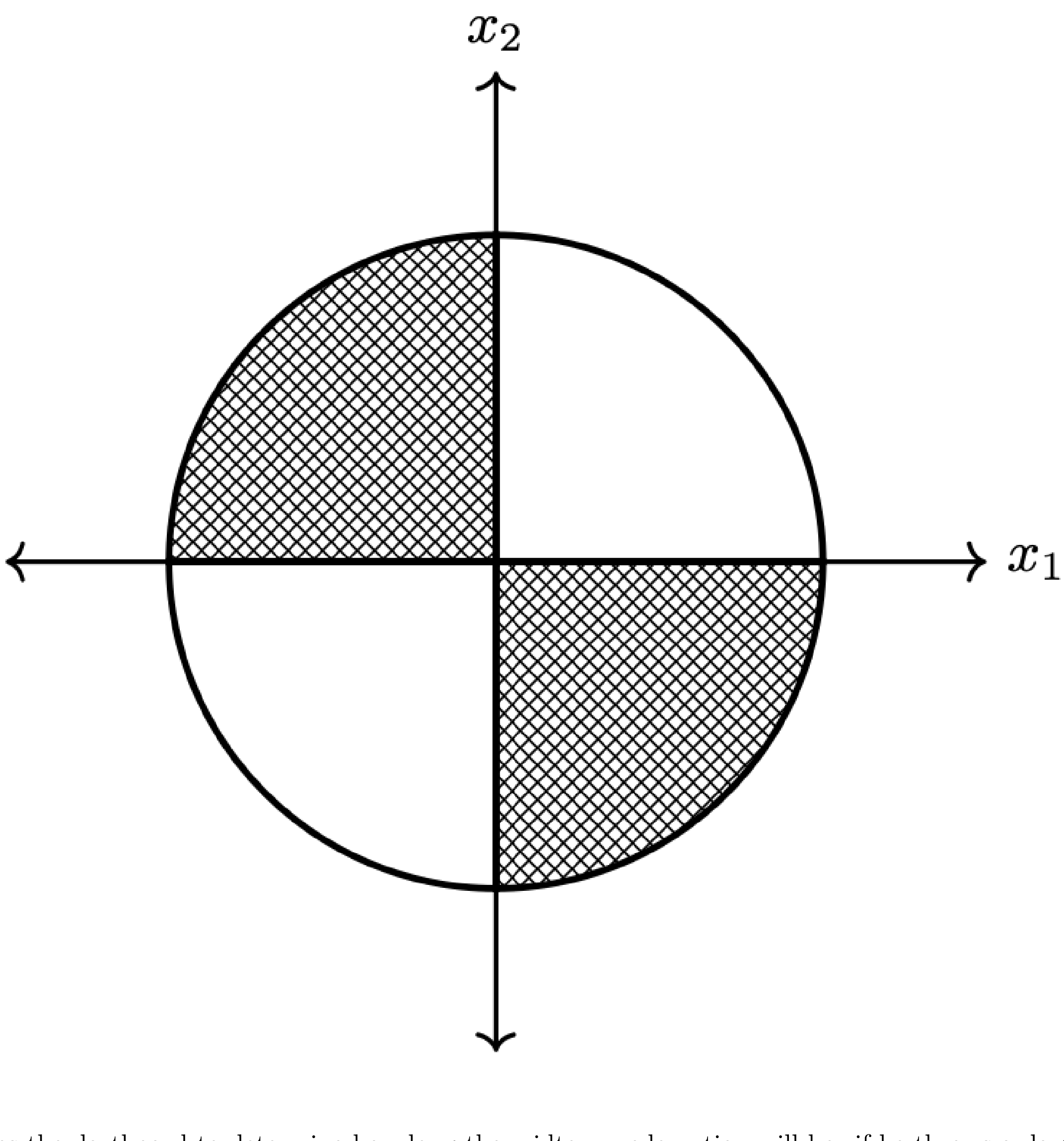
DSC 140A - Homework 09

Homework 09 Instructions. You may collaborate with whomever, including AI. Each student must submit their own homework even if they worked with others. We will not be grading for correctness, but we will be grading for completeness. Please turn out a single PDF file that contains your solutions to all of the problems. We recommend using LATEXwith our template. You can use markdown but make sure each problem is clearly labeled. For coding problems, we recommend using Jupyter notebooks, covert them into PDF, and then merge the PDF with your math problems solutions generated in LATEX.

Problem 1.

Suppose Jake has a dartboard at home that looks like the below:



Jake uses the dartboard to determine how long the midterm redemption will be: if he throws a dart and it lands in the shaded region, the midterm will have 17 questions; if it lands in the unshaded region, it will have 16 questions.

Assume that Jake's dart throws are drawn from a uniform distribution on the dartboard, and that the dart always hits the board (that is, the density function is constant everywhere on the dartboard, and zero off of the dartboard). Let X_1 be the horizontal component of a dart throw and X_2 be the vertical component. Let Q be the number of questions on the exam; since it is chosen randomly, it is also a random number.

a) True or False: X_1 and X_2 are independent.

Solution: False.

b) True or False: X_1 and X_2 are conditionally independent given Q.

Solution: False.

c) True or False: X_1 and Q are independent.

Solution: True.

d) True or False: X_1 and Q are conditionally independent given X_2.

Solution: False.

Solution:

Video explanation: https://youtu.be/b4XlZsePCgU

Problem 2.

Suppose the underlying class-conditional densities in a binary classification problem are known to be multivariate Gaussians.

Suppose a Quadratic Discriminant Analysis (QDA) classifier using full covariance matrices for each class is trained on a data set of n points sampled from these densities.

True or False: As the size of the data set grows (that is, as $n \to \infty$), the training error of the QDA classifier must approach zero.

Solution: False.

Video explanation: https://youtu.be/t40ex-JCYLY

Problem 3.

Suppose Quadratic Discriminant Analysis (QDA) is used to train a classifier on the following data set of (x_i, y_i) pairs, where x_i is the feature and y_i is the class label:

x_i	y_i
0	0
2	0
4	0
1	1
3	1
5	1

Univariate Gaussians are used to model the class conditional densities, each with their own mean and variance.

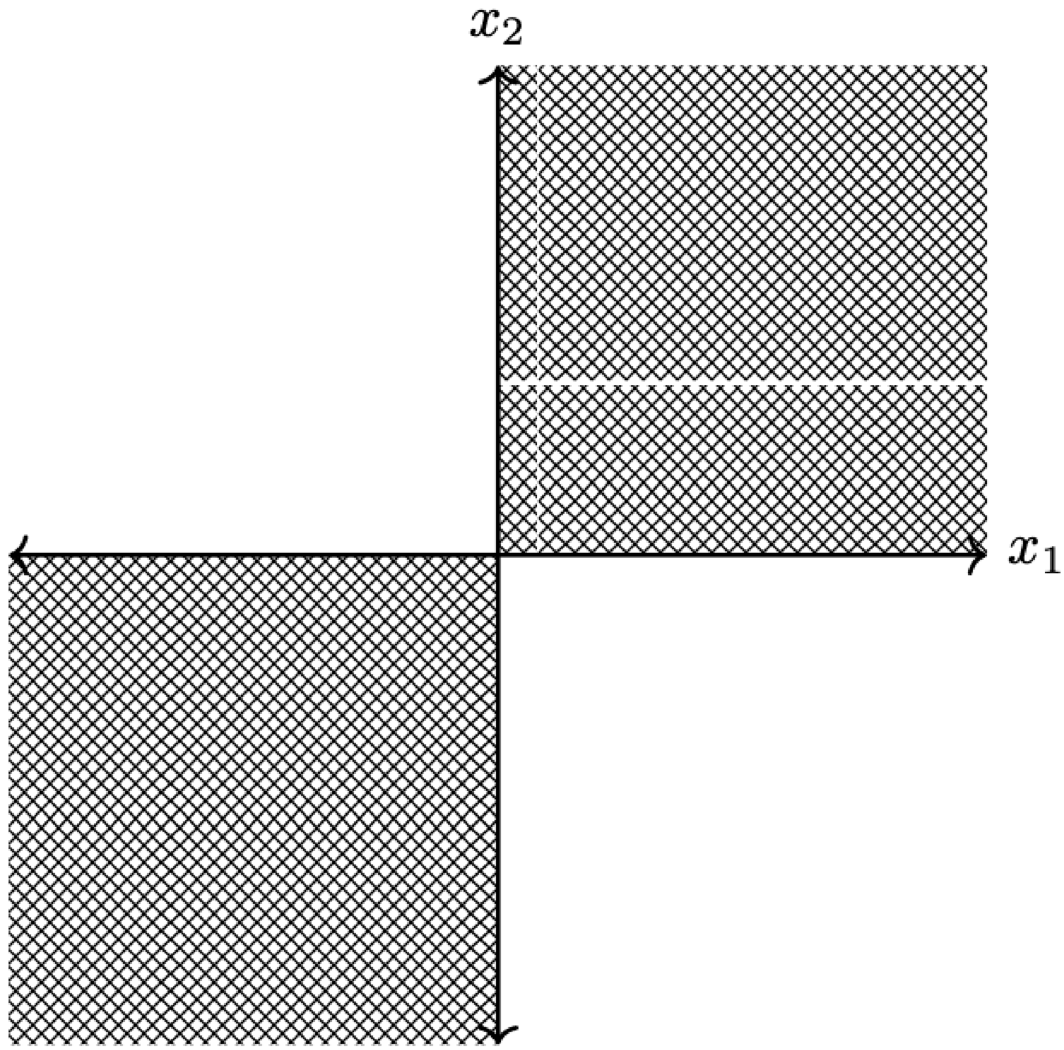
What is the prediction of the QDA classifier at $x = 2.25$?

Solution: Class 0.

Video explanation: https://youtu.be/5VxizVoBHsA

Problem 4.

Suppose that Jake has a wall at home that he has painted to look like the below:



Jake uses the wall to determine how long the Final will be: if he throws a dart at the origin and it lands in the shaded region, the final have 14 questions; if it lands in the unshaded region, it will have 2 questions.

Assume that Jake's dart throws are drawn from a spherical Gaussian whose mean is at the origin. Let X_1 be the horizontal component of a dart throw and X_2 be the vertical component. Let Q be the number of questions on the exam; since it is chosen randomly, it is also a random number. You can assume that the wall is infinitely large, and that the shaded regions extend infinitely up and to the right, and down and to the left.

a) True or False: X_1 and X_2 are independent.

Solution: True.

b) True or False: X_1 and X_2 are conditionally independent given Q.

Solution: False.

c) Let $D = \sqrt{X_1^2 + X_2^2}$, and note that D is the distance of a dart throw from the origin. True or False: D and Q are independent.

Solution: True.

d) True or False: X_1 and and X_2 are conditionally independent given D.

Solution: False.

Problem 5.

Suppose you are training a decision tree classifier to predict whether it will rain in the next hour or not. To do so, you will use two features: the current temperature (in Fahrenheit) and the current pressure (in millibars). Here is some training data that you have collected:

Temperature	**Pressure**	**Rain?**
65	1001	Yes
72	1003	Yes
79	1030	No
55	1022	Yes
62	1025	No
71	1010	Yes
73	1011	No

Train the decision tree to a depth of two. In other words, decide on a root question, splitting the data into two groups, then decide on another question for each group (if necessary). Your resulting decision tree should have two interior (question) nodes and three leaf nodes and should classify all of the training data correctly. Use the Gini coefficient to measure uncertainty. Every time you must decide on a question, choose the one that minimizes the uncertainty of the resulting split. Use as your questions thresholds of the form: "Is temp. < 55?", "Is temp. < 65?", ..., "Is pressure < 1030?".

You may solve this problem by hand, or you may write code to help you solve it. If you use code, you *may not* use any third party libraries besides `numpy` and `pandas`. If you use code, turn in your code with your solution by either screenshot or combining pdfs.

Your answer should include:

- A drawing of the resulting decision tree that labels each interior node with the question that it asks and each leaf node with the class that it predicts.
- For each interior node, a calculation of the uncertainty of all possible splits, with the best split highlighted in some way.

Note: you should choose the *best split* at every step. Namely, your root node should ask the question that minimizes uncertainty. It is possible to choose a different root question and still get a decision tree of depth 2 that perfectly classifies the training data, but this tree will not be considered correct because there is a unique root question minimizing uncertainty.

Solution: Recall, at each node we choose a 'question' or *split* that results in the lowest uncertainty. A split on a set S of labeled points will form left- and right-sets, S_L, S_R, as well as the proportion of points in S that fell into each side, p_L, p_R. The uncertainty of a given split is calculated as the weighted sum of uncertainty of each of the resulting sets. Define the following

$$u(S) = 2p(1-p) = 2(1-p)p$$

$$u(\text{split on } S) = p_L u(S_L) + p_R u(S_R)$$

Our root question will examine full data set. For fun, we can calculate this initial uncertainty as

$$u(S) = 2 \cdot \frac{3}{7} \cdot \frac{4}{7} = 0.490$$

We can calculate uncertainty of possible splits, starting with the lowest to highest x_1, then the lowest to highest x_2, where x_1, x_2 are the Temperature and Pressure features. We will calculate the first three below, and the rest using code.

Notation: The symbol : is often used in set notation, you can read it as meaning "such that".

$$\begin{aligned}
u(x_1 < 55\ ?) &\Rightarrow \text{No } [3,4],\ \text{Yes } [0,0] \\
&= u(S : x_1 \not< 55) + 0 \\
&= u(S) \\
&= 0.490
\end{aligned}$$

$$\begin{aligned}
u(x_1 < 62\ ?) &\Rightarrow \text{No } [3,3],\ \text{Yes } [0,1] \\
&= \frac{6}{7} u(S : x_1 \not< 62) + \frac{1}{7} u(S : x_1 < 62) \\
&= \frac{6}{7}(2 \cdot \frac{1}{2} \cdot \frac{1}{2}) + \frac{1}{7}(2 \cdot 1 \cdot 0) \\
&= 0.429
\end{aligned}$$

$$\begin{aligned}
u(x_1 < 65\ ?) &\Rightarrow \text{No } [2,3],\ \text{Yes } [1,1] \\
&= \frac{5}{7} u(S : x_1 \not< 65) + \frac{2}{7} u(S : x_1 < 65) \\
&= \frac{5}{7}(2 \cdot \frac{2}{5} \cdot \frac{3}{5}) + \frac{2}{7}(2 \cdot \frac{1}{2} \cdot \frac{1}{2}) \\
&= 0.486
\end{aligned}$$

$$u(x_1 < 71\ ?) \Rightarrow \ldots$$

Once performed over all possible splits on S, we arrive at the minimum uncertainty with the split $x_2 < 1011$, with

$$\begin{aligned}
u(x_2 < 1011\ ?) &\Rightarrow \text{No } [3,1],\ \text{Yes } [0,3] \\
&= 0.214
\end{aligned}$$

Note that uncertainty is zero at our node $\{S : x_2 < 1011\}$, thus we consider that node a leaf. However we must continue reducing uncertainty for our node $\{S : x_2 \not< 1011\}$. To do this, we repeat the above process, but examining only this subset of data.

Since the current node was split using x_2, it is wise to examine splits using x_1. The first two are calculated below.

$$
\begin{aligned}
u(x_1 < 55\ ?: x_2 \not< 1011) &= u(S : x_2 \not< 1011, x_1 \not< 55) + 0 \\
&= u(S : x_2 \not< 1011) \\
&= 2 \cdot \frac{3}{4} \cdot \frac{1}{4} \\
&= 0.375
\end{aligned}
$$

$$
\begin{aligned}
u(x_1 < 62\ ?: x_2 \not< 1011) &\Rightarrow \text{No } [0, 1],\ \text{Yes } [3, 0] \\
&= \frac{1}{4} u(S : x_2 \not< 1011, x_1 \not< 62) + \frac{3}{4} u(S : x_2 \not< 1011, x_1 < 62) \\
&= \frac{1}{4}(2 \cdot 0 \cdot 1) + \frac{3}{4}(2 \cdot 1 \cdot 0) \\
&= 0
\end{aligned}
$$

Aha! We've found another split that results in zero uncertainty. We can now consider the nodes $\{S : x_2 \not< 1011, x_1 < 62\}$ and $\{S : x_2 \not< 1011, x_1 \not< 62\}$ to be leaves!

Our final decision tree structure looks like