From January 2020

1) A cell phone repair shop considers the following binary features when making a decision about the action to take on a phone: whether the phone boots correctly (f_1), and whether the screen is cracked (f_2). Each feature may have value true (T) or false (F). The actions that the engineer can take are: replace the screen (R), update the operating system (U), or do nothing (N). The engineer only makes one decision at a time. The decision is based on the following dataset: D = {< F, F, U >, <T, F, N >, <F, T, R >, < T, T, R >}, where the first two elements of each vector are the feature values of f_1 and f_2 respectively, and the last one is the class. Build a decision tree for this problem with the algorithm ID3.

From January 2019

2) Consider the data set S = {<T, T, 0>, <F, T, 0>, <T, F, 1>, <T, F, 0>, <F,F,1>, <F,F,0>}, where each data point has two binary features A and B (whose values are either true (T) or false (F)), and the third value is the class. What is the first feature that the CART algorithm would split on? Justify your answer.

3) Given the following dataset: {<-1,2>, <0,1>, <1,2>, <2,5>} , where the first component of each vector is the input, and the second component the corresponding desired output, compute the least-squares solution to the linear regression problem for the function: $y=w_0+w_1x$

From question sheet

4) Given the dataset: <-1, 1.6>, <0,0.95>, <1,1.2>, <2,1.9>, find the least-squares solution for the function:

$$y(x,\mathbf{w})=w_0+w_1\frac{1}{1+e^{-(x+1)}}$$

$D = \{\langle F, F, U\rangle, \langle T, F, N\rangle, \langle F, T, R\rangle, \langle T, T, R\rangle\}$

First:

$$G(D) = -\tfrac{1}{4}\log\tfrac{1}{4} - \tfrac{1}{4}\log\tfrac{1}{4} - \tfrac{1}{2}\log\tfrac{1}{2} = 1.5$$

second:

$$G(D, f_1) = H(D) - \frac{\#(f_1 = F)}{4} H(D, f_1 = F)$$

$$- \frac{\#(f_1 = T)}{4} H(D, f_1 = T)$$

$$H(D, f_1 = F) = -\tfrac{1}{2}\log(\tfrac{1}{2}) - \tfrac{1}{2}\log(\tfrac{1}{2}) - 0\log 0 = 1$$

$$H(D, f_1 = T) = -0\log 0 - \tfrac{1}{2}\log\tfrac{1}{2} - \tfrac{1}{2}\log\tfrac{1}{2} = 1$$

$$G(D, f_1) = 1.5 - \tfrac{1}{2}\times 1 - \tfrac{1}{2}\times 1 = 0.5$$

$$G(D, f_2) = H(D) - \frac{\#(f_2 = F)}{4} H(D, f_2 = F)$$

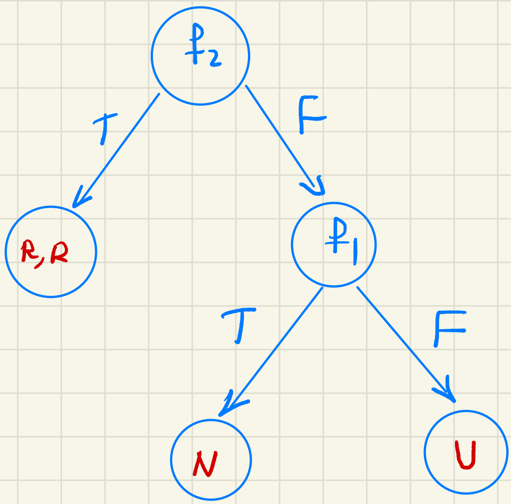
$$- \frac{\#(f_2 = T)}{4} H(D, f_2 = T)$$

$$H(D, f_2 = F) = -\tfrac{1}{2}\log\tfrac{1}{2} - \tfrac{1}{2}\log\tfrac{1}{2} - 0\log 0 = 1$$

$$H(D, f_2 = T) = -0\log 0 - 0\log 0 - 1\log 1 = 0$$

$$G(D, f_2) = 1.5 - \tfrac{2}{4}\times 1 - \tfrac{2}{4}\times 0 = 1$$

We see: $G(D, f_2) > G(D, f_1)$



---

a. Consider the data set S = {<T, T, 0>, <F, T, 0>, <T, F, 1>, <T, F, 0>, <F,F,1>, <F,F,0>}, where each data point has two binary features A and B (whose values are either true (T) or false (F)), and the third value is the class. What is the first feature that the CART algorithm would split on? Justify your answer.

[4 marks: 1 use of Gini Impurity, 1 impurity of the initial set, 1(x2) impurity of subsets per feature]

Answer:

With CART we use the Gini Impurity of a set G(S). The Gini split for A is:
G(S, A) = 0.44 -3/6 * 0.44 - 3/6 0.44 = 0

The Gini split for B is:

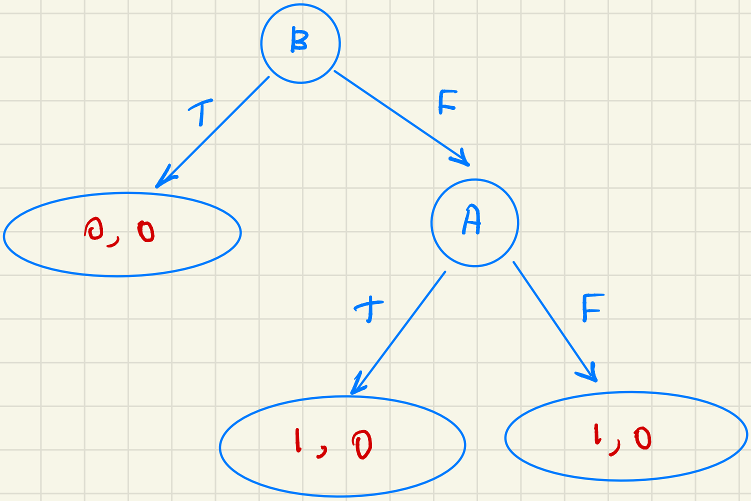
G(S, B) = 0.44 -2/6 * 0 - 4/6* 0.5 = 0.11
CART would split on B.

For example:

$$G(S) = \sum_{i=1}^{C} p_i(1-p_i) = 4/6\,(1-4/6) + 2/6\,(1-2/6) = 0.44$$

$$G(S, A=T) = 2/3\,(1-2/3) + 1/3\,(1-1/3) = 0.44$$

$$G(S, A=F) = 2/3\,(1-2/3) + 1/3\,(1-1/3) = 0.44$$



b. Given the following dataset: {<-1,2>, <0,1>, <1,2>, <2,5>}, where the first component of each vector is the input, and the second component the corresponding desired output, compute the least-squares solution to the linear regression problem for the function: $y=w_0+w_1x$ **[8 marks: 2 definition of the feature matrix, 4 use and calculation of the pseudoinverse, 2 vector w from the pseudoinverse and input]**

Answer:

We begin by defining the matrix of the features Φ:

$$\Phi=\begin{bmatrix}1 & -1\\ 1 & 0\\ 1 & 1\\ 1 & 2\end{bmatrix}$$

Then we compute the pseudoinverse of Φ:

CONTINUES ON FOLLOWING PAGE

$$\Phi^T\Phi=\begin{bmatrix}4 & 2\\ 2 & 6\end{bmatrix}$$

$$(\Phi^T\Phi)^{-1}=\begin{bmatrix}0.3 & -0.1\\ -0.1 & 0.2\end{bmatrix} \quad \Phi_p=\begin{bmatrix}0.4 & 0.3 & 0.2 & 0.1\\ -0.3 & -0.1 & 0.1 & 0.3\end{bmatrix}$$

Lastly, the least-squares solution is given by:

$$w=\Phi_p t=\begin{bmatrix}0.4 & 0.3 & 0.2 & 0.1\\ -0.3 & -0.1 & 0.1 & 0.3\end{bmatrix}\begin{bmatrix}2\\ 1\\ 2\\ 5\end{bmatrix}=\begin{bmatrix}2\\ 1\end{bmatrix}$$

The least-squares solution to the regression problem is therefore:

$y=2+x$

11. Given the dataset: <-1, 1.6>, <0,0.95>, <1,1.2>, <2,1.9>, find the least-squares solution for the function: $y(x,\mathbf{w})=w_0+w_1\frac{1}{1+e^{-(x+1)}}$

$$\mathbf{\Phi}=\begin{bmatrix}1 & 0.5\\ 1 & 0.73\\ 1 & 0.88\\ 1 & 0.95\end{bmatrix}$$

$$\mathbf{w}=\mathbf{\Phi}_p\mathbf{t}=\begin{bmatrix}1.96 & 0.48 & -0.49 & -0.94\\ -2.23 & -0.29 & 0.97 & 1.56\end{bmatrix}\begin{bmatrix}1.6\\ 0.95\\ 1.2\\ 1.9\end{bmatrix}=\begin{bmatrix}1.22\\ 0.28\end{bmatrix}$$