for predictions?

28. In a linear model $f(x)=w^Tx+b$, on which side of the decision boundary is the negative class, in relation to the direction of vector w? In which direction does the function f(x) increase most steeply? In which direction does the value of the function f(x) stay constant?

29. For a linear model $f(w,b;x)=w^Tx+b$ and a single sample with three features, with class y=+1 (or y=-1), if the feature values are x=[1,0,-1], what is the gradient of f() with respect to w_1, w_2, w_3 & b

30. What are the issues with error metric that allows negative values (i.e., rewards for correct predictions)?

31. What is the difference between “risk” and “empirical risk”, what is “loss”, and “empirical risk minimization”?

32. What is the shape of the bipolar/unipolar sigmioid functions? What value do they have at 0? To what values do they converge at +/- infinity? Are they differentiable?

33. GRAD: What is the problem with MSE over sigmoid? Which term in the update rule is problematic, and what the problem is?

34. In the formulat for binary cross-entropy loss: -[y log(a) + (1-y)log(1-a)], what values can y take? What values can a take? What do “y” and “a” mean, and where are they obtained from when they are plugged in into the formula?

35. What are the properties of logistic loss/binary cross entropy loss?