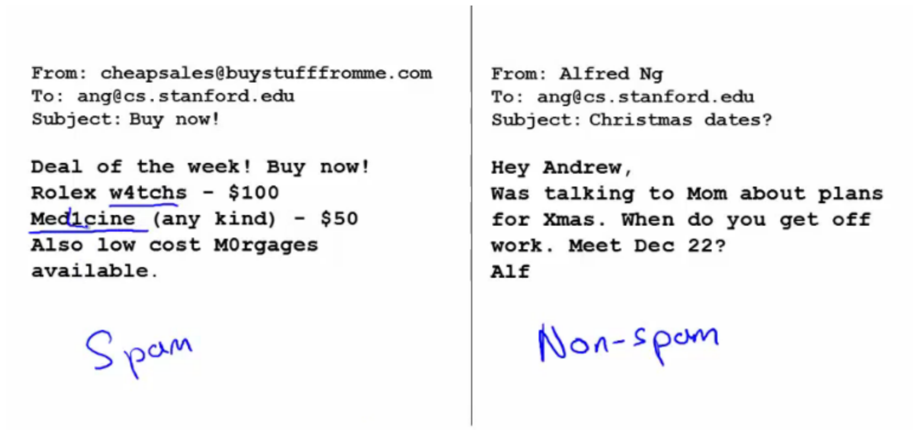
# Week 11

July 27, 2021

# Prioritizing what to work on - spam classification example

- Building a spam classifier.
- Misspelled word => Spam (1).
- Real content => Not spam (0).




## Select your own features

- Choose 100 words that indicate if an email is spam or not.
- Buy, discount, and deal are examples of spam.
- Andrew and now are examples of non-spam.
- All these words go into one long vector.
- If a matching word does not appear in the email, store 0 in the vector; otherwise, store 1.
- Check which word category has the most occurrences.

## How to improve system accuracy?

- Collect more data.
- Develop sophisticated features based on email routing data.
- Create a powerful algorithm for detecting misspellings.
- Plot learning curves to see whether extra data, features, and so on will help algorithmic optimization.

## Error analysis

- Examine the samples (in the cross validation set) on which your algorithm made errors manually.
- Try to figure out why.
- For example, you may find out that the most prevalent types of spam emails are pharmaceutical emails and phishing emails.
- What features would have helped classify them correctly?

## Error metrics for skewed analysis

- Suppose we're attempting to categorize cancer patients.
- We have 1% error. Looks good?
- But only 0.5% of people have cancer.
- Now, 1% error looks very bad!

### Precision and recall

| Classification | Guessed | Real |
|---|---|---|
| True positive | 1 | 1 |
| False positive | 1 | 0 |
| True negative | 0 | 0 |
| False negative | 0 | 1 |

- Precision: How often does our algorithm cause a false alarm?

$$\frac{true\ positives}{true\ positives\ +\ false\ positives}$$

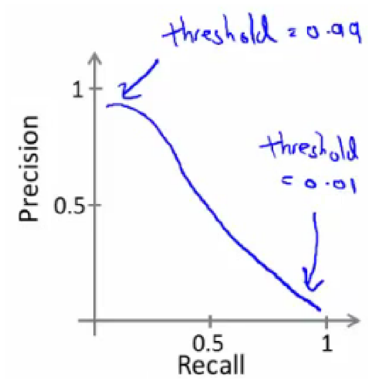
- Recall: How sensitive is our algorithm?

$$\frac{true\ positives}{true\ positives\ +\ false\ negative}$$

### Trading off precision and recall

- Trained a logistic regression classifier
  - Predict 1 if $h_\theta(x) >= 0.5$
  - Predict 0 if $h_\theta(x) < 0.5$

- We might change the prediction threshold such that we are more sure that a 1 is a true positive.
  - Predict 1 if $h_\theta(x) >= 0.8$
  - Predict 0 if $h_\theta(x) < 0.2$
- But classifier has lower recall - predict y = 1 for a smaller number of patients.




$F_{score}$ is calculated by averaging precision and recall and assigning a larger weight to the lower number.

$$F_{score} = 2\frac{PR}{P+R}$$

If you're attempting to establish the threshold automatically, one method is to test a variety of threshold values and assess them on your cross validation set. Then select the threshold that yields the highest $F_{score}$.