Quiz Submissions - Quiz 3 - Attempt 2

Nhat Le (username: hung.le@mail.mcgill.ca)

Attempt 1

Written: Feb 8, 2019 9:05 PM - Feb 8, 2019 9:26 PM

Submission View

Released: Jan 23, 2019 11:59 PM

View the quiz answers.

Question 1 1 / 1 point

What contains more information, the toss of a loaded die that has a 90% chance of rolling a six (and a 2% chance of every other number), or the outcome of a fair coin toss?

○ The loaded die contains more information.

✓ ○ The fair coin toss contains more information.

○ They contain the same amount of information

○ Not enough information is given

▼ Hide Feedback

Using the formula for entropy (i.e., information content) from lecture (Lecture 7, slide 22), we can get that the amount of information in the toss of the loaded die is

$$-0.9\log_2(0.9) - 5\times 0.02\log_2(0.02) \approx 0.701$$

{"version":"1.1","math":"-0.9\log_2(0.9)-5\times 0.02\log_2(0.02) \approx 0.701"}

A fair coin toss, on the other hand, provides 1 bit of information as

$$-0.5\log_2(0.5) - 0.5\log_2(0.5) = 1$$

{"version":"1.1","math":"-0.5\log_2(0.5)-0.5\log_2(0.5)=1"}

Question 2 0 / 1 point

Suppose you have a multi-class classification dataset with **4 possible classes and 30 binary features.** How many parameters do you need to learn to fit a Naive Bayes model to this dataset? Note that this dataset is multi-class, meaning that the domain of target *y* value is a discrete set with four possible values.

➡ ○ 123

○ 61

✗ ○ 124

○ 122

▼ Hide Feedback

We need to learn the class distribution P(y) and the conditional distribution of the features given each class. Since there are 4 classes, P(y) requires 3 parameters to learn, as specifying the marginal probability of three of classes determines the probability of the fourth class. (Note, for instance, that in the binary classification case we only needed 1 feature to represent P(y) when there was two classes).

In addition, to estimate P(x|y=k) for each class, we need 30 parameters. Thus, in total, there are 3+4*30=123 parameters to learn.

Question 3 1 / 1 point

Based on lecture, which of the following statements would you say is **false:**

○ Decision trees are a relatively interpretable machine learning model.

○ The learning process for decision trees can be sensitive to small changes in the input data.

○ Decision trees work well when fitting data with axis-orthogonal class boundaries.

✓ ○ Many efficient learning algorithms exist for decision trees with general non-linear, non-binary tests.

▼ Hide Feedback

As discussed in Lecture 7, slide 42 decision trees are interpretable (relative to other ML models). Moreover, decision trees are **good** at fitting data that is separable along axis-orthogonal boundaries (Lecture 7, slide 43), and the sensitivity of decision trees is discussed as a limitation in Lecture 7, slide 43. However, efficient learning algorithms only exist for decision trees with binary tests.

Question 4 1 / 1 point

Which of the following describes a true benefit of post-pruning compared to early stopping for regularizing decision trees:

○ Post pruning is generally more computational efficient and saves on runtime.

✓ ○ Post pruning is better at accommodating complex dependencies between the input features.

○ Post pruning is overall less prone to overfitting.

○ None of the above

▼ Hide Feedback

As discussed in Lecture 7, slide 39 post pruning "allows you to deal with cases where a single attribute is not informative, but a combination of attributes is informative", which means that is better at accommodating complex dependencies between the input features.

Question 5 1 / 1 point

Suppose you are learning a decision tree for email spam classification. Your current sample of the training data has the following distribution of labels:

- *[43+, 30-]*

i.e., the training sample has 43 examples that are spam and 30 that are not spam. Now, you are choosing between two candidate tests.

Test 1 (T1) tests whether the number of words in the email is greater than 30 and would result in the following splits:

- *num_words > 30 : [5+, 15-]*
- *num_words <= 30: [38+, 15-]*

Test 2 (T2) tests whether the email contains an external URL link and would result in the following splits:

- *has_link: [25+, 5-]*
- *not_has_link: [18+, 25-]*

Which test should you use to split the data? I.e., which test provides a higher information gain?

○ Choose T1, since it provides higher information gain.

✓ ○ Choose T2, since it provides higher information gain.

○ They provide the same information gain.

○ Not enough information is provided.

▼ Hide Feedback

The conditional entropy for test 1 is given by:

$$H(\text{data}\mid T1) = -\frac{20}{73}\left(\frac{5}{20}\log_2\left(\frac{5}{20}\right) + \frac{15}{20}\log_2\left(\frac{15}{20}\right)\right) - \frac{53}{73}\left(\frac{38}{53}\log_2\left(\frac{38}{53}\right) + \frac{15}{53}\log_2\left(\frac{15}{53}\right)\right)$$
$$\approx 0.846$$

{"version":"1.1","math":"H(\textrm{data}\: |\: T1) = -\frac{20}{73}\left(\frac{5}{20}\log_2\left(\frac{5}{20}\right) + \frac{15}{20}\log_2\left(\frac{15}{20}\right) \right) -\frac{53

The conditional entropy for test 2 is given by:

$$H(\text{data}\mid T2) = -\frac{30}{73}\left(\frac{25}{30}\log_2\left(\frac{25}{30}\right) + \frac{5}{30}\log_2\left(\frac{5}{30}\right)\right) - \frac{43}{73}\left(\frac{18}{43}\log_2\left(\frac{18}{43}\right) + \frac{25}{43}\log_2\left(\frac{25}{43}\right)\right)$$
$$\approx 0.845$$

{"version":"1.1","math":"H(\textrm{data}\: |\: T2) = -\frac{30}{73}\left(\frac{25}{30}\log_2\left(\frac{25}{30}\right) + \frac{5}{30}\log_2\left(\frac{5}{30}\right) \right) -\frac{43

Since the conditional entropy of the data given T2 is lower, this test provides a higher information gain.

Attempt Score: 4 / 5 - 80 %

Overall Grade (highest attempt)**:** 4 / 5 - 80 %

Done