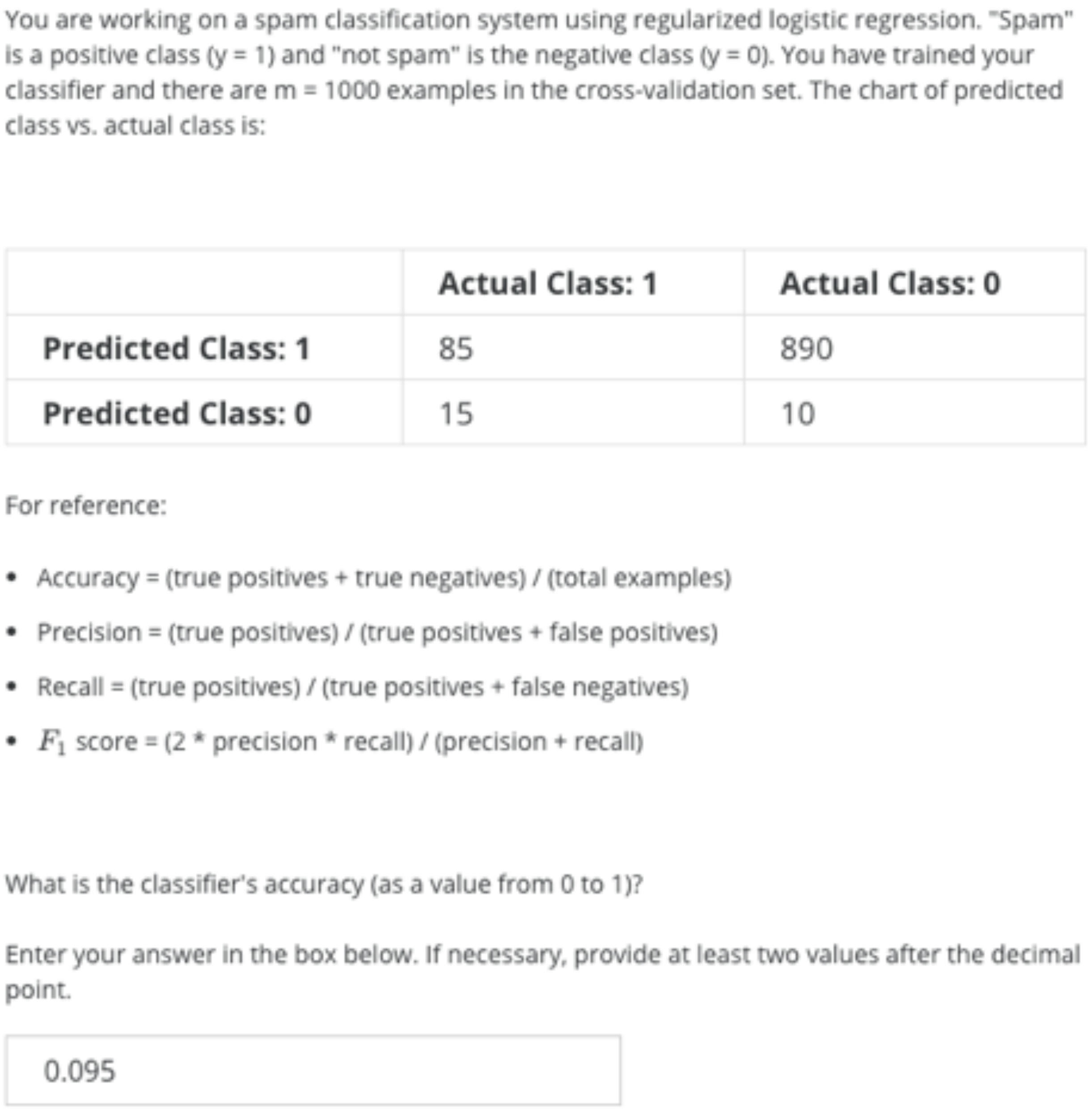
Week 6

Quiz 2- Machine Learning System Design

Ques1.



You are working on a spam classification system using regularized logistic regression. "Spam" is a positive class (y = 1) and "not spam" is the negative class (y = 0). You have trained your classifier and there are m = 1000 examples in the cross-validation set. The chart of predicted class vs. actual class is:

	Actual Class: 1	**Actual Class: 0**
Predicted Class: 1	85	890
Predicted Class: 0	15	10

For reference:

- Accuracy = (true positives + true negatives) / (total examples)
- Precision = (true positives) / (true positives + false positives)
- Recall = (true positives) / (true positives + false negatives)
- F_1 score = (2 * precision * recall) / (precision + recall)

What is the classifier's accuracy (as a value from 0 to 1)?

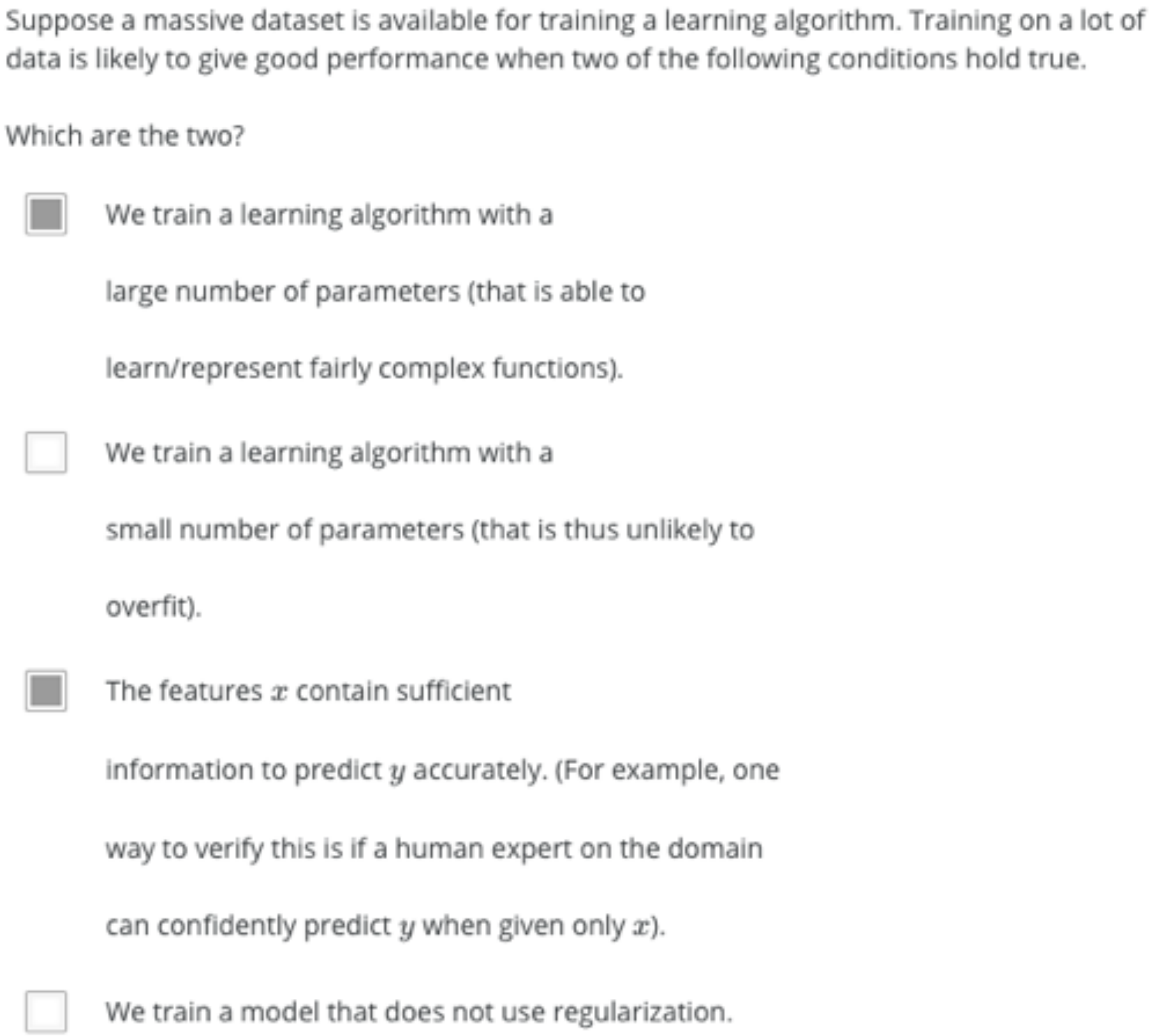
Enter your answer in the box below. If necessary, provide at least two values after the decimal point.

0.095

Answer:

Accuracy = (true positives + true negatives) / (total examples) = (85+10) / (1000) = 0.095

Ques2.



Suppose a massive dataset is available for training a learning algorithm. Training on a lot of data is likely to give good performance when two of the following conditions hold true.

Which are the two?

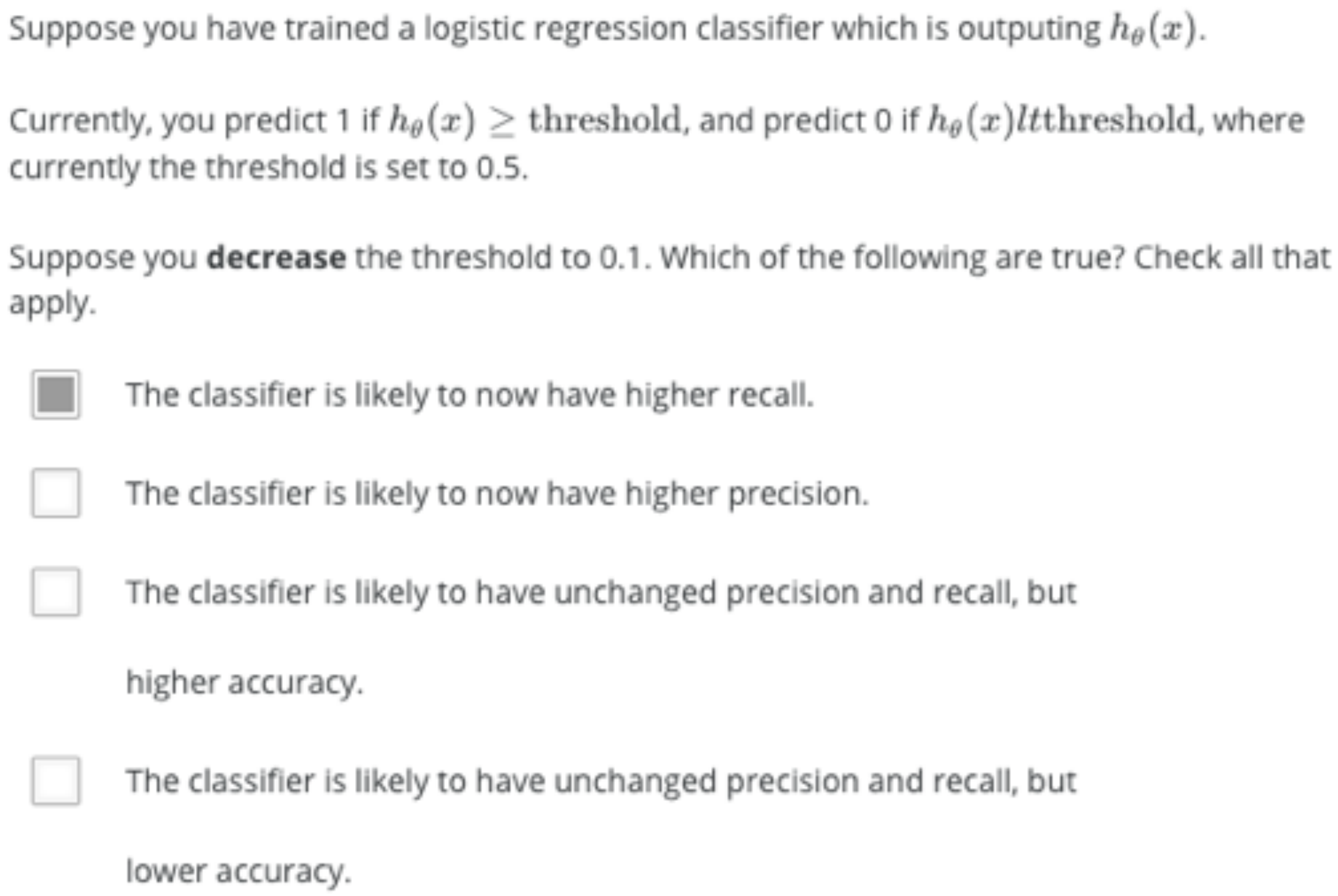
☑ We train a learning algorithm with a large number of parameters (that is able to learn/represent fairly complex functions).

□ We train a learning algorithm with a small number of parameters (that is thus unlikely to overfit).

☑ The features x contain sufficient information to predict y accurately. (For example, one way to verify this is if a human expert on the domain can confidently predict y when given only x).

□ We train a model that does not use regularization.

Ques3.



Suppose you have trained a logistic regression classifier which is outputing $h_\theta(x)$.

Currently, you predict 1 if $h_\theta(x) \geq \mathrm{threshold}$, and predict 0 if $h_\theta(x) lt\mathrm{threshold}$, where currently the threshold is set to 0.5.

Suppose you **decrease** the threshold to 0.1. Which of the following are true? Check all that apply.

☑ The classifier is likely to now have higher recall.

□ The classifier is likely to now have higher precision.

□ The classifier is likely to have unchanged precision and recall, but

higher accuracy.

□ The classifier is likely to have unchanged precision and recall, but

lower accuracy.

Answer:

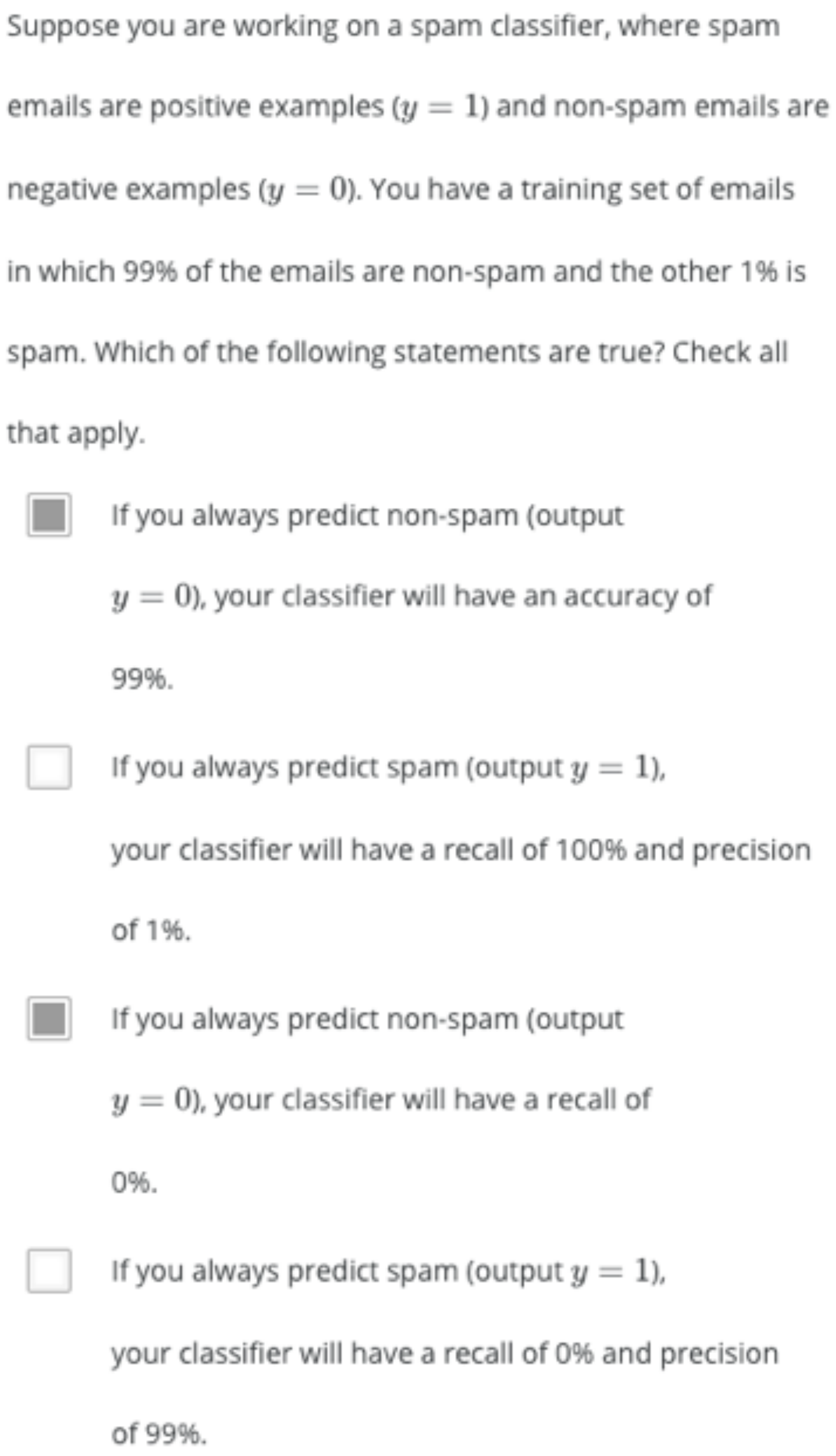
When the threshold has been decreased to 0.1,

Recall = (true positives) / (true positives + false negatives), true positives are more, (true positives + false negatives) remains unchanged. Thus, recall increases.

Precision = (true positives) / (true positives + false positives), true positives are more, (true positives + false positives) are more also. Thus, precision undetermined.

Accuracy = (true positives + true negatives) / (total examples), true positives are more, (total examples) remains unchanged. Thus, accuracy increases.

Ques4.



Suppose you are working on a spam classifier, where spam emails are positive examples ($y = 1$) and non-spam emails are negative examples ($y = 0$). You have a training set of emails in which 99% of the emails are non-spam and the other 1% is spam. Which of the following statements are true? Check all that apply.

☑ If you always predict non-spam (output $y = 0$), your classifier will have an accuracy of 99%.

□ If you always predict spam (output $y = 1$), your classifier will have a recall of 100% and precision of 1%.

☑ If you always predict non-spam (output $y = 0$), your classifier will have a recall of 0%.

□ If you always predict spam (output $y = 1$), your classifier will have a recall of 0% and precision of 99%.

Answer:
When always predicting non-spam:

Accuracy = (true positives + true negatives) / (total examples), (true positives + true negatives) is (0 + 99%), total example is 100%. Accuracy = 99%.

Recall = (true positives) / (true positives + false negatives), true positive is 0, (true positives + false negatives) = actual positive = 1%, thus recall = 0.

Precision = (true positives) / (true positives + false positives), true positive is 1%, (true positives + false positives) = predicted positive = 0%, thus precision = 0.

When always predicting spam:

Accuracy = (true positives + true negatives) / (total examples), (true positives + true negatives) is (1% + 0), total example is 100%. Accuracy = 1%.

Recall = (true positives) / (true positives + false negatives), true positive is 1%, (true positives + false negatives) = actual positive = 1%, thus recall = 100%.

Precision = (true positives) / (true positives + false positives), true positive is 1%, (true positives + false positives) = predicted positive = 100%, thus precision = 1%.

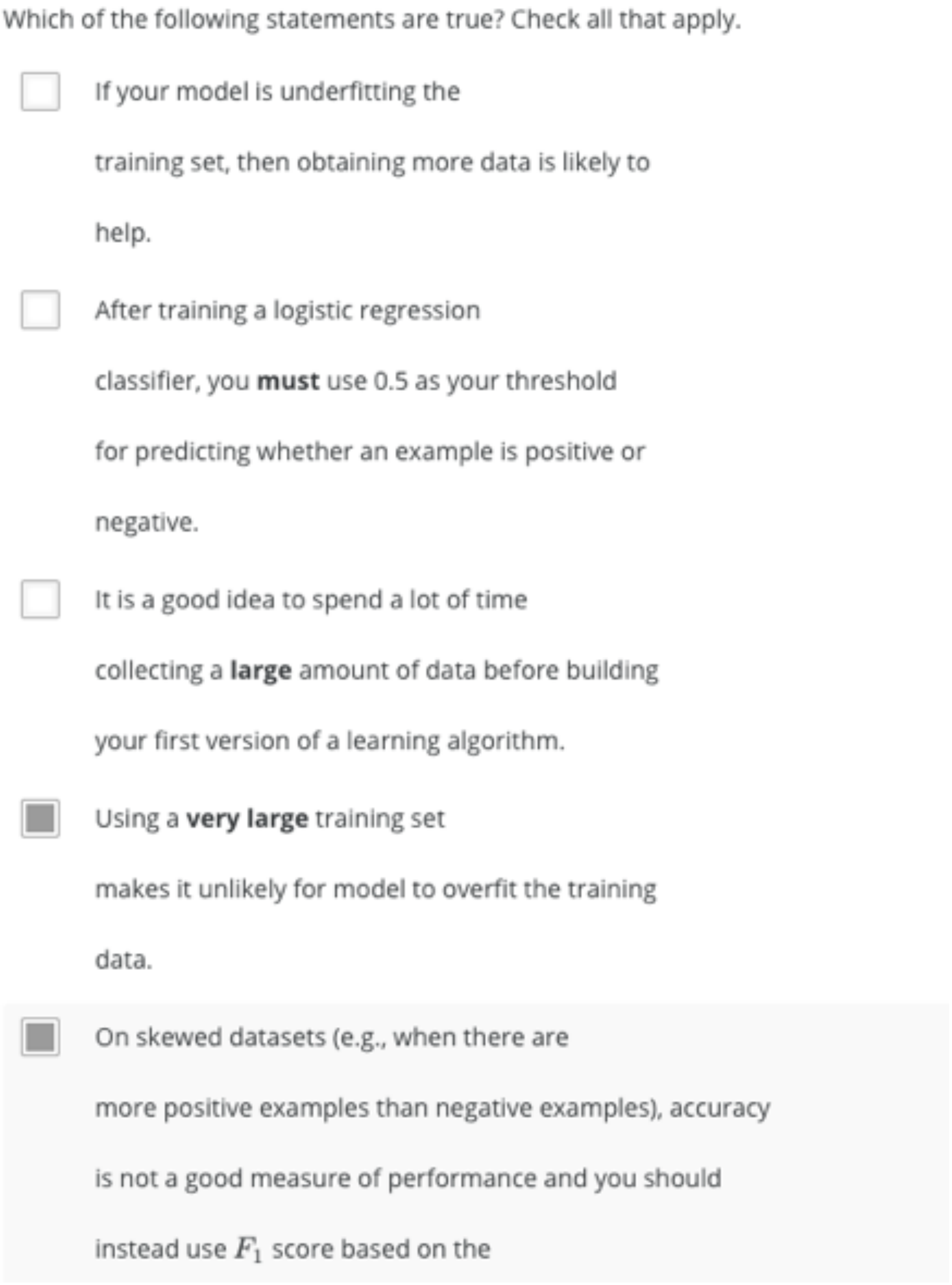
A) Correct.

B) Correct.

C) Correct.

D) Wrong.

Ques5.



Which of the following statements are true? Check all that apply.

□ If your model is underfitting the training set, then obtaining more data is likely to help.

□ After training a logistic regression classifier, you **must** use 0.5 as your threshold for predicting whether an example is positive or negative.

□ It is a good idea to spend a lot of time collecting a **large** amount of data before building your first version of a learning algorithm.

☒ Using a **very large** training set makes it unlikely for model to overfit the training data.

☒ On skewed datasets (e.g., when there are more positive examples than negative examples), accuracy is not a good measure of performance and you should instead use F_1 score based on the

Answer:
A) When there are not enough useful features, feeding more data will not help.
B) Not necessary.
C) Not always the case.
D) Large dataset will not overfit the model.
E) True.