## Machine Learning System Design

## Latest Submission Grade 80%

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:



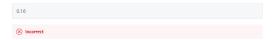
	Actual Class: 1	Actual Class: 0
Predicted Class: 1	85	890
Predicted Class: 0	15	10

## For reference:

- Accuracy = (true positives + true negatives) / (total examples)
- Precision = (true positives) / (true positives + false positives)
- Recall = (true positives) / (true positives + false negatives)
- F<sub>1</sub> score = (2 \* precision \* recall) / (precision + recall)

What is the classifier's precision (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.



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.



We train a learning algorithm with a

small number of parameters (that is thus unlikely to

information to predict  $\boldsymbol{y}$  accurately. (For example, one

way to verify this is if a human expert on the domain

can confidently predict  $\boldsymbol{y}$  when given only  $\boldsymbol{x}$ ).

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.

☐ When we are willing to include high

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

 $x_1x_2$ , etc.).

We train a learning algorithm with a

large number of parameters (that is able to

learn/represent fairly complex functions).

© Correct
You should use a "low bias" algorithm with many parameters, as it will be able to make use of the large dataset provided. If the model has too few parameters, it will underfit the large training set.

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

1/1 point

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

Suppose you  ${\bf increase}$  the threshold to 0.7. Which of the following are true? Check all that apply.

☐ The classifier is likely to now have higher recall.

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

 $\ensuremath{ \ensuremath{ \begin{tabular} \$ 

© correct Increasing the threshold means more y = 0 predictions. This will decrease both true and false positives, so precision will increase.

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

thus the same  $F_1$  score.

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

1/1 point

If your model is underfitting the

training set, then obtaining more data is likely to