Decision Trees - extra credit
Return by May 4 to get credit

We want to construct a decision tree using the information gain to decide which features to select. We are given the following data set (from Yoh-Han Pao, Adaptive Pattern Recognition and Neural Networks, 1980). The data are represented by attribute–value pairs and there are two classes. We want to construct a decision tree that will enable us to later classify new unlabeled data.

height	hair	eyes	class
tall	dark	blue	yes
short	dark	blue	yes
tall	blond	blue	no
tall	red	blue	no
tall	blond	brown	yes
short	blond	blue	no
short	blond	brown	yes
tall	dark	brown	yes

1. Compute the entropy of the set S, which is defined as follows:
   $Entropy(S) \equiv -p \log_2 p - n \log_2 n$
   where p is the fraction of positive examples and n of negative examples in the set S.
   $Entropy(S) = \ldots$

2. for each attribute A (in the example *height*, *hair*, and *eyes*)

   (a) Split the examples into d disjoint subsets, E_k for $k = 1, \ldots d$ according to the value v of the attribute A. Each of those subsets corresponds to a branch in the decision tree from node A. For each subset compute its entropy, where p_k and n_k are respectively the fraction of positive and negative examples in the subset E_k:
   $Entropy(S, A = v) \equiv -p_k \log_2 p_k - n_k \log_2 n_k$
   In our example, if we start with *height* we have two values, *tall* and *short*. For the value *tall* there are 3 positive examples and 3 negative:

$$\begin{aligned} Entropy(S, height = tall) &= -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} \\ &= 0.971 \end{aligned}$$

   Compute in the same way the entropy for the remaining value of *height*:
   $Entropy(S, height = short) = \ldots$

(b) After you have computed the entropy of each subset for each value of attribute A compute the $Gain$, which is the expected reduction in entropy due to sorting on attribute A. This is computed as:

$$Gain(S, A) \equiv Entropy(S) \; - \sum_{v \in Values(A)} \frac{|S_v|}{|S|} Entropy(S_v)$$

where S_v is the subset of S for which A has value v, $Values(A)$ is the set of all possible values for A, $|S_v|$ is the number of elements in S with value v.

In our example, since A is $height$ then $Values(A) = \{tall, short\}$, $|S_{tall}| = 5$ and $|S_{short}| = 3$

$Gain(A, height) = \ldots$

Repeat the process for the other attributes, *hair* and *eyes*:

$Entropy(S, hair = dark) = \ldots$

$Entropy(S, hair = blond) = \ldots$

$Entropy(S, hair = red) = \ldots$

$Gain(A, hair) = \ldots$

and then

$Entropy(S, eyes = blue) = \ldots$

$Entropy(S, eyes = brown) = \ldots$

$Gain(A, eyes) = \ldots$

3. Select the attribute that has the largest value for $Gain$. This becomes the node at the top level of the decision tree. If any of the subsets obtained by sorting on the first attribute is mixed, then repeat the process recursively, starting from the mixed subset(s). Notice that each subset is a subset of the initial one, so the size of the subsets decreases at each level. Once an attribute is used on a path, it does not get used again, so the number of attributes to examine decreases at each level.

   Continue until all subsets have 0 entropy (i.e. none of the classes is mixed) or until you run out of attributes.