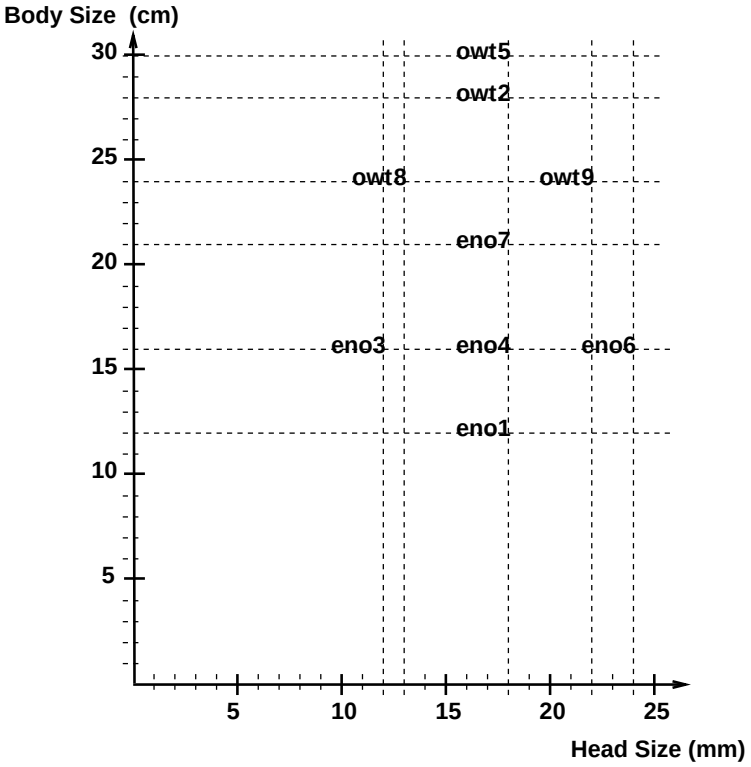
## Problem 3 Learning (20 Points)

As an avid reader of important Web sites (e.g, www.petersononline.com) you are no doubt aware that in July the world lost Roger Tory Peterson, a pioneer whose lavishly illustrated books introduced bird-watching to thousands of people worldwide.

Anxious to try out your newly acquired skills at machine learning and inspired by Peterson's example, you set out to automate the fine art of bird identification.

Wisely, you start off modestly, watching just two species of birds, two rare breeds called Eno and Owt. You manage to catch glimpses of each of these birds on a number of occasions, and have time only to get estimates of their overall size and their head size. Because they are related species (both in the Regetni family) they are not easily distinguished. You decide to use some machine learning techniques to allow you to predict the identity of the next bird you see that seems to be in that family.

| Sample | head size(mm) | body size (cm) | identity |
|---|---|---|---|
| 1 | 18 | 12 | Eno |
| 2 | 18 | 28 | Owt |
| 3 | 12 | 16 | Eno |
| 4 | 18 | 16 | Eno |
| 5 | 18 | 30 | Owt |
| 6 | 24 | 16 | Eno |
| 7 | 18 | 21 | Eno |
| 8 | 13 | 24 | Owt |
| 9 | 22 | 24 | Owt |




## Part A (2 Points)

Using nearest neighbor, and a distance metric where 1mm in head size is equivalent to 1cm in body size, what would you predict about a bird in the Regetni family that had a head size of 18 and a body size of 24?

## Part B (6 Points)

The idea of k-nearest neighbors suggests that you use not just the nearest neighbor, but some collection of nearest neighbors to identify the unknown, by taking a "vote" of the k-nearest neighbors.

1. What is the prediction if you use 3-nearest neighbors as the prediction technique (and the same distance metric as in Part A)?

2. Why should k be an odd number?

## Part C (8 Points)

You know that identification trees can sometimes be successful in reducing the complexity of prediction problems.

1. Generate an identification tree for the data in the table above. Show the tree here:

2. What prediction does it make about the identity of the unknown bird?

## Part D (4 Points)

You have now made three predictions about the bird identity, using nearest neighbor, 3-nearest neighbors, and an identification tree. Based on the training samples you have, describe how you would estimate which classification method gives the best predictive accuracy on new data points.

Recall that the *predictive accuracy* is a measure of how well a method classifies *new* data, that is, data not in the training set.

Answer:

## Problem 4 Identification Tree (30 Points)

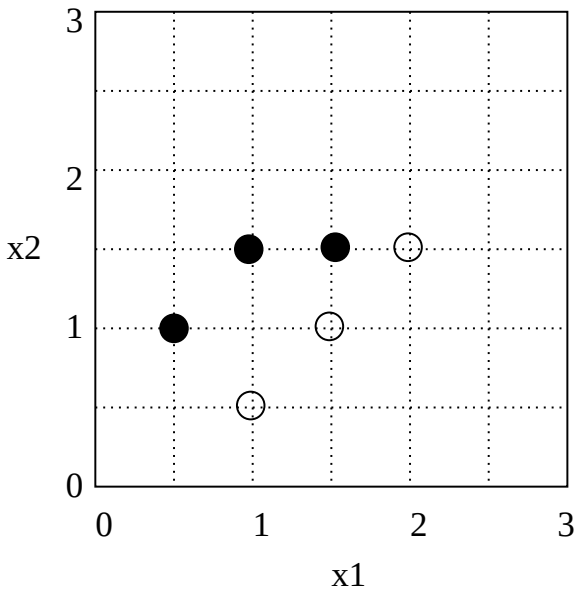
Having mastered the material in recent lectures in 6.034, you are determined to enter the fast-paced world-wide-web business. Knowing a good opportunity when you see one, you have started your own one-person company and now analyze data records for a living. Since your first client is interested not only in the predictive accuracy of your classification method, but also in an intelligible predictive model, you decide to use identification trees to analyze the client's data.

Let us assume the record attributes are arranged in a preassigned order: $x_1, x_2, \ldots, x_n$. Each training case is described by a vector of attribute values. Such a vector corresponds to a point in the $n$-dimensional space, the *measurement space*, defined by the record attributes. A classification task is then the partition of the measurement space into disjoint regions $R_i$ each of which is assigned a class label.

A new measurement $\mathbf{x}$ is classified by assigning it the class label associated with the region into which $\mathbf{x}$ falls.

### Part A (6 Points)

Your first task is to explain to your client how the identification tree works. Assume you have a training set of 6 cases belonging to two classes described by two attributes $x_1$ and $x_2$. Draw in the diagram below the boundaries of the decision regions implied by the `make-tree` procedure in PS 8.




| p | p log p |
|---|---|
| $\frac{1}{6}$ | -0.431 |
| $\frac{1}{5}$ | -0.464 |
| $\frac{1}{4}$ | -0.5 |
| $\frac{1}{3}$ | -0.528 |
| $\frac{2}{5}$ | -0.529 |
| $\frac{1}{2}$ | -0.5 |
| $\frac{3}{5}$ | -0.442 |
| $\frac{2}{3}$ | -0.39 |
| $\frac{3}{4}$ | -0.311 |
| $\frac{4}{5}$ | -0.258 |
| $\frac{5}{6}$ | -0.219 |

Figure 1: Training cases for Part A. You may find the table on right useful in estimating the decision boundaries.

## Part B (4 Points)

Your client likes your explanation and asks you to try your algorithm on a larger data set.

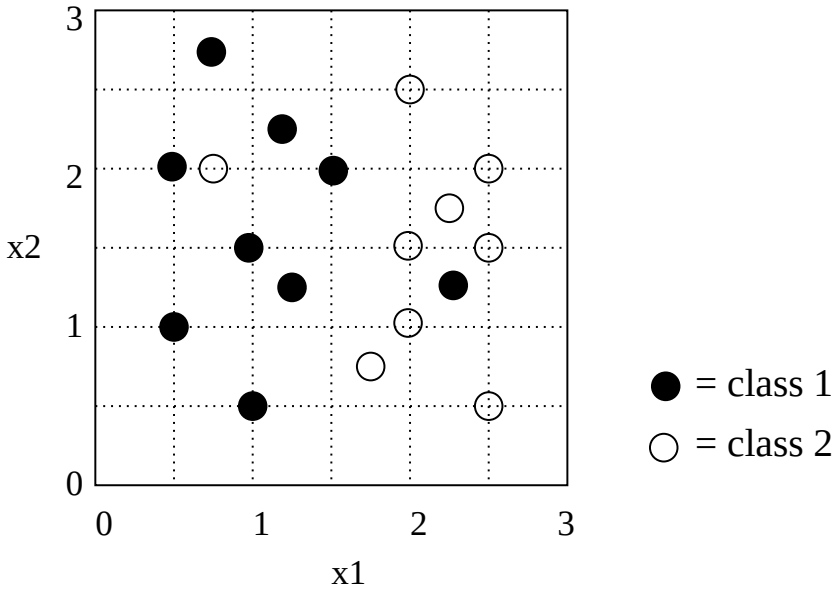



Figure 2: Training cases for Part B

You pull out your laptop and are ready to dazzle your client. To your dismay, your dog (the same one in PS 8) has completely ruined the floppy containing the identification tree program. Draw clearly in the diagram above the decision boundary corresponding to the top level split of your identification tree. Sketch the remaining decision boundaries corresponding to the lower level splits.

## Part C (4 Points)

Your client is not quite pleased with the classification results even though `make-tree` achieves a 100% accuracy in classifying the training cases. Why?

Answer:

## Part D (3 Points)

Not to let your client down, you said there is an option setting for the `make-tree` that controls the aggressiveness in splitting nodes in the identification tree. What you have in mind is that a node in the identification tree is split only if the gain in the reduction of average disorder exceeds certain predetermined (positive) threshold, i.e.,

```
disorder (node) - average-disorder (split-node) > threshold
```

Describe **concisely** the benefits of using the aggressiveness option for the data set in Part B.

Answer:

## Part E (4 Points)

Encouraged by your results but still not entirely convinced, your client gave you a second data set to try.

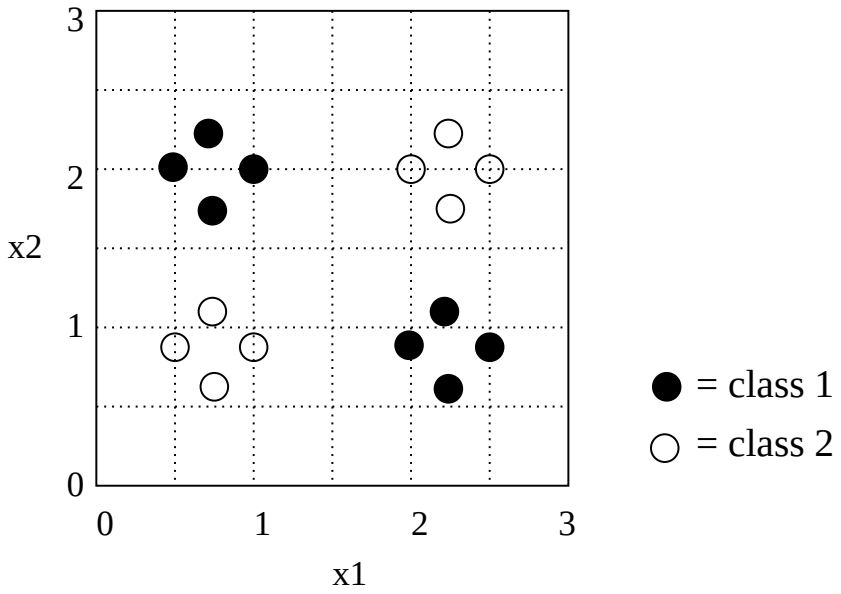



Figure 3: Training cases for Part E

Would you use the aggressiveness option for this data set? Why or why not?

Answer:

## Part F (5 Points)

At last your client is pleased with your presentation. Your first contract is to classify data sets such as the following.

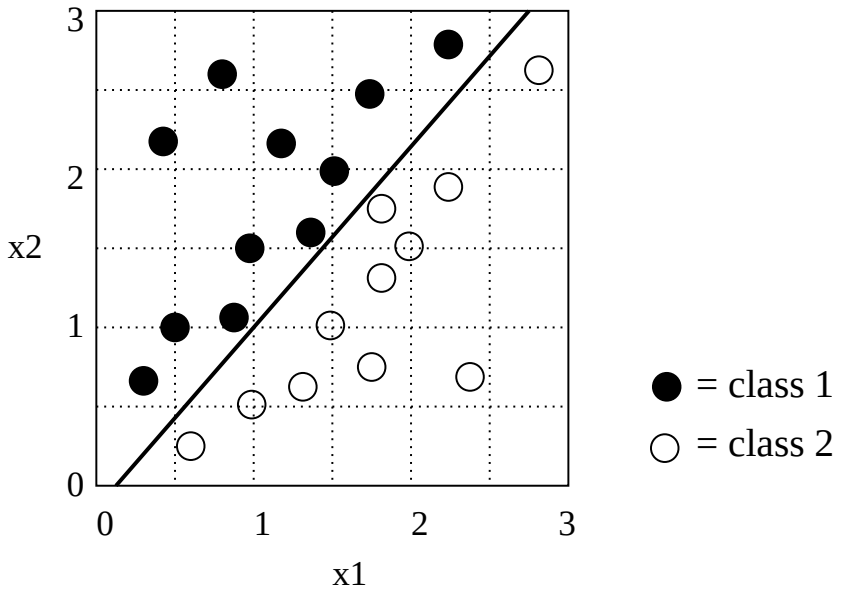



Figure 4: Training cases for Part F

Recall that the *predictive accuracy* is a measure of how well a tree classifies *new* data, that is, data not in the training set.

Suppose the true boundary separating the classes is the solid diagonal line. How would the predictive accuracy of an identification tree vary as a function of the number of training cases N used to build the tree? Assume that the training cases are all properly classified. Circle the correct answer.

```
A. roughly constant accuracy
B. decrease accuracy with N initially and then increase afterwards
C. increase accuracy with N initially and then decrease afterwards
D. decrease accuracy with N initially and then roughly constant afterwards
E. increase accuracy with N initially and then roughly constant afterwards
F. None of the above.
```

Explain **concisely** your answer:

## Part G (4 Points)

Most of the client's data sets are more complicated than that in Part F: They have noisy training samples, they have multiple decision regions, and the decision boundaries can be nonlinear.

Do you expect your answer in Part F to the predictive accuracy of the identification tree as a function of the number of training cases to hold true even for these more complicated data sets? Why or why not?

Answer: