| **Name:** | **UCSC ID:** |
|---|---|
| | |

# Practice Midterm Exam Questions

## CSE 142 – Machine Learning

## November 1, 2021

## 14:40-15:40pm

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

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 15:45pm sharp.

Note that we provide as many questions as we can for you to practice, but the midterm exam will have less questions to fit in 1-hour duration.

Good luck!

## Sample Midterm Questions

***NOTE: The answers here are rather brief – you may need to explain your answers more thoroughly than these!***

1. [4 points] What is machine learning? Briefly describe the three key components of a learning problem?

Machine learning is the design and analysis of algorithms that improve their performance at some task with experience. (Description of the task, experience/data, performance.)

2. [2 points] Give an example of an unsupervised discrete learning problem.

Clustering

3. [3 points] How are training, validation, and testing data sets typically used in developing a machine learning algorithm or application?

Training data is used to train ML models, validation data to choose among the alternative models, and testing data to evaluate the chosen model's performance,

4. [4 points] Describe inductive learning and deductive learning. Which do we focus on in machine learning and why?

See 1/5 lecture notes.

5. [4 points] Briefly discuss the relationship between overfitting and generalization in machine learning. What is likely to have lower error on the training data, a linear model or a higher-order polynomial model? How about on the testing data?

See 1/7 lecture notes; overfitting leads to poor generalization. A linear model will generally have higher error on the training data and lower error on the testing data (because it doesn't overfit the training data).

6. [3 points] What is the intrinsic dimensionality of a set of data?

The real (and relevant) dimensionality of the problem without noise or irrelevant data.

7. [3 points] Give an example of a predictive machine learning task and an example of a descriptive machine learning task.

Predictive ML – classification, regression
Descriptive ML – clustering

8. [6 points] A test for a new, deadly strain of anthrax (that has no symptoms) is known to be 99.9% accurate. The chances of any random person having this strain are one in a million. You get tested for anthrax during a routine medical exam, and your test comes back positive. If A is the variable that describes whether you have anthrax (true) or not (false), and T is the variable that describes the output of your anthrax test (true if you test positive, false if you test negative), what is the relatively likelihood that you have anthrax? Use Bayes' Rule.

Given:
P(T=true | A=true) = P(T=false | A=false) = 0.999
P(A=true) = 0.000001
Find:
Relative likelihood of P(A=true | T=true) and P(A=false | T=true)

P(A=true | T=true) = kP(T=true | A=true)P(A=true) = k(0.999)(0.000001) = k(.000000999)
…where k = 1/P(T=true)
P(A=false | T=true) = kP(T=true | A=false)P(A=false) = k(0.001)(0.999999) = k(0.000999999)
So it's about 1,000 times more likely that you don't have anthrax (i.e., 0.000999999/0.000000999)
(Although, note that before you got tested it was about 1,000,000 more times likely that you don't have anthrax. So your chances have gone up.)

9. [4 points] In the basic binary linear classifier (with the linear discriminant function midway between the class centroids), the centroid of our positive class is at (2, 1, 4) and the centroid of our negative class is at (4, 4, 6). The decision boundary is a plane defined by the vector **w** and the threshold t. If we now add a new positive sample at (0, -2, 2) to the training data and recompute, how will this affect the placement of the decision plane?

It will move toward the positive class, since the new sample is on the far (positive) side of the centroid of the positive class. (Thus it will move the centroid further away from the negative class, bringing the halfway decision boundary with it.)

10. [3 points] You are given the probability tables for P(data | hypotheses) and P(hypotheses). You need to choose the best model (hypothesis) from this data. What kind of decision rule is this?

This is a maximum a posteriori (MAP) decision rule, since you can calculate *k* P(hypotheses | data) from what is given and choose the hypothesis that maximizes this.

11. [4 points] Why is feature selection often performed in a machine learning problem before learning the model?

To reduce dimensionality, eliminate unneeded features, decorrelate features…

12. [3 points] In testing the binary classification model you learned from the training data, you get 70 out of 100 instances correct. 40 of those correctly estimated the positive class (the concept), and the rest correctly estimated the negative class. There were a total of 60 positive examples in the test set. What is the false positive rate? The false negative rate? The precision? The recall?

TP = 40, TN = 30, FP = 10, FN = 20, FPR = 10/40 = 0.25, FNR = 20/60 = 0.33, precision = 40/50 = 0.80, recall = TPR = TP/P = 40/60 = 0.67

13. [4 points] In learning a classifier, you use a loss function in weighing the effects of various training data instances. If the classification margin of an instance is very high, what should the loss function for that instance be (qualitatively, not a specific value)? If the margin of an instance is very low (negative), what should the loss function for that instance be? If an instance is badly misclassified, what should the loss function for that instance be?

For a large margin, it should be very low or zero. For a very low (very negative) margin, it should be high (maybe not too high to avoid bad effect of outliers). A badly misclassified instance is the same situation as the previous answer, for a very low margin.

14. You need to estimate prior probabilities for your 5-class classifier, and you have { 10, 8, 14, 9, and 5 } samples from your classes. What are your estimated prior probabilities, using Laplace correction?

11/51 = 0.22, 9/51 = 0.18, 15/51 = 0.29, 10/51 = 0.20, 6/51 = 0.12

15. Describe how a loss function (for classification or regression) can be made robust to outliers.

Instead of increasing monotonically as the margin becomes more negative, increase at first (to penalize errors) but then decrease (to not penalize large errors too much, assuming they are errors that should have minimal impact).

16. In concept learning, what is the difference between a *hypothesis* and a *concept*?

Nothing, they're the same thing. Or in some terminology, a hypothesis is an estimate of the (true) concept.

17. Give an example of a hypothesis that is not learnable by the conjunctive hypothesis space representation.

H : feature1=A or feature2=B  (CHS cannot learn disjunctions of features)

18. In our conjunctive hypothesis space learning, we generally seek the least general generalization. Compared with a more general generalization, what effect does choosing a less general hypothesis have on our false positive rate?

It reduces the false positive rate – less general hypothesis means it's less likely to label something as belonging to the concept. Thus there will be fewer non-concept instances mistakenly classified as belonging to the concept.

19. What CHS hypothesis will guarantee a *consistent* concept for any problem?

The concept $h(x)$ = False.
This is guaranteed not to allow any false positives, and thus is consistent. (Recall the definitions of complete and consistent.)

20. In a decision tree learning routine, a particular node has 8 examples of the positive class and 2 examples of the negative class. What is the impurity measure of that node?

0.72, using the entropy measure

21. A decision tree for concept *c* has five leaves with the following training examples in each leaf:

    L1: (5 pos, 2 neg)
    L2: (6 pos, 1 neg)
    L3: (3 pos, 4 neg)
    L4: (0 pos, 3 neg)
    L5: (1 pos, 2 neg)

    Using Laplace correction, give the ranking order of the leaves (from highest ranked to lowest).

Empirical probabilities:
L1: 6/9 = 0.67
L2: 7/9 = 0.78
L3: 4/9 = 0.44
L4: 1/5 = 0.2
L5: 2/5 = 0.4

Rank (high to low): L2 > L1 > L3 > L5 > L4

22. In a univariate linear regression problem, what is the geometric interpretation of the regression coefficient?

The slope of the regression line.