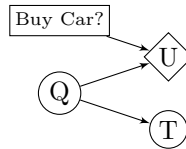


# 1 Decision Networks and VPI

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  $Q = -q$ ), 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.

$$\begin{aligned}
 EU(\text{buy}) &= P(Q = +q) \cdot U(+q, \text{buy}) + P(Q = -q) \cdot U(-q, \text{buy}) \\
 &= 0.7 \cdot 500 + 0.3 \cdot (-200) = 290
 \end{aligned}$$

- (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:

$$\begin{aligned}
 P(T = \text{pass} \mid Q = +q) &= 0.9 \\
 P(T = \text{pass} \mid Q = -q) &= 0.2
 \end{aligned}$$

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.

$$\begin{aligned}
 P(T = \text{pass}) &= \sum_q P(T = \text{pass}, Q = q) \\
 &= P(T = \text{pass} \mid Q = +q)P(Q = +q) + P(T = \text{pass} \mid Q = -q)P(Q = -q) \\
 &= 0.9(0.7) + 0.2(0.3) = 0.69 \\
 P(T = \text{fail}) &= 0.31
 \end{aligned}$$

$$\begin{aligned}
 P(Q = +q \mid T = \text{pass}) &= \frac{P(T=\text{pass} \mid Q=+q)P(Q=+q)}{P(T=\text{pass})} = \frac{0.9 \cdot 0.7}{0.69} = \frac{21}{23} \approx 0.91 \\
 P(Q = +q \mid T = \text{fail}) &= \frac{P(T=\text{fail} \mid Q=+q)P(Q=+q)}{P(T=\text{fail})} = \frac{0.1 \cdot 0.7}{0.31} = \frac{7}{31} \approx 0.22
 \end{aligned}$$

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

$$\begin{aligned}
 EU(\text{buy} \mid T = \text{pass}) &= P(Q = +q \mid T = \text{pass})U(+q, \text{buy}) + P(Q = -q \mid T = \text{pass})U(-q, \text{buy}) \\
 &\approx 0.91 \cdot 500 + 0.09 \cdot (-200) \approx 437 \\
 EU(\text{buy} \mid T = \text{fail}) &= P(Q = +q \mid T = \text{fail})U(+q, \text{buy}) + P(Q = -q \mid T = \text{fail})U(-q, \text{buy}) \\
 &\approx 0.22 \cdot 500 + 0.78 \cdot (-200) = -46 \\
 EU(\neg\text{buy} \mid T = \text{pass}) &= 0 \\
 EU(\neg\text{buy} \mid T = \text{fail}) &= 0 \\
 \text{Therefore: } MEU(T = \text{pass}) &= 437 \text{ (with buy) and } MEU(T = \text{fail}) = 0 \text{ (using } \neg\text{buy)}.
 \end{aligned}$$

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

$$\begin{aligned}
 VPI(T) &= (\sum_t P(T = t)MEU(T = t)) - MEU(\phi) \\
 &= 0.69 \cdot 437 + 0.31 \cdot 0 - 290 \approx 11.53 \\
 \text{You shouldn't pay for it, since the cost is } & \$50.
 \end{aligned}$$

## 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

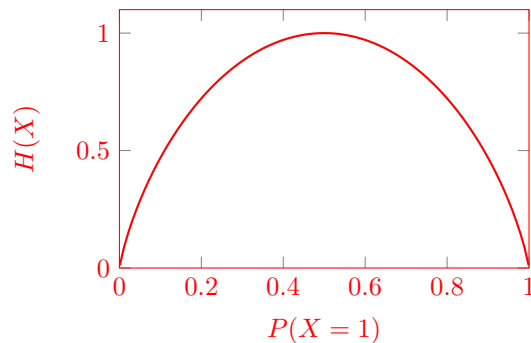
Before we get started, let's review the concept of entropy.

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

$$H(X) = \sum_x P(x) \log_2(1/P(x)).$$

You can see this as the expected information content of the distribution.

- (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?

To calculate the entropy for a random variable, we estimate the probability distribution and use the formula from the part above.

$$P(S = \text{football}) = 1/2, P(S = \text{basketball}) = 1/2$$

$$P(P = \text{guard}) = 1/2, P(P = \text{kicker}) = 1/4, P(P = \text{center}) = 1/4$$

$$H(S) = \frac{\log_2(2)}{2} + \frac{\log_2(2)}{2} = 1$$

$$H(P) = \frac{\log_2(2)}{2} + \frac{\log_2(4)}{4} + \frac{\log_2(4)}{4} = 3/2$$

## 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.

Since the variable that we want to predict is Sport, we want to be calculating the entropy with respect to the variable Sport.

(a) i. Distribution before: 8 examples with (1/2, 1/2). (here the first number in the tuple is P(basketball), and the second number is P(football)).

ii. Entropy before:  $(\frac{\log(2)}{2} + \frac{\log(2)}{2}) = 1$

(b) i. Distribution after: 4 examples with (1, 0), 4 examples with (0, 1)

ii. Entropy after:  $\frac{4}{8}(\frac{\log(1)}{1}) + \frac{4}{8}(\frac{\log(1)}{1}) = 0$

So, the information gain is  $(1 - 0) = 1$ , which is the greatest possible.

- (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?

Note that here we will be splitting on different variables but still need to look at the entropy of the distribution of the variable we need to predict which is sport. So, the before case remains same as before.

(a) **Position:** Distribution after: 4 examples with (3/4, 1/4), 2 examples with (1/2, 1/2), 2 examples with (0, 1).

Entropy after:  $\frac{4}{8}(\frac{\log(4/3)}{4/3} + \frac{\log(4)}{4}) + \frac{2}{8}(\frac{\log(2)}{2} + \frac{\log(2)}{2}) + \frac{2}{8}(\frac{\log(1)}{1}) = 0.66$

(b) **Name:** Distribution after: 1 example with (1, 0), 2 examples with (1/2, 1/2), 1 example with (0, 1), 2 examples with (1/2, 1/2), 1 example with (0,1), 1 example with (0,1).

Entropy after: 0.5

(c) **College:** Distribution after: 2 examples with (1, 0), 1 example with (1, 0), 3 examples with (1/3, 2/3), 1 example with (0, 1), 1 example with (0,1).

Entropy after: 0.34

Note that none of these variables completely classifies the data.

Regarding using the Name as a feature to use in classifying data: since we expect people’s names to be unique, using them as a feature in learning is akin to using the unique ID of each data point. That is to say, it’s quite a bad idea you will overfit to the training data.

- (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?

Use an inequality relation, Attribute  $> a$ , where  $a$  is a split point chosen to give the highest information gain. E.g., an initial split on Age  $> 34$  will perfectly classify the training data.

- (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?

We use the algorithm as given in the slides, and for each split use the variable that gives us the maximum information gain. In this given problem, as we observed above, the variable Age correctly classifies all of the training data, so that is the first variable that gets picked up, and the algorithm stops at that.

This decision tree would predict test example 1 (Age 41) to be Basketball and test example 2 (Age 32) to be Football.

- (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?

When we come to a split on a variable whose value for the test subject is missing in the tree, we could just choose the most likely branch of the split (the branch that leads to the node with the greatest number of items).