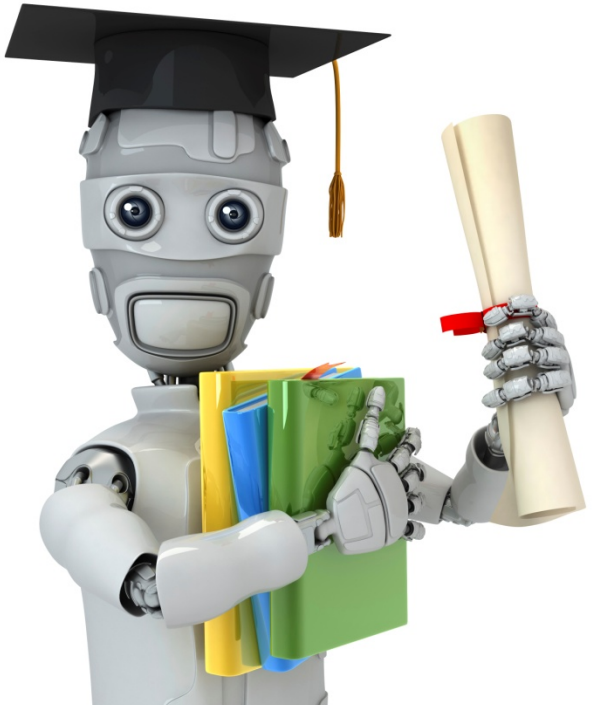

Machine Learning

Machine learning system design

Error analysis

Recommended approach

1 - Start with a simple algorithm that you can implement quickly. Implement it and test it on your cross-validation data.

2 - Plot learning curves to decide if more data, more features, etc. are likely to help. high bias? high variance?

3 - Error analysis: Manually examine the examples (in cross validation set) that your algorithm made errors on. See if you spot any systematic trend in what type of examples it is making errors on.

examinar manualmente os erros

Error Analysis

$m_{CV} =$ 500 examples in cross validation set

Algorithm misclassifies 100 emails.

Manually examine the 100 errors, and categorize them based on:

(i) What type of email it is — pharma, replica, steal passwords, ...

(ii) What cues (features) you think would have helped the algorithm classify them correctly.

Pharma: 12

Replica/fake: 4

Steal passwords: 53

Other: 31

Deliberate misspellings: 5
(m0rgage, med1cine, etc.)

Unusual email routing: 16

Unusual (spamming) punctuation: 32

The importance of numerical evaluation

- Should discount/discounts/discounted/discounting be treated as the same word?

  Can use "stemming" software (E.g. "Porter stemmer") soluçao? este software

  universe/university.

  Error analysis may not be helpful for deciding if this is likely to improve performance. Only solution is to try it and see if it works.

  Need numerical evaluation (e.g., cross validation error) of algorithm's performance with and without stemming.

  Without stemming: 5% error With stemming: 3% error

  Distinguish upper vs. lower case (Mom/mom): 3.2%