HMC CS 158
Quiz 6: Machine Learning System Design

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 $n = 1000$ 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
	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)

(a) What is the classifier's accuracy (as a value from 0 to 1)?

(b) What is the classifier's precision (as a value from 0 to 1)?

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 which of the following condition hold true?

(a) A human expert on the application can confidently predict y when given only the features x (or more generally, if we have some way to be confident that x contains information to predict y accurately.

(b) The classes are not too skewed.

(c) We train a model that does not use regularization.

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

(e) We train a learning algroithm with a large number of parameters (that is able to learn / represent fairly complex functions.

This quiz is adapted from course material by Andrew Ng (Stanford).

3. Suppose you have trained a logistic regression classifier which is outputting $h_{\boldsymbol{\theta}}(\boldsymbol{x})$. Currently, you predict 1 if $h_{\boldsymbol{\theta}}(\boldsymbol{x}) \geq$ threshold and predict 0 if $h_{\boldsymbol{\theta}}(\boldsymbol{x}) <$ threshold, where currently the threshold is set to 0.5. Suppose you **increase** the threshold to 0.9. Which of the following are true? Check all that apply.

   (a) The classifier is likely to now have higher precision.

   (b) The classifier is likely to now have higher recall.

   (c) The classifier is likely to have unchanged precision and recall and thus the same F_1 score.

   (d) The classifier is likely to have unchanged precision and recall but lower 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.

   (a) A good classifier should have both a high precision and high recall on the cross-validation set.

   (b) 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.

   (c) If you always predict non-spam (output $y = 0$), your classifier will have an accuracy of 99%.

   (d) 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.

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

   (a) If you model is underfitting the training set, then obtaining more data is likely to help.

   (b) After training a logistic regression classifier, you **must** use 0.5 as your threshold for predicting whether an example is positive or negative.

   (c) 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.

   (d) 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.

   (e) 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.