Week 6-2: Machine Learning System Design

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:

	Actual Class: 1	Actual Class: 0
Predicted Class: 1	85	890
Predicted Class: 0	15	10

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

Precision= $\frac{a}{a+b}$ where a is true positives, b is false positives, c is false negatives, d is true negatives.
$\frac{85}{85+890} = \frac{85}{975} = \boxed{0.087}$

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?

(a) 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).

(b) 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).

3. Suppose you have trained a logistic regression classifier which is outputting $h_\theta(x)$. Currently, you predict 1 if $h_\theta(x) \geq$ threshold and predict 0 if $h_\theta(x) <$ 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.

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

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.

Assume the training data consists of 200 emails, where 198 are non-spam and 2 are spam.

Table 1: Always Predicts Non-Spam $y = 0$:

	Actual Class: 1	Actual Class: 0
Predicted Class: 1	0	0
Predicted Class: 0	2	198

Accuracy $= \frac{a+d}{n} = \frac{0+198}{200} = \frac{198}{200} = 99\%$
Precision $= \frac{a}{a+b} = \frac{0}{0+0} =$ Undefined
Recall $= \frac{a}{a+c} = \frac{0}{0+2} = \frac{0}{2} = 0\%$

Table 2: Always Predicts Spam $y = 1$:

	Actual Class: 1	Actual Class: 0
Predicted Class: 1	0	0
Predicted Class: 0	2	198

Accuracy $= \frac{2}{200} = 1\%$
Precision $= \frac{2}{200} = 1\%$
Recall $= \frac{2}{2} = 100\%$

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

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

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

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

(a) On skewed data (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.

(b) Using a **very large** training set makes it unlikely for the model to overfit the training data.