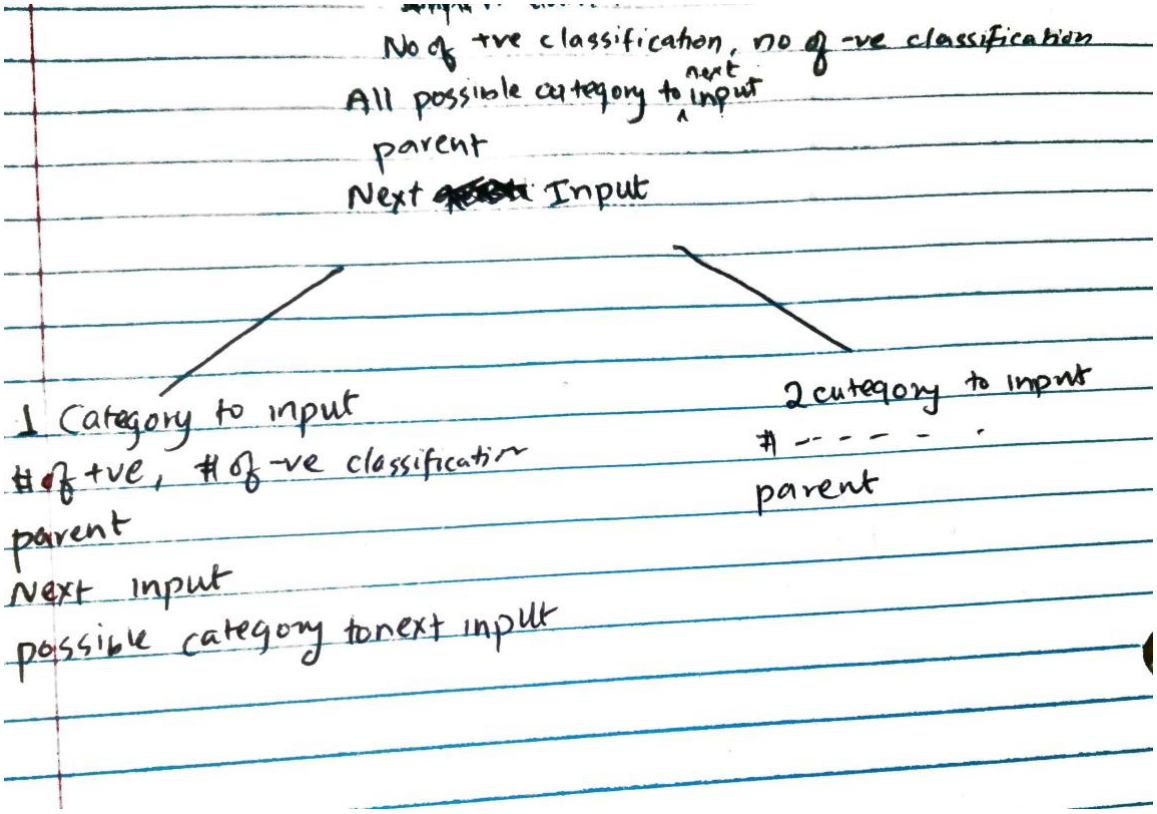
Muskan Uprety
Lab C Design Plan
March 27, 2020

We are trying to implement a decision tree classifier that gives us a binary classification answer based on multiple input variables. We are going to have a tree structure where nodes are expanded, and the child nodes are all the possible categories to the input feature. A diagram describing the approach is given below:



As we can see, the root node has information about how many positive or 'yes' classification we have and how many negative or 'no' classification we have. Nodes also store info about their parent which would be none for the root. The crucial information of the node is the 'next input'

which is the input variable that is going to be expanded to try and categorize the overall problem. We choose the input variable to expand by minimizing the entropy function on all available input variables. Entropy functions describes the level of uncertainty we have on the classification problem by looking at a particular input variable. Since we are looking at binary classification, entropy will be between zero and one, zero meaning no uncertainty and one meaning input variable contributes nothing to reduce uncertainty. We want to use the input variable whose entropy is closest to zero. The number of children for this node would be the number of all possible categories in this input variable, which is also stored inside the node. After storing all this information, we expand the node where children of the nodes, like mentioned previously, are all possible options for the input variable that we are expanding.

All these nodes also store all the information mentioned above for the root node. All individual child node calculates the entropy value for the remaining input variable based on the instances it holds (this is the no. of positive and negative classifications). We continue expanding nodes in this manner until we have reached a node where there are only positive or only negative classifications. That would mean the entropy value for the input variable is zero. Then we stop expanding that node. We also stop expanding a node if we run out of input variables. If we run out of input variables but still are not at a point of full certainty, we use probability to predict the classification of an instance. In the diagram above, the right node has only one type of classification (let's say all positive classification) so it will not be expanded further. We don't need a next input to expand or its possible categories.