Feedback — XI. Machine Learning System Design

Help

You submitted this quiz on **Mon 5 May 2014 12:04 AM PDT**. You got a score of **5.00** out of **5.00**.

Question 1

You are working on a spam classification system using regularized logistic regression. "Spam" is the 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	**0**
Predicted Class	**1**	85	890
	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 recall (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.

You entered:

0.85

Your Answer		Score	Explanation
0.85	✔	1.00	There are 85 true positives and 15 false negatives, so recall is 85 / (85 + 15) = 0.85.
Total		1.00 / 1.00	

Question 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?

Your Answer		Score	Explanation
☑ 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).	✔	0.25	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.
☑ We train a learning algorithm with a large number of parameters (that is able to learn/represent fairly complex functions).	✔	0.25	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.
☐ We train a learning algorithm with a small number of parameters (that is thus unlikely to overfit).	✔	0.25	If the model has a small number of parameters, then it will underfit the large training set and not make good use of all the data.
☐ The classes are not too skewed.	✔	0.25	The problem of skewed classes is unrelated to training with large datasets.
Total		1.00 / 1.00	

Question 3

Suppose you have trained a logistic regression classifier which is outputing $h_\theta(x)$. 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 **decrease** the threshold to 0.3. Which of the following are true? Check all that apply.

Your Answer		Score	Explanation
☐ The classifier is likely to have unchanged precision and recall, and thus the same F_1 score.	✔	0.25	By making more y = 1 predictions, we increase true and false positives and decrease true and false negatives. Thus, precision and recall will certainly change.
☐ The classifier is likely to now have higher precision.	✔	0.25	Lowering the threshold means more y = 1 predictions. This will increase both true and false positives, so precision will decrease, not increase.
☑ The classifier is likely to now have higher recall.	✔	0.25	Lowering the threshold means more y = 1 predictions. This will increase the number of true positives and decrease the number of false negatives, so recall will increase.
☐ The classifier is likely to now have lower recall.	✔	0.25	Lowering the threshold means more y = 1 predictions. This will increase the number of true positives and decrease the number of false negatives, so recall will increase, not decrease.
Total		1.00 / 1.00	

Question 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.

Your Answer		Score	Explanation
☑ If you always predict non-spam (output $y = 0$), your classifier	✔	0.25	Since every prediction is y = 0, there will be no true positives, so recall is 0%.

will have a recall of 0%.			
☑ A good classifier should have both a high precision and high recall on the cross validation set.	✔	0.25	For data with skewed classes like these spam data, we want to achieve a high F_1 score, which requires high precision and high recall.
☑ If you always predict spam (output $y = 1$), your classifier will have a recall of 100% and precision of 1%.	✔	0.25	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 non-spam (output $y = 0$), your classifier will have 99% accuracy on the training set, but it will do much worse on the cross validation set because it has overfit the training data.	✔	0.25	The classifier achieves 99% accuracy because of the skewed classes in the data, not because it is overfitting the training set. Thus, it is likely to perform just as well on the cross validation set.
Total		1.00 / 1.00	

Question 5

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

Your Answer		Score	Explanation
☐ After training a logistic regression classifier, you **must** use 0.5 as your threshold for predicting whether an example is positive or negative.	✔	0.20	You can and should adjust the threshold in logistic regression using cross validation data.
☐ It is a good idea to	✔	0.20	You cannot know whether a huge dataset will be

spend a lot of time collecting a **large** amount of data before building your first version of a learning algorithm.			important until you have built a first version and find that the algorithm has high variance.
☑ Using a **very large** training set makes it unlikely for model to overfit the training data.	✔	0.20	A sufficiently large training set will not be overfit, as the model cannot overfit some of the examples without doing poorly on the others.
☑ 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.	✔	0.20	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.
☐ If your model is underfitting the training set, then obtaining more data is likely to help.	✔	0.20	If the model is underfitting the training data, it has not captured the information in the examples you already have. Adding further examples will not help any more.
Total		1.00 / 1.00	