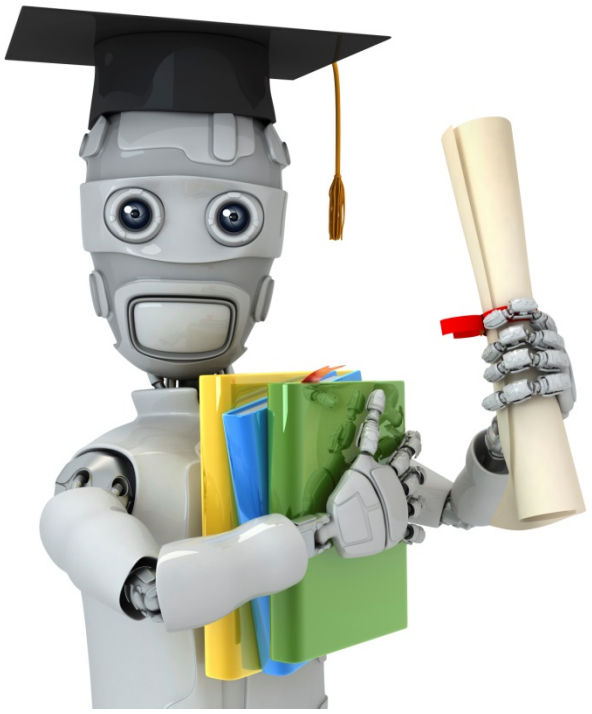


Machine Learning

Machine learning system design

Prioritizing what to work on: Spam classification example

确定研究的优先级：哪些特征去用到

Building a spam classifier

From: cheapsales@buystufffromme.com
To: ang@cs.stanford.edu
Subject: Buy now!

Deal of the week! Buy now!
Rolex w4tchs - $100
Medicine (any kind) - $50
Also low cost M0rgages
available.

From: Alfred Ng
To: ang@cs.stanford.edu
Subject: Christmas dates?

Hey Andrew,
Was talking to Mom about plans
for Xmas. When do you get off
work. Meet Dec 22?
Alf

Building a spam classifier

Supervised learning. x = features of email. y = spam (1) or not spam (0).

Features x: Choose 100 words indicative of spam/not spam.

E.g. deal, buy, discont, andrew, now, ...

$$x = \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ \vdots \\ 1 \\ \vdots \end{bmatrix} \begin{matrix} \text{andrew} \\ \text{buy} \\ \text{deal} \\ \text{discont} \\ \vdots \\ \text{now} \\ \vdots \end{matrix} \quad x \in \mathbb{R}^{100}$$

$$x_j = \begin{cases} 1 & \text{if word } j \text{ appears in email} \\ 0 & \text{otherwise.} \end{cases}$$

From: cheapsales@buystufffromme.com
To: ang@cs.stanford.edu
Subject: Buy now!

Deal of the week! Buy now!

Note: In practice, take most frequently occurring n words (10,000 to 50,000) in training set, rather than manually pick 100 words.

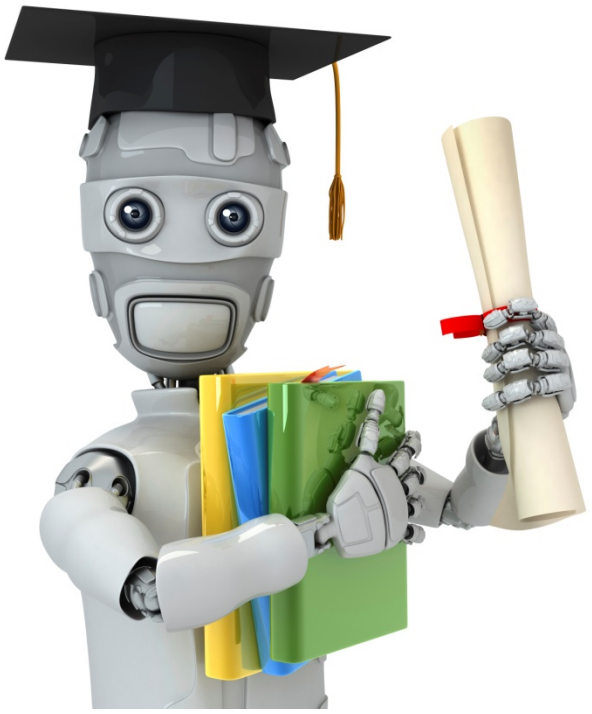
Andrew Ng

Building a spam classifier

How to spend your time to make it have low error?

- Collect lots of data
  E.g. “honeypot” project.
- Develop sophisticated features based on email routing information (from email header).
- Develop sophisticated features for message body, e.g. should “discount” and “discounts” be treated as the same word? How about “deal” and “Dealer”? Features about punctuation?
- Develop sophisticated algorithm to detect misspellings (e.g. m0rtgage, med1cine, w4tches.)

Andrew Ng



Machine Learning

Machine learning system design

Error analysis

误差分析

Recommended approach

- Start with a simple algorithm that you can implement quickly. Implement it and test it on your cross-validation data.
- Plot learning curves to decide if more data, more features, etc. are likely to help.
- 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.

Andrew Ng

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

(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

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

A: Can use "stemming" software (E.g. "Porter stemmer")

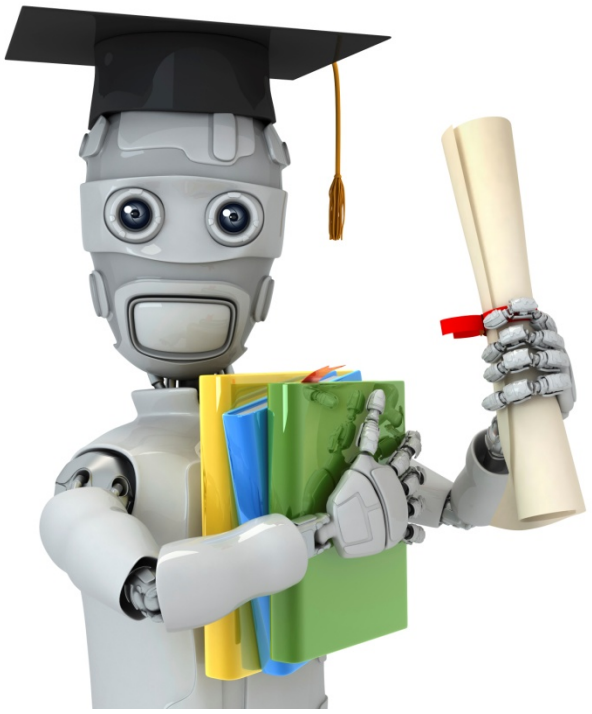
universe/university.

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%



Machine Learning

Machine learning system design

Error metrics for skewed classes

倾斜类问题(有时对算法造成影响)

Cancer classification example

Train logistic regression model $h_\theta(x)$. ($y = 1$ if cancer, $y = 0$ otherwise)

1. 机器学习：测试集得出1%的错误率

Find that you got 1% error on test set.
(99% correct diagnoses)

正例、负例的比率非常接近于一个极端情况
比例：正例也负例很多） → 偏斜类
$y=1$ $y=0$

得癌症的患者只有0.5%

Only 0.50% of patients have cancer. → skewed classes.

```
function y = predictCancer(x)
  y = 0; %ignore x!
return
```

0.5% error

2. 非机器学习：0.5%错误率
忽略了输入值x
始终使y=0（预测没有人得癌症）

→ 99.2% accuracy (0.8% error)
→ 99.5% accuracy (0.5% error) (假)

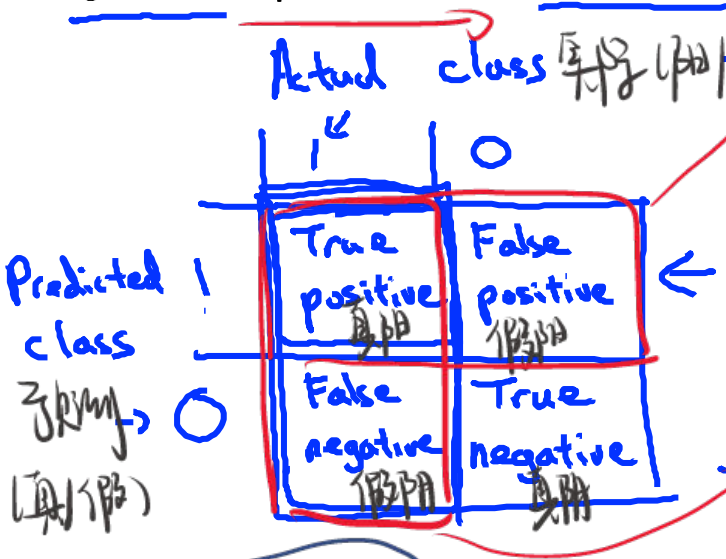
分类误差/分类精确度作为评估度量所出现的问题

也就有个偏斜类，用分类精确度并不能很好的度量算法

⇒ 不同误差度量值：Precision/Recall

Precision/Recall

$y = 1$ in presence of rare class that we want to detect



Precision (Of all patients where we predicted $y = 1$, what fraction actually has cancer?)

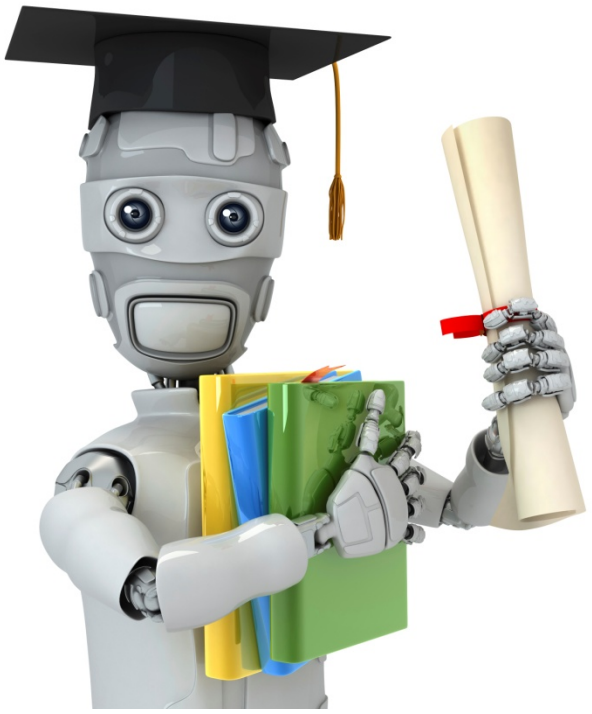
$$\frac{\text{True positives}}{\text{\#predicted positive}} = \frac{\text{True positives}}{\text{True pos + False pos}}$$

Recall (Of all patients that actually have cancer, what fraction did we correctly detect as having cancer?)

$$\frac{\text{True positives}}{\text{\#actual positives}} = \frac{\text{True positives}}{\text{True pos + False neg}}$$

$$\frac{TP}{TP + FN}$$

Andrew Ng



Machine Learning

Machine learning system design

Trading off precision and recall

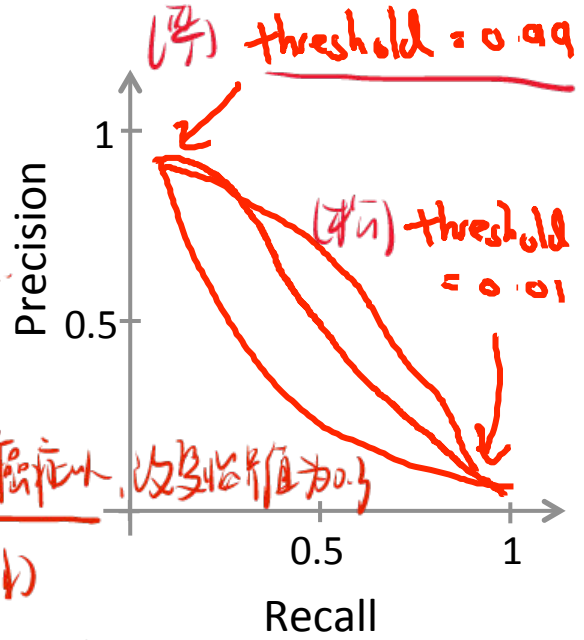
平衡查准率及召回率

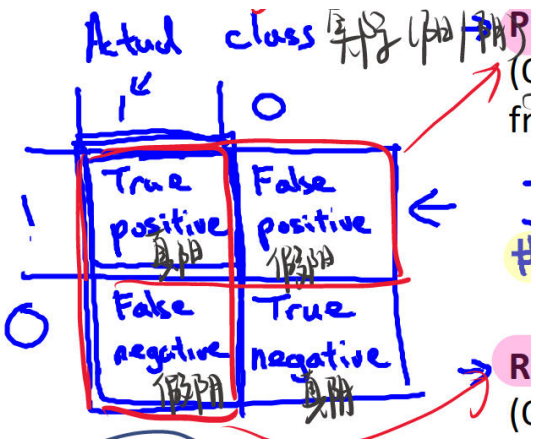
Trading off precision and recall

- Logistic regression: $0 \le h_\theta(x) \le 1$
- Predict 1 if $h_\theta(x) \ge 0.5$
- Predict 0 if $h_\theta(x) < 0.5$
- Suppose we want to predict $y = 1$ (cancer) only if very confident.
  - Higher precision, lower recall
- Suppose we want to avoid missing too many cases of cancer (avoid false negatives).
  - Higher recall, lower precision.
- More generally: Predict 1 if $h_\theta(x) \ge$ threshold.

precision = true positives / no. of predicted positive

recall = true positives / no. of actual positive





F_1 Score (F score)

How to compare precision/recall numbers?

	Precision(P)	Recall (R)	Average	F_1 Score
Algorithm 1	0.5	0.4	0.45	0.444
Algorithm 2	0.7	0.1	0.4	0.175
Algorithm 3	0.02	1.0	0.51	0.0392

Predict y=1 all the time

Average: $\frac{P+R}{2}$

F_1 Score: $2\frac{PR}{P+R}$

P=0 or R=0 $\Rightarrow$ F-score = 0

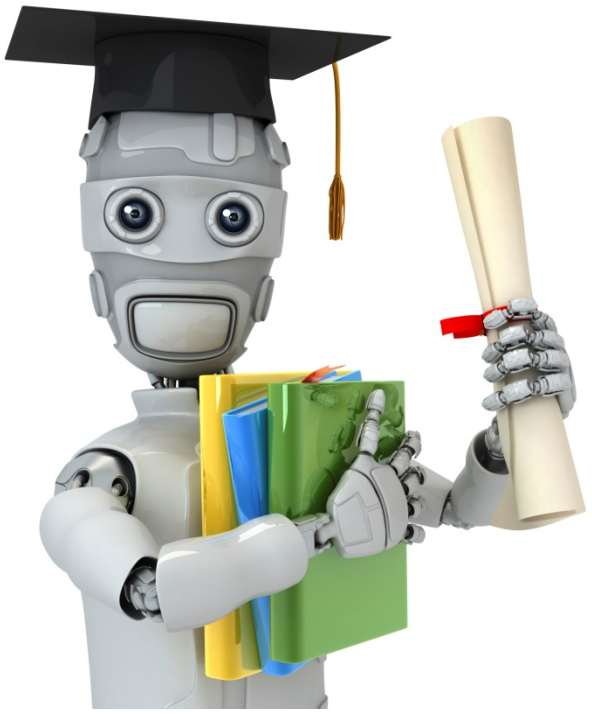
P=1 and R=1 $\Rightarrow$ F-score = 1

Precision: TP / predicted ↓ Recall: TP / actual ↑

Predict y=0 all the time

Precision: TP↓ / predicted ↑ Recall: TP↓ / actual ↓

Andrew Ng



Machine Learning

Machine learning system design

Data for machine learning

用于机器学习的数据

Designing a high accuracy learning system

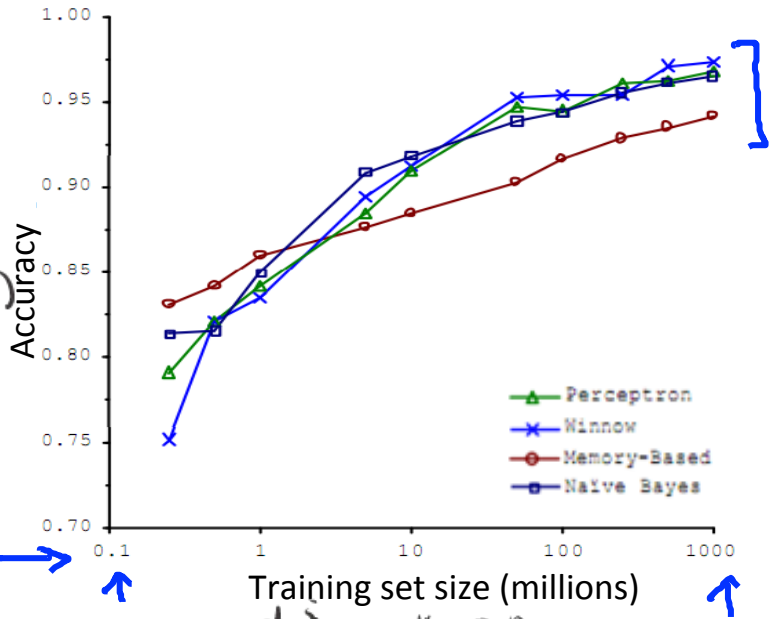
E.g. Classify between confusable words.

{to, two, too}, {then, than}

For breakfast I ate _____ eggs.

Algorithms

- Perceptron (Logistic regression)
- Winnow
- Memory-based
- Naïve Bayes



"It's not who has the best algorithm that wins.
It's who has the most data."

[Banko and Brill, 2001]

Large data rationale

Assume feature $x \in \mathbb{R}^{n+1}$ has sufficient information to predict y accurately.

Example: For breakfast I ate two eggs.

Counterexample: Predict housing price from only size (feet^2) and no other features.

Useful test: Given the input x, can a human expert confidently predict y?

Large data rationale

Use a learning algorithm with many parameters (e.g. logistic regression/linear regression with many features; neural network with many hidden units).

low bias algorithms

$J_{train}(\theta)$ will be small

Use a very large training set (unlikely to overfit)

low variance

$J_{train}(\theta) \approx J_{test}(\theta)$

$J_{test}(\theta)$ will be small