Practice Midterm 2

1. *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.

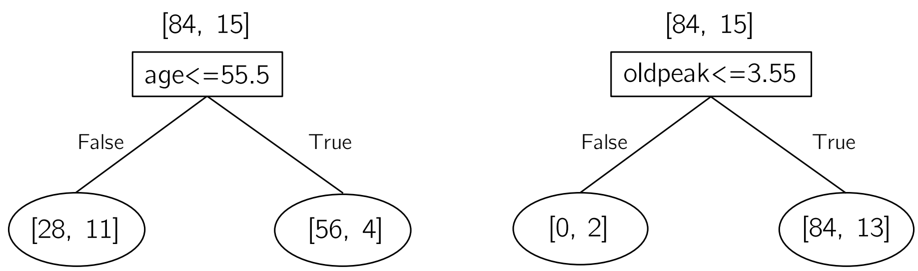
   (a) If the class years are represented as the values [0,1,2,3] (respectively), what is the mean (expected value) $E[Y]$ of this distribution?

   (b) Compute the variance $\text{Var}(Y)$ 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). Sketch out the position of the test statistic on a standard normal distribution. Shade the area(s) representing the p-value. What do you conclude about your observed data?

2. *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?

3. *Logistic regression.* Say I train a logistic regression model and obtain the following weights $\vec{w} = [1, 4]^T$.

   (a) Compute the decision boundary. Your answer should be an inequality describing when $\hat{y} = 1$ (i.e. predict 1).

   (b) Sketch the model, labeling the decision boundary and graph axes.

   (c) How would you classify a new point $x_{\text{test}} = -0.3$?

   (d) If the weight vector had instead been $\vec{w} = [2.5, 10]^T$, would the decision boundary change? Would the prediction change?

4. We are performing SGD to train a logistic regression model. We start with $\vec{w} = [w_0, w_1]^T = [0, 0]^T$ and $\alpha = 0.01$. What are the new weights after analyzing data point $(x, y) = (-3, 1)$?

5. *Bayesian probability.* For a specific disease, the incidence in the general population is $\frac{1}{500}$. Say I have a clinical test for this disease that comes back either positive or negative. Given a positive test result, there is an 80% chance the person has the disease. What is the *accuracy* of the test? In other words, compute the probability of a positive test result, given that the person has the disease. You may assume this value equals the probability of a negative test result, given the person is healthy.

6. *Naive Bayes.*

   (a) Write out a generic Bayesian model for $p(y|\vec{x})$, labeling the likelihood, prior, evidence, and posterior.

   (b) Now say that $\vec{x} = [x_1, x_2, x_3]^T$ (i.e. 3 features). Rewrite the likelihood, applying the Naive Bayes assumption. Challenge question: redo the steps of the derivation in this small case.

   (c) Explain in words how we compute $p(x_2 = v|y = k)$ (i.e. given y is some class label k, what is the probability feature x_2 takes on the value v).

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

8. What is the runtime of Naive Bayes, assuming each feature has a small, constant number of possible values and the outcome has K possibilities? Also assume that we use **dictionaries** throughout the implementation. Answer for the two steps below in terms of n, p, and K:

   (a) Estimating the prior probabilities (i.e. computing θ_k for $k = 1, \cdots, K$).

   (b) Estimating the likelihood probabilities (i.e. computing $\theta_{k,j,v}$ for all k, j, v).

9. *Disparate impact.* 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.