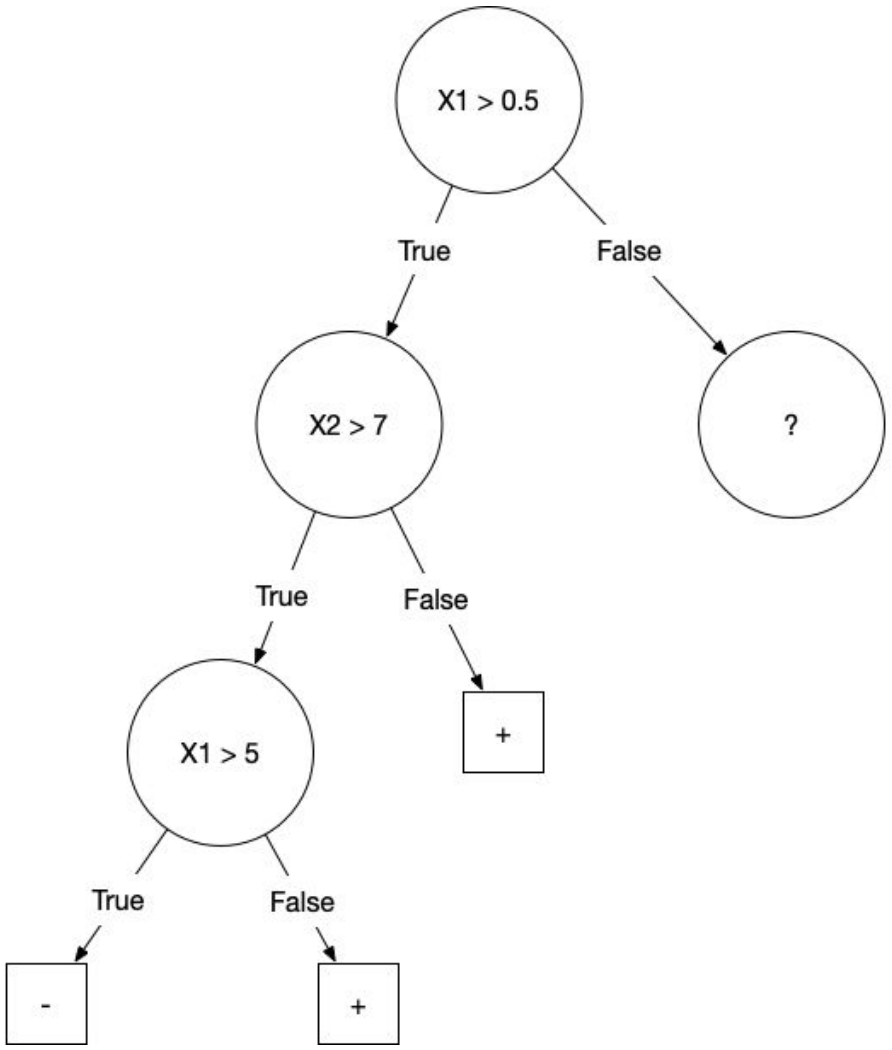
Homework 1 - Supervised Learning

To give you an idea of the kinds of questions I might ask on an exam, I've put together a few examples below. These questions are typically longer than what I would expect you to be able to do in one class period, but they should serve as good practice and indicate the level of understanding I'm expecting.

Decision Trees



X	Y
(7.0, 8.0)	-
(3.0, 7.5)	+
(-1.0, 6.5)	+
(-3.0, 6.0)	-
(2.0, 5.5)	+
(-0.5, 3.0)	+
(-2.0, 2.5)	+
(6.5, 1.0)	+
(-0.2, 0.5)	-
(-1.0,-1.0)	-

For this question, refer to the decision tree given graphically above. In this example, the features are continuous, so the attributes used for splitting (circular nodes) are greater-than/less-than tests. Square nodes are leaves, and return the indicated class (+/-). The table to the right gives 10 data points.

1. In order to fill the node labeled with “?”, assume you have three possible split choices:
    a. X2 > 6.5
    b. X2 > 0.5
    c. X2 > 0.0

    Using the given data, compute the best split among these three using Information Gain.
2. Sketch a graph that shows the 10 data points and the decision boundary of this tree using the split you found in the first part. Clearly indicate which regions correspond with which labels
3. Assuming that the children of the “?” node are leaves,
    a. What’s the misclassification error on the given data for each of the possible splits?
    b. Does the split with the lowest misclassification rate match the split with the highest Info Gain? Why or Why not?
4. The splitting attributes in part 1 were chosen by picking one of the training data points in the current subset,
    a. Give an alternative method for picking a splitting attribute, and explain (in words) how to compute it for any similar dataset (2D, real valued).
    b. Would your method produce a better set of splits for the given data? Why or why not?
5. Use the “pick-a-data-point” method to add one additional split to the tree that minimizes misclassification on the given data.
    a. Give the value for the split, and show it on your graph.
    b. What’s the new misclassification rate? Is it possible to perfectly fit the given data?

Polynomial Regression

In class our examples all assumed X and Y were both single numbers (X,Y \in R).

1. Now let the inputs be 2D (X \in R^2). How many different ways can we combine the X components to get features with **degree exactly 2**?
2. To fit a polynomial of degree **at most** 2, how many features do we need to include when we construct our matrix?
3. Generalize your answers for the first two parts to inputs in d dimensions (X \in R^d) and polynomials of degree p. Give an equation in terms of p and d.

Neural Networks

Notation:

- g: threshold function
- h: output of the learner
- $w^{(i)}_j$: the “j”th weight for the “i”th layer.

1. Using squared error your loss function looks like: E(W,X,y) = ½ (y - h(X))^2

a. Expand out h(x) in terms of a generic threshold function g() for a network with two layers.
b. Using the chain-rule, give an equation for the partial derivative of E with respect to one of the weights. Note that the equations may simplify differently for each layer

k-Nearest Neighbors (and Cross Validation)

1. For the same training data given in the first question on decision trees, sketch the decision boundary for kNN when k=1 and distance = Euclidean.
2. Using Leave-One-Out (10-fold) cross validation, compute the misclassification rate for k=1 to k=5. Which k has the best error?

Ensemble methods

1. We discussed in class that bagging tends to help with overfitting, while boosting helps with underfitting. Given the results of the cross validation that you computed in the previous section, would you expect bagging to help when k=1? Explain why or why not.
2. Part of boosting involves training the weak learners on a computed distribution, D_t. One way to make this work when using weak learners that don't naturally consider training data with "weights" is to re-sample the data according to D_t . Assuming that the underlying "weak" learner is kNN, explain how this kind of sampling and the choice of k interact.