

Midterm 2 Practice Exam

1. *PCA: duplicated from last class.* In terms of n , p , and r , what is the runtime of each step of PCA? (you can skip the step related to computing eigenvalues and eigenvectors)

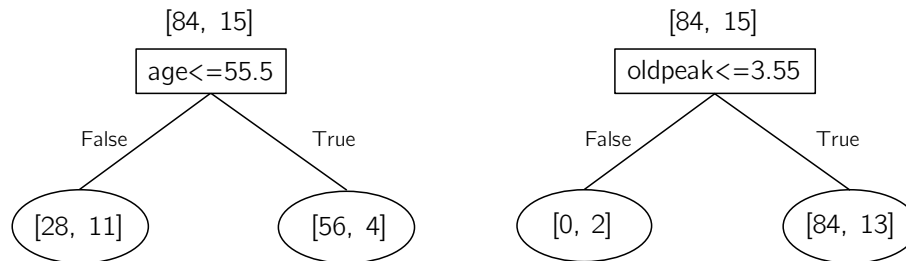
2. *Central Limit Theorem.* Going back to our class year example, say we expect the following probabilities of each class year: $[0.125, 0.125, 0.25, 0.5]$ for [first-year, sophomore, junior, senior]. Let Y denote this random variable for year.
 - (a) If the class years are represented as the values $[0,1,2,3]$ (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?
4. *Logistic Regression*. Say I run logistic regression on a dataset with $p = 1$ and obtain the following weights: $\vec{w} = [3, -2]$.
- (a) Compute the decision boundary. Your answer should be an inequality describing when $\hat{y} = 1$ (i.e. predict 1).
- (b) Sketch the logistic function, labeling the decision boundary and the axes.
- (c) For a new point $x_{\text{test}} = 2$, what label would you predict?
- (d) If the weight vector had instead been $\vec{w} = [6, -4]$, would the decision boundary change? Would the prediction change?

5. 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 $(x, y) = (-3, 1)$?
6. *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.
7. *Naive Bayes.* Before moving to the “naive” part, write out a generic Bayesian model for $p(y|\vec{x})$, labeling the likelihood, prior, evidence, and posterior.

$$p(y|\vec{x}) =$$

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

$$p(x_1, x_2, x_3|y) =$$

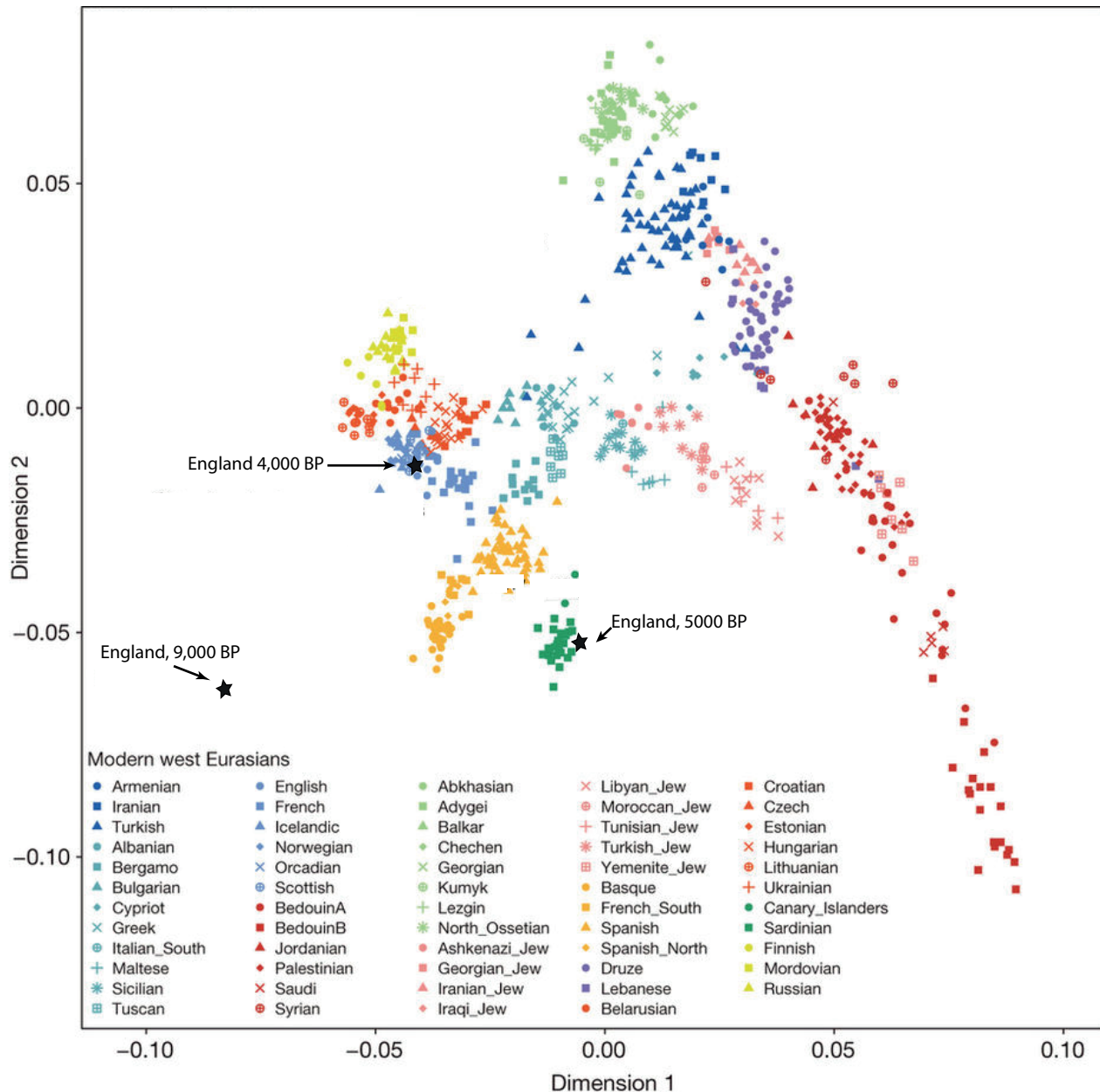
Finally, explain in words how we compute $p(x_2 = v|y = k)$ (i.e. given y is some class label k , what is the probability feature x_2 takes on the value v).

8. 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.

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

9. 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.

10. *PCA*. The figure below shows the first two PCs of genome-wide data from 777 present-day people from West Eurasia, along with three ancient British people who lived 9000, 5000 and 4000 years ago (labeled stars, “BP” means “[years] Before Present”).



- (a) What can you infer about the relationship between each of the ancient people and present-day Europeans?
- (b) What does this figure suggest about the history of Britain, and the people living there, over the past ten thousand years?

Acknowledgements: Iain Mathieson