Practice Midterm 2

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.

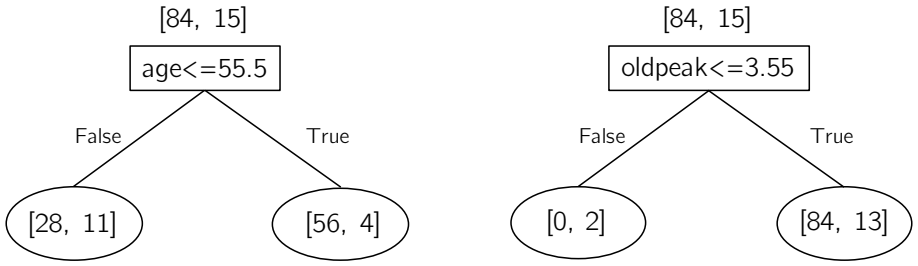
   (a) If the class years are represented as the values [0,1,2,3] (respectively), what is the mean (expected value) $E[Y]$ of this distribution?

   (b) Compute the variance of this distribution.

   (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.

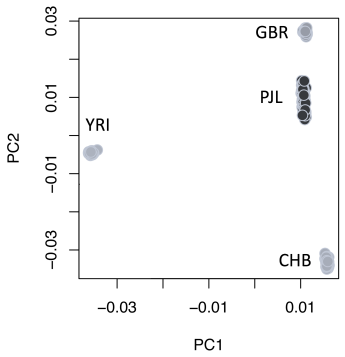


(a) After splitting the data based on each feature, what is the *classification error* for each tree (assuming that we are classifying based on the majority class)?

(b) Before considering the feature, what is $H(Y)$, the entropy of the initial partition?

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

4. Consider the following four populations from the 1000 Genomes Project: CHB ("Chinese from Beijing"), GBR ("Great Britain"), PJL ("Punjabi from Lahore"), YRI ("Yoruba from Nigeria"). The first two principal components for this data are shown below:



   (a) Which two populations are most closely related?

   (b) In the space above, draw a tree showing the relationship between these four populations that is consistent with the PCA plot.

5. Say I train a binary logistic regression model (i.e. outcomes $\in \{0, 1\}$) and end up with $\hat{w} = [\hat{w}_0, \hat{w}_1]^T = [1, 4]^T$. What is the decision boundary? Sketch a graph of this logistic model and label the decision boundary. How would you classify a new point $x_{\text{test}} = -0.3$?

6. Hypothetically, of the applicants for loans at a bank, 27.5% of the Black applicants got a loan compared to 35% for white applicants. Is there disparate impact in the bank's decisions? Explain your reasoning.

7. Say we have the following training data with $p = 2$ features. Feature f_1 can take on three values $\{1, 2, 3\}$ and f_2 can take on five values $\{A, B, C, D, E\}$. Using Naive Bayes, which class $y \in \{0, 1\}$ would you predict for the test example $\vec{x} = [1, D]$? Show all work.

$\boldsymbol{x}$	f_1	f_2	y
$\boldsymbol{x}_1$	3	A	0
$\boldsymbol{x}_2$	2	B	1
$\boldsymbol{x}_3$	1	C	0
$\boldsymbol{x}_4$	2	E	0
$\boldsymbol{x}_5$	1	A	1