Congratulations! You passed!

Grade received 80% To pass 80% or higher

Go to next item

Machine Learning System Design

Latest Submission Grade 80%

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: **1 / 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 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.

0.85

Correct
There are 85 true positives and 15 false negatives, so recall is 85 / (85 + 15) = 0.85.

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?

- [x] 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.

- [ ] The classes are not too skewed.
- [x] 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.

- [ ] When we are willing to include high order polynomial features of x (such as x_1^2, x_2^2, x_1x_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.

- [x] The classifier is likely to now have higher precision.

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 now have higher recall.
- [ ] The classifier is likely to have unchanged precision and recall, and thus the same F_1 score.
- [ ] The classifier is likely to have unchanged precision and recall, but higher 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. **0 / 1 point**

- [x] 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.

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

You didn't select all the correct answers

5. Which of the following statements are true? Check all that apply. **1 / 1 point**

- [x] The "error analysis" process of manually examining the examples which your algorithm got wrong can help suggest what are good steps to take (e.g., developing new features) to improve your algorithm's performance.

Correct
This process of error analysis is crucial in developing high performance learning systems, as the space of possible improvements to your system is very large, and it gives you direction about what to work on next.

- [ ] 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.
- [ ] 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.
- [x] 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.