**Midterm 2 Review**

1. *Bootstrap.* With $n = 2$ training examples, how many unique datasets can I generate with bagging? What about $n = 3$?

2. *Central Limit Theorem.* Going back to our class year example, say we expect the following probabilities of each class year: [0.125, 0.125, 0.25, 0.5] for [first-year, sophomore, junior, senior]. Let $Y$ denote this random variable for year.

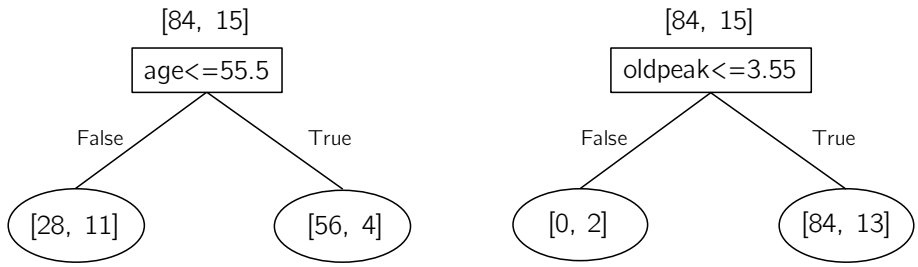
   (a) If the class years are represented as the values [0,1,2,3] (respectively), what is the mean (expected value) $E[Y]$ of this distribution?

   (b) Set up a computation for the variance of this distribution. The result of this computation (double check after class) is $\text{Var}(Y) = 1.109375$.

   (c) In reality we observe a class with $n = 40$ students and sample mean $\bar{Y}_n = 1.9$. We wish to test the hypothesis that there are more first-years and sophomores in the class than we expected. First, use the CLT to compute the associated Z-score.

   (d) The associated p-value is 0.08833 (double check after class). What do you conclude about your observed data?

3. *Entropy.* Consider the two feature choices below (for the heart disease dataset), and their associated splits. Counts of label -1 vs. 1 are shown in brackets.




(a) After splitting the data based on each feature, what is the *classification error* for each tree?

(b) Before considering the feature, what is $H(Y)$, the entropy of the initial partition? (don't need to find a value, just set up the equation)

(c) Which tree do you think produces more information gain?