Quiz Submissions - Quiz 3 - Attempt 1

✕

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

Attempt 1

Written: Feb 6, 2019 10:55 PM - Feb 6, 2019 11:42 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 75% chance of rolling a six (and a 5% 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 a loaded die is

$$-0.75\log_2(0.75) - 5 \times 0.05\log_2(0.05) \approx 1.392$$

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

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

Question 2 **0 / 1 point**

Suppose you have two **binary** classification datasets: Dataset A has *m* binary features and Dataset B has *m* continuous (i.e., real-valued) features. You plan to run "Binary Naive Bayes" (i.e., Naive Bayes with binary features) on Dataset A and Gaussian Naive Bayes on Dataset B. Which dataset/model requires more parameters to learn?

➡ ○ Gaussian Naive Bayes requires more parameters

○ Binary Naive Bayes requires more parameters

✗ ○ They require the same number of parameters

○ Not enough information

▼ Hide Feedback

Both methods require that you learn the class distribution P(y), so there is no difference there. Since the output is binary, estimating P(y) requires a single parameter (i.e., we need to estimate

$$\theta_y$$

which gives the estimated marginal probability that y is equal to 1).

For Binary Naive Bayes, we also need to estimate P(x|y=1) and P(x|y=0) for each feature. Each of these estimates is a single parameter value, so we need to estimate 2m+1 parameters in total for the binary case.

In the Gaussian Naive Bayes case, we also need to estimate P(x|y=1) and P(x|y=0) for each feature. In this case, we estimate P(x|y=k) as a Gaussian and we need to learn the mean and variance for each feature, which requires 2 parameters to learn. Thus in total we have (2 classes)*(2 parameters to learn the conditional distribution of each feature for each class)*(m features)+(1 parameter to learn the marginal likelihood of the target class)=4m+1 parameters.

Question 3 **1 / 1 point**

Which of the following is a drawback of decision trees:

○ Learned decision trees are generally uninterpretable

○ The learning process is very expensive when binary tests are used.

✓ ○ The learning process is sensitive to small changes in the input data.

○ Decision trees struggle to fit data that is separable along axis-orthogonal boundaries.

▼ Hide Feedback

As discussed in Lecture 7, slide 42 the interpretability and fast learning algorithms are **benefits** not limitations. Moreover, decision trees are **good** at fitting data that is separable along axis-orthogonal boundaries (Lecture 7, slide 43). The sensitivity of decision trees is discussed as a limitation in Lecture 7, slide 43.

Question 4 **1 / 1 point**

Which of the following is **not** a standard approach to regularize the learning of decision trees:

○ Early stopping

✓ ○ L2 regularization

○ Post pruning

○ None of the above (i.e., all three approaches are standard regularization techniques for decision trees)

▼ Hide Feedback

As discussed in Lecture 7, slide 39 early stopping and post pruning are both approaches to regularize decision trees. L2 regularization does not naturally apply to decision trees since there is no parameter vector to apply the penalty to.

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 20 and would result in the following splits:

- *num_words > 20 : [13+, 20-]*
- *num_words <= 20: [30+, 10-]*

Test 2 (T2) tests whether the email contains spelling errors and would result in the following splits:

- *spelling_error: [30+, 15-]*
- *no_spelling_error: [13+, 15-]*

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:

$$H(\text{data} \mid T1) = -\frac{13+20}{13+20+30+10}\left(\frac{13}{13+20}\log_2\left(\frac{13}{13+20}\right)+\frac{20}{13+20}\log_2\left(\frac{20}{13+20}\right)\right)$$

$$-\frac{30+10}{13+20+30+10}\left(\frac{30}{30+10}\log_2\left(\frac{30}{30+10}\right)+\frac{10}{30+10}\log_2\left(\frac{10}{30+10}\right)\right) \approx 0.882$$

The conditional entropy for test 2 is given by:

$$H(\text{data} \mid T2) = -\frac{30+15}{30+15+13+15}\left(\frac{30}{30+15}\log_2\left(\frac{30}{30+15}\right)+\frac{15}{30+15}\log_2\left(\frac{15}{30+15}\right)\right)$$

$$-\frac{13+15}{30+15+13+15}\left(\frac{13}{13+15}\log_2\left(\frac{13}{13+15}\right)+\frac{15}{13+15}\log_2\left(\frac{15}{13+15}\right)\right) \approx 0.948$$

Since the conditional entropy of the data given T1 is lower, this test provides a higher information gain.

Attempt Score: 4 / 5 - 80 %

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

Done