Machine Learning System Design

LATEST SUBMISSION GRADE

100%

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 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

✓ **Correct**

The classifier correctly predicted the true positives and the true negatives = 85 + 10, so the accuracy is 95/1000 = 0.095

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] 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).

✓ **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.

- [ ] We train a model that does not use regularization.
- [ ] We train a learning algorithm with a small number of parameters (that is thus unlikely to overfit).
- [x] 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.

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 0.9. Which of thefollowing 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, but higher accuracy.
- [ ] The classifier is likely to have unchanged precision and recall, and thus the same F_1 score.

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**

- [x] If you always predict non-spam (output $y = 0$), your classifier will have 99% accuracy on the training set, and it will likely perform similarly on the cross validation set.

  Correct
  The classifier achieves 99% accuracy on the training set because of how skewed the classes are. We can expect that the cross-validation set will be skewed in the same fashion, so the classifier will have approximately the same accuracy.

- [ ] 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.
- [x] A good classifier should have both a high precision and high recall on the cross validation set.

  Correct
  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.

- [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.

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

- [ ] 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.
- [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.

- [ ] 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.
- [x] 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.