CSE 417T (Machine Learning): Exam 1 Practice Questions

1. You are a reviewer for the International Mega-Conference on Machine Learning for Everything, and you read papers with the following main claims. Would you accept or reject the paper? Provide a one-to-two sentence justification.
   (a) **accept / reject**. "My algorithm is better than yours since my algorithm has a lower in-sample error!"
   (b) **accept / reject**. "My algorithm is better than yours since my algorithm has a higher VC dimension!"
   (c) **accept / reject**. "My algorithm is better than yours since my algorithm has a low bias!"

2. Machine Learning Whiz Kid (MLWK) comes up to you with the following proposal for the Best Learner Ever (BLE). Given training dataset $\mathcal{D}$ with binary labels ± 1, BLE learns the following hypothesis $g(\mathbf{x})$: If $\mathbf{x} =$ some $\mathbf{x}_n \in \mathcal{D}$, then $g(\mathbf{x}) = y_n$, else $g(\mathbf{x}) = +1$. MLWK claims that since $E_{\text{in}} = 0$, as N gets large, BLE is guaranteed to get excellent generalization performance because of Hoeffding's inequality. Do you agree with MLWK? If not, then explain why not.

3. You work for Orange, a fictional maker of smartphones, and you have to develop a classifier that predicts whether some input fingerprint matches the fingerprint of a given phone's owner. Suppose classifying an input as $+1$ means that it matches while classifying an input as -1 means that it does not match. Through intensive market research, you know that everytime your classifier incorrectly says two fingerprints do not match when in fact they do, Orange loses 1 cent or 0.01 dollars. However, when your classifier incorrectly says two fingerprints match when in fact they don't, Orange loses 20 dollars.
   (a) Write down the cost matrix (see LFD 1.4.1) for this setting (you can assume that correct predictions incur 0 cost).
   (b) You decide to use logistic regression to predict the probability that an input fingerprint matches the phone owner's fingerprint. You train your model and get some hypothesis g. Suppose for some input, g tells you that the probability of a match is p. What is the expected cost (according to g) of classifying this input as $+1$? What is the expected cost of classifying it as -1?

4. Alice is trying to learn a classifier for a specific problem. She tries two different learning methods, A and B, and two different levels of regularization for each: Level 1, which is weaker regularization, and Level 2, which is stronger regularization. She has a lot of data so she splits it into training and validation sets that are large enough to give good estimates. Then she finds the training and validation errors for all four models. She sends this information to her friend, Bob. Unfortunately, it gets corrupted on the way, and some of the error numbers get erased (those that remain are correct). Bob receives the following information:

Method	Training Error	Validation Error
A1	9%	
A2		
B1		7%
B2	10%	

Unfortunately, Bob has no way of getting in touch with Alice and has to simply use this data to decide which method to use on some test data. If you were Bob, which method would you choose and why?

5. To show that the VC dimension of H is at least $d+1$, what do we have to prove.
   - ◯ There is a set of $d+1$ points that can be shattered by H.
   - ◯ There is a set of $d+1$ points that cannot be shattered by H.
   - ◯ Every set of $d+1$ points can be shattered by H.
   - ◯ Every set of $d+1$ points cannot be shattered by H.

6. In performing updates, the perceptron algorithm does not take into account the distance of an incorrectly classified example from the current hypothesis $\mathbf{w}$.
   - ◯ True
   - ◯ False

7. The selection of the initial weight vector does not affect the final output of the perceptron algorithm.
   - ◯ True
   - ◯ False

8. Suppose I have a dataset with 1000 data points, and I am interested in performing a linear regression, and computing training and test error. For training set size K, I use the methodology of randomly selecting K training examples, and using the remaining $1000-K$ as my test set. In expectation, what would you expect to happen to my training and test errors as K increases?
   - ◯ They both increase
   - ◯ Training error increases and test error decreases
   - ◯ Training error decreases and test error increases
   - ◯ They both decrease

9. You are participating in a machine learning competition. You are given a public dataset of size $100,000,000$ and have chosen a classification algorithm (running on some finite hypothesis set). Now your friend also decides to participate in the competition, and she has a private dataset of size $100,000,000$. Combined with the public dataset, she has twice the amount of the data that you have. Assume the two datasets are independently drawn.

   Assume she uses exactly the same learning algorithm you do. Suppose you have learned a hypothesis g_1, and your friend has learned a hypothesis g_2. You are interested in how well g_1 and g_2 makes predictions. In particular, you want to know the ratio:

   $$\frac{E_{out}(g_1) - E_{in}(g_1)}{E_{out}(g_2) - E_{in}(g_2)}$$

   Which of the following is the most likely value for this ratio? Why?

   $$4, \;\; 2, \;\; \sqrt{2}, \;\; 1, \quad -1, \;\; \frac{1}{\sqrt{2}}, \;\; \frac{1}{2}, \;\; \frac{1}{4}$$

10. The VC-dimension of the family of finite unions of positive intervals (i.e., predict $+1$ if the point is inside one of the intervals, predict -1 otherwise) over the real line is

    - ◯ 1
    - ◯ 2
    - ◯ 3
    - ◯ ∞

11. In general for most cases, as we increase the amount of training data available to a learning algorithm, what happens to the bias of the learning algorithm?

    - ◯ It increases
    - ◯ It decreases
    - ◯ It stays the same
    - ◯ It first increases, then decreases