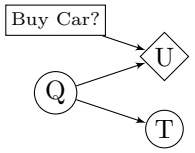
CS 188
Spring 2026

Regular Discussion 9

1 Decision Networks and VPI

A used car buyer can decide to carry out various tests with various costs (e.g., kick the tires, take the car to a qualified mechanic) and then, depending on the outcome of the tests, decide which car to buy. We will assume that the buyer is deciding whether to buy car c and that there is time to carry out at most one test which costs \$50 and which can help to figure out the quality of the car. A car can be in good shape (of good quality $Q = +q$) or in bad shape (of bad quality $Q = \neg q$), and the test might help to indicate what shape the car is in. There are only two outcomes for the test T: pass ($T = \text{pass}$) or fail ($T = \text{fail}$). Car c costs \$1,500, and its market value is \$2,000 if it is in good shape; if not, \$700 in repairs will be needed to make it in good shape. The buyers estimate is that c has 70% chance of being in good shape. The Decision Network is shown below.



(a) Calculate the expected net gain from buying car c, given no test.

(b) Tests can be described by the probability that the car will pass or fail the test given that the car is in good or bad shape. We have the following information:

$$P(T = \text{pass} \mid Q = +q) = 0.9$$
$$P(T = \text{pass} \mid Q = \neg q) = 0.2$$

Calculate the probability that the car will pass (or fail) its test, and then the probability that it is in good (or bad) shape given each possible test outcome.

(c) Calculate the optimal decisions given either a pass or a fail, and their expected utilities.

(d) Calculate the value of (perfect) information of the test. Should the buyer pay for a test?

2 Decision Trees

You are a geek who hates sports. Trying to look cool at a party, you join a discussion that you believe 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

2.1 Entropy

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

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

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

(g) What is the entropy of the distribution of Sport in the training data? What about Position?

2.2 Decision Trees

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

(h) To review the concept of "information gain", calculate it for a split on the Sport variable.

(i) 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?

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

(k) 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?

(l) 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?