Discussion 9

DSC 80

2024-05-31

Section 1

WI24 Final Problem 10

WI24 Final Problem 10

One of the nodes in Harshi's decision tree has 12 points, with the following distribution of classes.

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

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

Problem

Which node has the same entropy as the original node?

$$\text{entropy} = -\sum_{C} p_C \log_2 p_C$$

What does entropy try to capture?

Solution

- We can do this question without having to actually calculate entropy by observing something about entropy in binary classification settings.

Solution

- We can do this question without having to actually calculate entropy by observing something about entropy in binary classification settings.
- In binary classification, two nodes will have the same entropy if they have the same relative distribution of classes, or if the distributions of classes are complements of each other.

Solution

- We can do this question without having to actually calculate entropy by observing something about entropy in binary classification settings.
- In binary classification, two nodes will have the same entropy if they have the same relative distribution of classes, or if the distributions of classes are complements of each other.
- For example, a node with 75% positives and 25% negatives has the same entropy as a node with 25% positives and 75% negatives.

Solution

- So, we just need to find nodes with the same proportions as the original node, which has 4 positives and 8 negatives
- Which ones are those?

Part 2

Which of the six nodes above have the lowest entropy? If there are multiple correct answers, select them all.

$$\text{entropy} = -\sum_{C} p_C \log_2 p_C$$

What situation leads to the minimal entropy – once again, try doing this without evaluating the formula!

Which of the provided nodes would cause this to happen?

Section 2

WI24 Final Problem 12

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.

Problem

How many times is a `GPTRegression` model fit?

Our ingredients: - k folds - 5 possible values of one hyperparameter - 1200 rows in our dataset

Let's try to step through k-fold CV to see what happens.

Solution

- How many times do we fit a model for each value of the hyperparameter?

Solution

- How many times do we fit a model for each value of the hyperparameter?
- We fit a model for every partition of the dataset with one fold held out

Solution

- How many times do we fit a model for each value of the hyperparameter?
- We fit a model for every partition of the dataset with one fold held out
- So, we get k fits for each hyperparameter value

Part 2

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)`?

Solution

- Will this number vary for each model fit?

Solution

- Will this number vary for each model fit?
- If not, then we can multiply our previous answer by the number of points in the training set each time – what is that value in terms of k?

Solution

- Will this number vary for each model fit?
- If not, then we can multiply our previous answer by the number of points in the training set each time – what is that value in terms of k?
- $\frac{k-1}{k} \cdot 1200$

Section 3

SP22 Final Problem 12

Problem

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

Recall

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

- What is the formula for recall? What is it trying to capture?
- What elements in the table do we need to calculate that?

True Negatives

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

We need to formulate accuracy as a function of the values we have in this table – what will that be?

Even Numbers

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.

- Think about the equations for precision and recall. What is "the number of mistakes" a classifier makes in those same terms?
- What needs to be true for precision and recall to be equal?

Soulja Boy

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."

Classifier

Which of the given statements is true about the classifier?

- What is a true positive? A false positive? A true negative? A false negative?
- Without thinking about the formulas: what would good precision look like here? Good recall?

Section 4

WI23 Final Problem 8.4

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.

Problem

- 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?

Section 5

Attendance

Attendance

Once I give you a number, fill out the following Google form: https://forms.gle/3br6oaLshaZx1oot9