CS 188 Fall 2015 Section 10: Decision Trees

You are a geek who hates sports. Trying to look cool at a party, you join a discussion that you belive to be about football and basketball. You gather information about the two main subjects of discussion, but still cannot figure out what sports they play.

Sport	Position	Name	Height	Weight	Age	College
?	Guard	Charlie Ward	6'02"	185	41	Florida State
?	Defensive End	Julius Peppers	6'07"	283	32	North Carolina

Fortunately, you have brought your CS 188 notes along, and will build some classifiers to determine which sport is being discussed.

You come across a pamphlet from the Atlantic Coast Conference Basketball Hall of Fame, as well as an Oakland Raiders team roster, and create the following table:

Sport	Position	Name	Height	Weight	Age	College
Basketball	Guard	Michael Jordan	6'06"	195	49	North Carolina
Basketball	Guard	Vince Carter	6'06"	215	35	North Carolina
Basketball	Guard	Muggsy Bogues	5'03"	135	47	Wake Forest
Basketball	Center	Tim Duncan	6'11"	260	35	Oklahoma
Football	Center	Vince Carter	6'02"	295	29	Oklahoma
Football	Kicker	Tim Duncan	6'00"	215	33	Oklahoma
Football	Kicker	Sebastian Janikowski	6'02"	250	33	Florida State
Football	Guard	Langston Walker	6'08"	345	33	California

1 Entropy

Before we get started, let's review the concept of entropy.

1. Give the definition of entropy for an arbitrary probability distribution $P(X)$.

2. Draw a graph of entropy $H(X)$ vs. $P(X = 1)$ for a binary random variable X.

3. What is the entropy of the distribution of *Sport* in the training data? What about *Position*?

2 Decision Trees

Central to decision trees is the concept of "splitting" on a variable.

1. To review the concept of "information gain", calculate it for a split on the *Sport* variable.

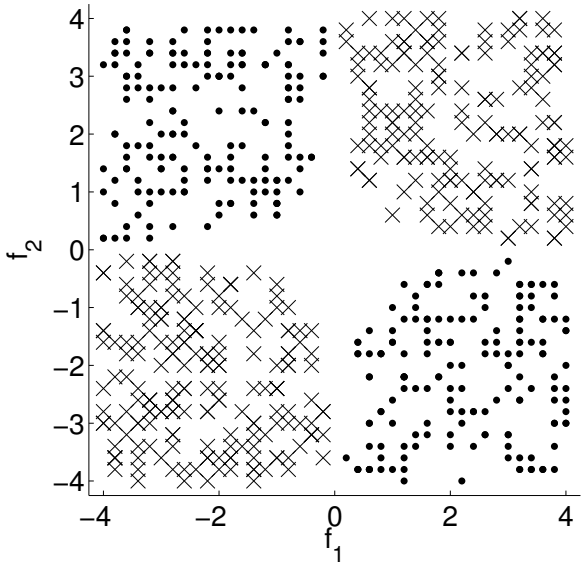
2. Of course, in our situation this would not make sense, as *Sport* is the very variable we lack at test time. Now calculate the information gain for the decision "stumps" (one-split trees) created by first splitting on *Position*, *Name*, and *College*. Do any of these perfectly classify the training data? Does it make sense to use *Name* as a variable? Why or why not?

3. Decision trees can represent any function of discrete attribute variables. How can we *best* cast continuous variables (*Height*, *Weight*, and *Age*) into discrete variables?

4. Draw a few decision trees that each correctly classify the training data, and show how their predictions vary on the test set. What algorithm are you following?

5. You may have noticed that the testing data has a value for *Position* that is missing in training data. What could we do in this case?

3 Pruning Decision Trees



1. Following your tree construction algorithm from the previous question, construct a decision tree to separate the two classes. Is this possible?

2. If you construct a decision tree all the way, such that each leaf node has a single class, are the two classes separated?