DSC 80 Discussion 9 Worksheet

1 WI24 Final Problem 10

Harshi is trying to build a decision tree that predicts whether or not a hotel has a swimming pool. Suppose + represents the "has pool" class and ∘ represents the "no pool" class. One of the nodes in Harshi's decision tree has 12 points, with the following distribution of classes.

$$+ + + + \circ \circ \circ \circ \circ \circ \circ \circ$$

Consider the following three splits of the node above.

- **Split 1: Yes:** $+ + \circ \circ \circ \circ$ **No:** $+ + \circ \circ \circ \circ$
- **Split 2: Yes:** $+ + \circ \circ \circ \circ \circ \circ \circ$ **No:** $+ + \circ$
- **Split 3: Yes:** $\circ$ **No:** $+ + + + \circ \circ \circ \circ \circ \circ \circ$

Which of the six nodes above have the same entropy as the original node? **Select all that apply.**

[ ] Split 1's "Yes" node

[ ] Split 1's "No" node

[ ] Split 2's "Yes" node

[ ] Split 2's "No" node

[ ] Split 3's "Yes" node

[ ] Split 3's "No" node

Which of the six nodes above have the lowest entropy? (There can be multiple correct answers.)

[ ] Split 1's "Yes" node

[ ] Split 1's "No" node

[ ] Split 2's "Yes" node

[ ] Split 2's "No" node

[ ] Split 3's "Yes" node

[ ] Split 3's "No" node

2 WI24 Final Problem 12

Diego wants to build a model that predicts the number of open rooms a hotel has, given various other features. He has a training set with 1200 rows. Diego fits a regression model using the `GPTRegression` class.

`GPTRegression` models have several hyperparameters that can be tuned, including `context_length` and sentience. To choose between 5 possible values of `context_length`, Diego performs k-fold cross-validation.

How many total times is a `GPTRegression` model fit?

[ ] $4k$

[ ] $5k$

[ ] $240k$

[ ] $6000k$

[ ] $4(k-1)$

[ ] $5(k-1)$

[ ] $240(k-1)$

[ ] $6000(k-1)$

Suppose that every time a `GPTRegression` model is fit, it appends the number of points in its training set to the list `sizes`. Note that after performing cross-validation, `len(sizes)` is equal to your answer to the previous subpart.
What is `sum(sizes)`?

[ ] $4k$

[ ] $5k$

[ ] $240k$

[ ] $6000k$

[ ] $4(k-1)$

[ ] $5(k-1)$

[ ] $240(k-1)$

[ ] $6000(k-1)$

3 SP22 Final Problem 12

After fitting a classifier, we use it to make predictions on an unseen test set. Our results are summarized in the following confusion matrix.

	Predicted Negative	Predicted Positive
Actually Negative	???	30
Actually Positive	66	105

What is the recall of our classifier? Give your answer as a fraction (it does not need to be simplified).

_____

The accuracy of our classifier is $\frac{69}{117}$. How many **true negatives** did our classifier have?

_____

True or False: In order for a binary classifier's precision and recall to be equal, the number of mistakes it makes must be an even number.

_____

Suppose we are building a classifier that listens to an audio source (say, from your phone's microphone) and predicts whether or not it is Soulja Boy's 2008 classic "Kiss Me thru the Phone." Our classifier is pretty good at detecting when the input stream is "Kiss Me thru the Phone", but it often incorrectly predicts that similar sounding songs are also "Kiss Me thru the Phone."
Our classifier has:

`[ ]` low precision and low recall.

`[ ]` low precision and high recall.

`[ ]` high precision and low recall.

`[ ]` high precision and high recall.

4 WI23 Final Problem 8.4

Let's say we're performing k-fold cross-validation to find the best combination of two hyperparameters called `height` and `color`.
For the purposes of this question, assume that:

- Our training set contains n rows, where n is greater than 5 and is a multiple of k.
- There are h_1 possible values of `height` and h_2 possible values of `color`.

Solve for the following quantities, **in terms of** n, k, h_1, **and** h_2:

- What is the size of each fold? _____
- How many times is row 5 in the training set used for training? _____
- How many times is row 5 in the training set used for validation? _____