Recall the weather example we discussed in weka. Here is a typical line from an arff table (it is slightly different from the one seen in class), each representing a data point.

name="joe", day of week="fri", time of day=3:07 pm, outlook=sunny, temp=68, humidity=52, windspeed=5. A humidity count of 0-10 represents an extremely dry day, while a humidity of 100 means it is raining.

Imagine you have thousands of such data points, each representing a record of data for members of a golf club who showed up to play at different times of day, on different days of the week, under different weather conditions.

1. Which of the following is a typical question that can be answered using machine learning? (mark all that apply)
   
a. Did Jane ever play on a friday?
b. Will there be at least 5 persons showing up to play between 3 and pm when it is sunny and 75 degrees?
c. Will it rain when Bob shows up to play next Sunday?
d. How many times was it extremely dry when Sam came to play?

ANS: b and c

A betting company is taking bets that next Thursday, at 3 pm, if the temperature is 62, it is sunny and dry, and the windspeed is between 5 and 10 mph, Fred will show up to play. I decide to use a machine learning system ML1 to decide whether it is a good bet to take or not.

2. List the input and output variables of a target function f for ML1.
   
ANS: Input variables: day of week, time, temp, weather_outlook, wind; Output variable: Boolean: 1 if Fred shows up, 0 if not.

3. True or False. With my ML1 system, I can always win any bet of the kind that the bookie is taking.
   
ANS: False, you can only try to optimize your chances of winning the bet

4. Training data for the ML1 system can consist of (mark all that apply)
   
a. all the data points currently available
b. all the data points currently available, plus future data points gathered after the ML1 system is released.
c. Some significant percentage of the data points currently available.
d. All data points currently available, after they have been corrected for statistical noise.

ANS: a, c, d
5. Another gambling company is offering bets on whether 5 or more persons will show up to play between 3 and 4 pm when it is raining and 45 degrees, with wind speed greater than 10 mph. Show the inputs and outputs for the target function for a new ML system (ML2) to help me bet and win with this second company.

**ANS:** ML2 inputs: time, weather_outlook, temperature, wind; ML2 output: Boolean: 1 if 5 or more persons show up, 0 if not

6. What preprocessing, if any do you need to make to the data to make it useful for the second ML system (ML2)?

**ANS:** 1. Convert time to an integer representing the hour (e.g., map all times between 3 and 4 pm to 15 and so on); 2. Convert windspeed to a Boolean (0 if < 10, 1 if >10); 3. Convert list of people who showed up into a boolean for each hour, 1 if >-5, 0 if <5

Let us say that our ML algorithm is a simple perceptron (the PLA), and that I would like to predict whether or not at least 3 people show up for golf at any given point in time under any weather conditions.

7. Show a target function for this PLA that outputs 1 (for yes) or -1 (for no)

**ANS:** f: \{timestamp, weather_outlook, temp, humidity, windspeed\} \rightarrow \{1,-1\}

8. List the features of interest in your data for this PLA. Considering your data for the PLA as a vector, show the dimensions of the vector.

**ANS:** features of interest in the data: time of day, weather outlook, temp, humidity, windspeed, num_persons

Dimensions of vector: <time_of_day, weather_outlook, temp, humidity, windspeed, num_persons>

9. A hypothesis generated by the PLA is

\( (w_0, w_1, w_2, w_3, w_4 \text{ are real numbers, } \text{num(persons)} \text{ is the number of arriving persons for a given day, time, weather conditions; numeric_code(outlook) is an encoding of sunny as 1, cloudy as 2 and so on.}) \)

a. A set of data points, e.g., (Fred, tuesday, 3 pm, cloudy, 68 degrees, 34%, 7mph)
b. An expression of the form:

\[ w_0*(\text{num(persons)}) + w_1*(\text{temp})^2 + w_2*(\text{humidity})-1 + w_3*\text{windspeed} + w_4*(\text{numeric_code(outlook)}) \]
c. An expression of the form:

\[ w_0*(\text{num(persons)}) + w_1*(\text{temp}) + w_2*(\text{humidity}) + w_3*\text{windspeed} + w_4*(\text{numeric_code(outlook)}) \]
d. An expression of the form:

\[ w_0*(\text{excess_persons}) + w_1*(\text{temp}) + w_2*(\text{humidity}) + w_3*\text{windspeed} + w_4*(\text{numeric_code(outlook)}) \text{ where excess_persons is num(peson}-3). \]

**ANS:** d

10. A learning step in the above PLA will (mark all that apply)

a. update the data points based on the target function value
b. update the coefficients w_i, for every misclassified data point

c. Compare the hypothesis expression at a data point with the output value in the training data set

d. All of the above

ANS: b and c

11. Hoeffding's inequality implies that (mark all that apply)

a. supervised learning is not possible but unsupervised learning may be

b. given a large enough training data set of size N, we can probably produce a workable approximation to the target function

c. For larger training data sets, we have a better likelihood of reducing the error of our prediction

d. The smaller the prediction errors, the less likely we are to get to it.

ANS: b,c,d

12. Linear regression (mark all that apply)

a. is a prediction algorithm.

b. always produces a linear fit that exactly matches every data point in the training data set

c. typically uses mean square error minimization to fit the curve to the data

d. can work for at most 3 dimensional data points.

ANS: a and c

13. Error measure is (mark all that apply)

a. usually a choice made by the customer of the ML system

b. determined by the relative costs of false positives and false negatives

c. dictated by the probability distribution of noise

d. influenced by mathematical feasibility.

ANS: a,b,c,d