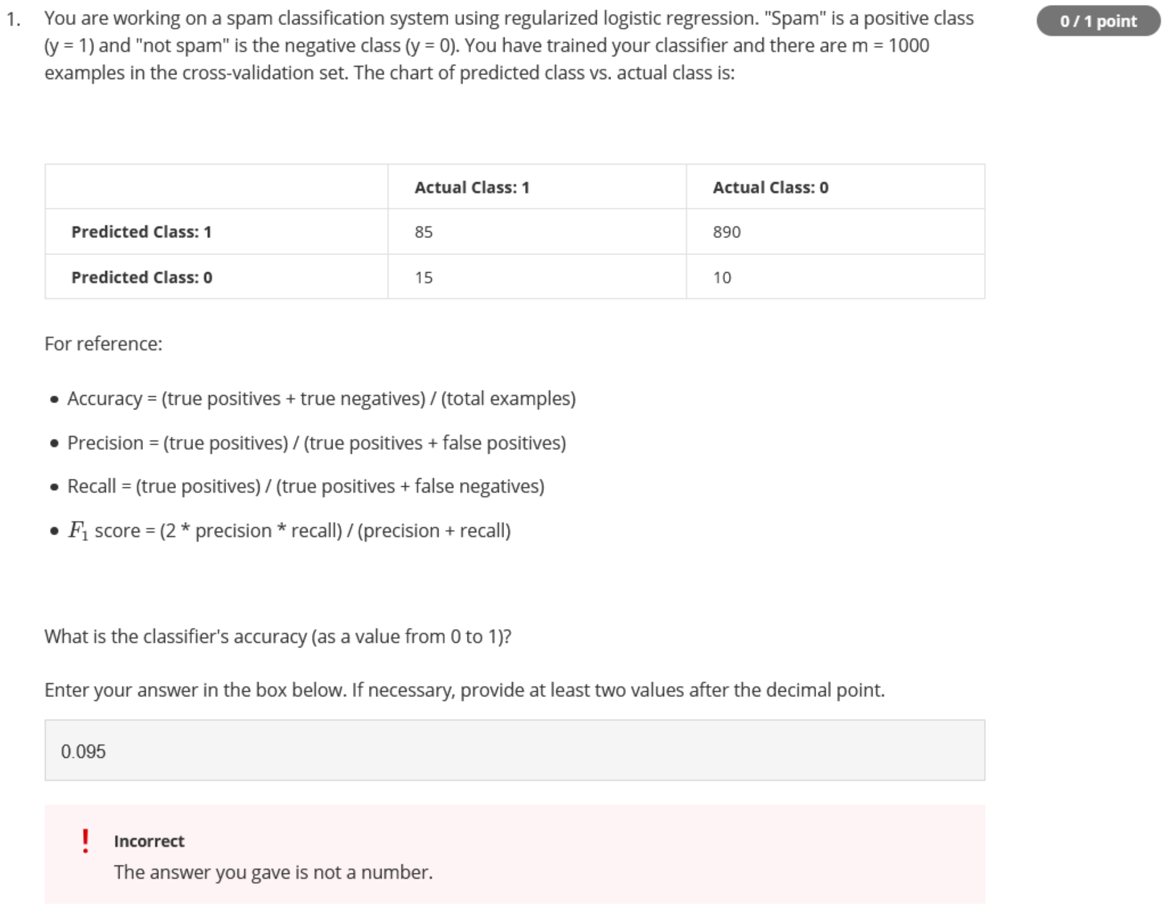
Quiz

Machine Learning System Design



1. 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:

0 / 1 point

	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

! **Incorrect**

The answer you gave is not a number.

2. 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.

1 / 1 point

Which are the two?

☑ A human expert on the application domain

can confidently predict y when given only the features x

(or more generally, if we have some way to be confident

that x contains sufficient information to predict y

accurately).

✓ **Correct**
It is important that the features contain sufficient information, as otherwise no amount of data can solve a learning problem in which the features do not contain enough information to make an accurate prediction.

□ The classes are not too skewed.

☑ Our learning algorithm is able to

represent fairly complex functions (for example, if we

train a neural network or other model with a large

number of parameters).

✓ **Correct**
You should use a complex, "low bias" algorithm, as it will be able to make use of the large dataset provided. If the model is too simple, it will underfit the large training set.

□ When we are willing to include high

order polynomial features of x (such as x_1^2, x_2^2,

$x_1 x_2$, etc.).

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

1 / 1 point

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

Suppose you **increase** the threshold to 0.7. Which of the following are true? Check all that apply.

☑ The classifier is likely to now have lower recall.

✓ **Correct**

Increasing the threshold means more y = 0 predictions. This will increase the decrease of true positives and increase the number of false negatives, so recall will decrease.

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

□ The classifier is likely to now have lower precision.

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

4. 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.

1 / 1 point

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

✓ **Correct**

Since every prediction is y = 1, there are no false negatives, so recall is 100%. Furthermore, the precision will be the fraction of examples with are positive, which is 1%.

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

☑ If you always predict non-spam (output

$y = 0$), your classifier will have a recall of

0%.

✓ **Correct**
Since every prediction is y = 0, there will be no true positives, so recall is 0%.

☑ If you always predict non-spam (output

$y = 0$), your classifier will have an accuracy of

99%.

✓ **Correct**
Since 99% of the examples are y = 0, always predicting 0 gives an accuracy of 99%. Note, however, that this is not a good spam system, as you will never catch any spam.

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

1 / 1 point

☑ 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

precision and recall.

✓ **Correct**
You can always achieve high accuracy on skewed datasets by predicting the most the same output (the most common one) for every input. Thus the F_1 score is a better way to measure performance.

☑ Using a **very large** training set

makes it unlikely for model to overfit the training

data.

✓ **Correct**
A sufficiently large training set will not be overfit, as the model cannot overfit some of the examples without doing poorly on the others.

□ 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.

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

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