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!"

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 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. In Fall 2018, there were a large number of students taking 417T (234 students), the instructor decided to hold the exam in two rooms. Students with an even student-ID will report to Wilson 214 and the students with odd student-IDs will report to Hillman 60. Assuming that a student-ID's parity is generated at an equal rate, μ:

    1. What probabilistic minimum can we guarantee that we will see no more than 55% of the students going to one classroom?
    2. If we would like to guarantee that for at least 90% of future exam splits where the results have no more than 55% of students going to one classroom, what minimum number of students would we need to make this guarantee?

6. The VC-dimension of the family of finite unions of closed intervals over the real line is

    ◯ 1
    ◯ 2
    ◯ 3
    ◯ ∞

7. 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

8. The selection of the initial weight vector does not affect the final output of the perceptron algorithm.

    ◯ True
    ◯ False

9. Suppose I have a dataset with 1000 data points, and I am interested in performing a linear regression, and computing training and test error (average sum of squares errors). 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. 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