1 point 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)
- F₁ 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.

0.99

point

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.

Which are the two?

When we are willing to include high order polynomial features of x (such as x_1^2 , x_2^2 , x_1x_2 , etc.).

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

large number of parameters (that is able to learn/represent fairly complex functions).

We train a learning algorithm with a

point

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

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 have unchanged precision and recall, but higher accuracy.

The classifier is likely to now have lower recall.

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

The classifier is likely to now have lower precision.

point

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

If you always predict non-spam (output y=0), your classifier will have 99% accuracy on the

that apply.

training set, but it will do much worse on the cross validation set because it has overfit the training data.

y=0), your classifier will have 99% accuracy on the

A good classifier should have both a

If you always predict non-spam (output

training set, and it will likely perform similarly on the cross validation set.

high precision and high recall on the cross validation

y=0), your classifier will have an accuracy of 99%.

If you always predict non-spam (output

point

After training a logistic regression

classifier, you must use 0.5 as your threshold

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

for predicting whether an example is positive or negative.

makes it unlikely for model to overfit the training

Using a very large training set

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

precision and recall. If your model is underfitting the

training set, then obtaining more data is likely to

help.

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.

提交测试