Problem 1

Consider a learning problem with n ***boolean attributes. Let the hypothesis class be*** H ***Let*** $\mathbf{c} \epsilon H$ ***be the target concept, and*** D ***be a set of*** m ***independent, randomly drawn examples from c A hypothesis is said to be consistent with*** D ***if it has zero prediction error on the examples in*** D***. Let*** H' ***denote the subset of hypotheses such that*** $\forall\ h \epsilon H'$***generalization error*** ϵ ***() ¿*** γ ***Give an upper bound on the probability that*** $\exists\ \epsilon H'$ ***such that is consistent with*** D***. You can use the fact that*** $P\ (|\epsilon(h) - \widehat{\epsilon}(h)|{>}\gamma) \leq$ ***2****exp****(-2****$\gamma^2 m$***).***
Solution:

Assumption: $H' \epsilon H$ is finite. Let $h_1 \epsilon H'$.
We know that $\epsilon\ (h_1 > \gamma)$.
To find: The probability that h_1 correctly classifies m training examples randomly sampled from D.
Upon taking single training example such as (x_1, y_1) we get the probability of its classification with h_1 correctly as

$$P[h_1(x_1) = y_1] \leq (1 - \gamma)$$

From the above equation if we need to find the probability of h_1 being right m times -

$$P_D^m[h_1(x_i) = y_i] \leq (1 - \gamma)^m$$

Now let us consider a second hypothesis say $h_2 \epsilon H'$. The equation depiction the probability of classifying the m training correctly can be given as:

$$\begin{aligned} P_D^m[h_1 \cup h_2] &= P_D^m[h_1] + P_D^m[h_2] - P_D^m[h_1 \cap h_2] \\ &\leq P_D^m[h_1] + P_D^m[h_2] \\ &\leq 2(1 - \gamma)^m \end{aligned}$$

Now, if we have to evaluate for k hypotheses, the probability of correct classification of any one of those can be given is $\leq k(1 - \gamma)^m$ where $k < |H'|$. So,

$$P_D^m[h(x_i) = y_i] \leq |H'|(1 - \gamma)^m \text{ for every } h\ \epsilon H'$$

When we have the equations $0 \leq \gamma \leq 1$ and $(1 - \gamma) \leq e^{-\gamma}$. So,

$$|H'|(1 - \gamma)^m \leq |H'|e^{-\gamma m}$$

So, as a conclusion we can say that if we have a space of finite hypothesis $H' \epsilon H$, given a set of m randomly drawn and independent training examples according to D, where $c \epsilon H$ is the target concept, the probability is less than $|H'|e^{-\gamma m}$ for the existence of hypothesis(consistent with given m training set) with a generalization error of $\epsilon(h) > \gamma$. The Upper Bound found as result is the Blumer Bound.

Problem 2

Overfitting refers to the phenomenon of an algorithm overtly fitting the patterns in the training data which may not generalize well. On the other hand, underfitting refers to the phenomenon of an algorithm not being able to capture even the desirable patterns existing in the data (due to limitations on its representational power). In the description below, training data is the data over which the model is learned, and test data is some unseen data coming from the same distribution. Consider learning three different classifiers C_1; C_2 ; C_3 on a given data set such that C_1 has high training as well as test accuracies, C_2 has high training accuracy but low test accuracy, whereas C_3 has low training as well as test accuracies. Which one of the following statements is correct?

a. ***C_1 is overfitting whereas C_2 is underfitting.***

b. ***C_1 is overfitting whereas C_3 is underfitiing.***

c. ***C_2 is overfitting whereas C_3 is underfitting.***

d. ***C_2 is underfitting whereas C_3 is overfitting.***

e. ***None of the above.***

Justify your answer.
Solution:

c) C_2 is overfitting and C_3 is underfitting

In case of Overfitting and Underfitting we can generally refer to two types of Generalization error:

1. Square of Bias: The difference between the predicted hypothesis function and the original function is the bias component of generalization error.
   If we get a high error in training as well as in testing phase then the model is said to have a high bias and model Underfits the data. In this case the predicted hypothesis is not able to fit the training data well and hence has a high error with the training as well as test data and perform badly in both with high errors.

2. Variance: The variance in the prediction calculated using the predicted hypothesis function is the variance component of generalization error.
   If we get a low error in training phase and high error in testing phase then the model is said to have a high variance and model Overfits the data. In this case the hypothesis function fits the training data too well such that the algorithm also models the various noise in the data and hence it starts performing badly in test data.

A model that as a balanced or low bias as well as variance fits the data properly and gives good performance in terms of accuracy in both the training and testing phase as we can see in C_1. C_2 has high training accuracy but low test accuracy and hence it Overfits the data or has high Variance. C_3 has low training as well as test accuracies and hence it Underfits the data or has high Bias.

Problem 3

Suppose that you are working with a supervised machine learning (classification) problem and you have access to m examples $(x^{(i)}, y^{(i)})_{i=1}^{m}$ ***coming from some (unknown) underlying distribution. Assume that these are the only examples that you have access to. You would like to build a model to predict the y's from x's as to generalize well to future (unseen) data. You have a bunch of learning algorithms at hand and you would like to find out which of these is most suited for this problem. Since, you don't know yet about much about machine learning, the only thing that you can do is train each algorithm on a subset of the available data (called the training data), predict the accuracy of the learned model on a (possibly different) subset of the available data (called the test data) and finally declare as algorithm of choice the one having the highest accuracy on the test data. Now, consider the following scenarios to decide the train/test subsets:***

a. ***You choose the entire set of examples for training as well as for testing.***

b. ***You randomly choose half of the examples as the training set and then independently (at random) choose another half as the test set. The two sets could be overlapping since they are chosen independent of each other.***

c. ***You decide a number m' $\leq$ m.You randomly choose m' examples as your training set. The remaining m'- m examples become your test set. The two sets are disjoint in this case.***

Answer the following:

- ***Which of the above scenarios is likely to give you best performance on the unseen data? Argue for each case.***
- ***If one decides to go for alternative (c), describe the considerations that you will make in choosing m'.Note that choosing m' decides the size of the training as well as the test set (since the total number of examples is fixed)***

Solution:

a. Choosing the entire set of examples for training as well as for testing will give the worse error. Since training data is used to fit the model and testing data is used to provide and unbiased evaluation of the final model, using testing set same as the training set will give the lowest error (biased result).

b. Randomly choosing half of the examples as the training set and then independently (at random) choose another half as the test set with the uncertainty that the two sets might be overlapping will also lead to bad performance of model. Good practice says that the model should be unseen to the test data before hand and in this case due to the uncertainty that training set and testing set might get overlapped there are high chances of testing error to be lower than it should be. Also dividing the available m training set into two halves as training and test sets the model does not gets trained on the various noises present in the m examples and will give high testing error if tested on the unseen data.

c. If we randomly choose m' examples as training set. The remaining $m' - m$ examples become our test set it gives us the flexibility to decide upon the ratio depending upon the m examples present. This pattern of choosing the ratio tends to perform best as the model has not seen any test data before giving an unbiased evaluation of the model trained on the m' training set. Also the test set is chosen from the same distribution of examples which means it contains the same pattern of noise as the training set giving better estimations.