

Machine Learning

Machine learning system design

Prioritizing what to
work on: Spam
classification example

确定执行的优先级 = 邮件分类队列

Building a spam classifier 垃圾邮件分类器

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

故意推销单词

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

垃圾 Spam (1)

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

非垃圾 Non-spam (0)

Building a spam classifier

监督学习 (Supervised Learning)

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

Features x : Choose 100 words indicative of spam/not spam. 构建100个单词的列表

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

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

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

x 特征向量
 出现 $\rightarrow 1$
 未出现 $\rightarrow 0$
 对100个单词列表排序

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

Deal of the week! Buy now!

\rightarrow 编码成一个特征向量 x

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

训练集. 挑选出出现频率最高的 n 个单词 $\rightarrow n$ 个特征组成特征向量

$$n \in (10000, 50000)$$



用其表示邮件, 并对其进行分类

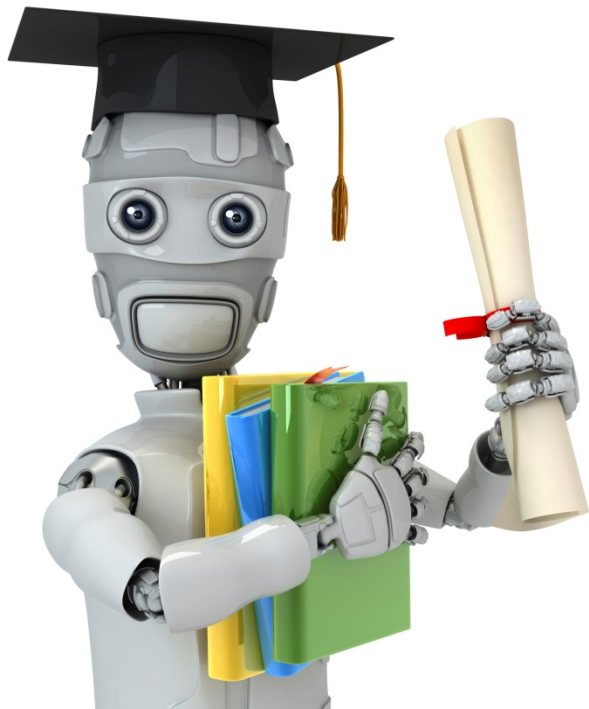
Building a spam classifier

短时间内, 垃圾邮件分类器具有高精度和低错误率

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

- Collect lots of data
搜集大量数据 E.g. "honeypot" project. 创建假邮箱地址, 并诱使恶意黑客向该地址发送邮件发送者以此得到大量垃圾邮件, 来训练分类算法
- Develop sophisticated features based on email routing information (from email header). 更多的特征 (特征) 如: 接收邮件的邮件包信其中通常出现在标题中
- 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? 更多的特征 (特征) 标点
- Develop sophisticated algorithm to detect misspellings (e.g. m0rtgage, med1cine, w4tches.) 识别拼写 识别故意拼写错误

发送者常去掩盖邮件来源
用假邮件标题
奇怪计算机服务器/路径发送邮件



Machine Learning

Machine learning system design

Error analysis

误差分析

Recommended approach

用简单算法快速实现，用交叉验证来测试数据

- 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. 画出训练曲线 (高偏差 高方差) 决定如何改进 (数据↑ 特征↓)
- 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.

误差分析: 手动检查交叉验证集中的例子

找出模型错误分类的例子

有共同特征/规律



设计新特征/想其它办法

编程中逐步迭代优化

应用新的特征指导决策

Error Analysis

$m_{CV} = 500$ examples in cross validation set 交叉验证集 500 样本

Algorithm misclassifies 100 emails. 错误分类 100

Manually examine the 100 errors, and categorize them based on 手动检查并分类

- 类型? → (i) What type of email it is pharma, replica, steal passwords, ...
特征/线索? → (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

统计错误邮件数量 分析错误原因

特征点式

Andrew Ng

The importance of numerical evaluation

保证机器学习算法有数值估计方法

例1:

Q: Should discount/discounts/discounted/discounting be treated as the same word? 是否将类似的不同单词

以此为例说明的 可返回数值评价指标

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

universe/university. 只有前两个英文字母不同 但是词义不同 / 但可能混淆为同义词

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. 误差分析无法决定此词干提取是否可行 最好的办法是尝试一下看是否有效

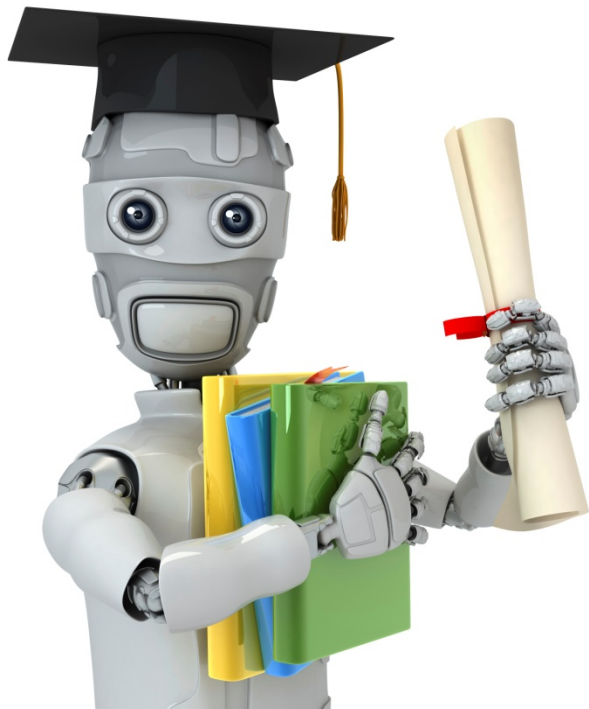
Without stemming: 5% error 不能 With stemming: 3% error 提升有40% (10-7=3)

例2: Distinguish upper vs. lower case (Mom/mom): 3.2% 提升有40% (10-6.8=3.2)

大小写是否看成个单词, 比较得分/不得分错误率

Andrew Ng

Tip: 一般用误差分析比较误差分析



Machine Learning

Machine learning system design

Error metrics for skewed classes

偏斜类问题(有时对算法会产生影响)

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)

Only 0.50% of patients have cancer.

得癌症的患者只有0.5%

→ skewed classes.

```
function y = predictCancer(x)
```

```
    → y = 0; %ignore x!
```

```
return
```

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)

例如: 负面比率非常高于-极端情况

例如: 正面比率非常低 → 偏斜类

y=1 y=0

0.5% error

→ 99.2% accuracy (0.8% error)

→ 99.5% accuracy (0.5% error) (假)

分类误差/分类精度作为评估度量所出现的问题

如果有偏斜类, 用分类精度并不能很好衡量算法

→ 不同误差度量值: precision/recall

Precision/Recall

癌症
稀少的情况用 y=1 表示

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

Actual class

Predicted class	1	0
1	True positive 真阳性	False positive 假阳性
0	False negative 假阴性	True negative 真阴性

Precision

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

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

Recall

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

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

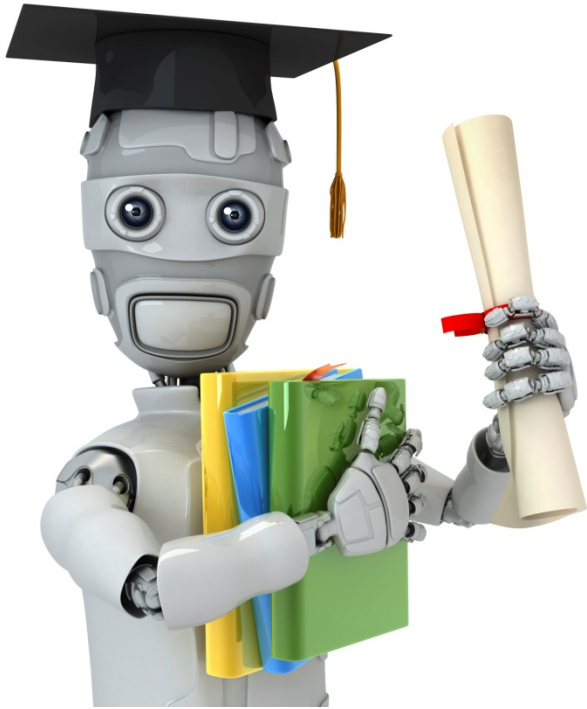
$$\frac{TP}{TP + FN}$$

Andrew Ng

y = 0

Recall = 0

如 P=3, FN=7, y=0 (癌症)
Precision = 0 / (0+7) = 0
Recall = 0 / (0+7) = 0
⇒ 该模型最糟糕, 啥都不对



Machine Learning

Machine learning system design

Trading off precision and recall

平衡查得率与召回率

$$h_{\theta}(x) = \frac{1}{1 + e^{-\theta^T x}}$$

Trading off precision and recall

\rightarrow precision = $\frac{\text{true positives}}{\text{no. of predicted positive}}$
 \rightarrow recall = $\frac{\text{true positives}}{\text{no. of actual positive}}$

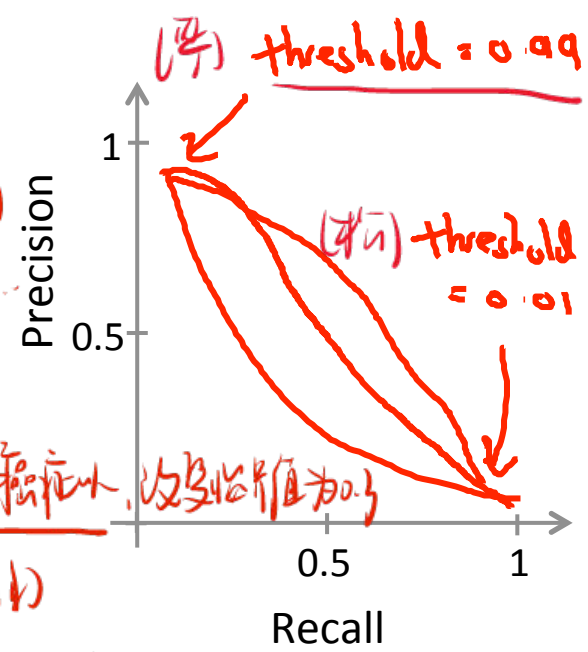
\rightarrow Logistic regression: $0 \leq h_{\theta}(x) \leq 1$
 Predict 1 if $h_{\theta}(x) \geq 0.5$ ~~0.7~~ ~~0.9~~ ~~0.3~~
 Predict 0 if $h_{\theta}(x) < 0.5$ ~~0.7~~ ~~0.9~~ ~~0.3~~

\rightarrow Suppose we want to predict $y = 1$ (cancer) only if very confident. *在非常确信时, 才判定为 $y=1$ (predict)*

\rightarrow Higher precision, lower recall *改变临界值为 0.1/0.9...*
TP_↓ predicted_↓ *TP_↓ actual_↓*

\rightarrow Suppose we want to avoid missing too many cases of cancer (avoid false negatives). *避免遗漏癌症病例, 改变临界值为 0.3*

\rightarrow Higher recall, lower precision.
TP_↑ actual_↑ *TP_↑ predicted_↑*



More generally: Predict 1 if $h_{\theta}(x) \geq \text{threshold}$. *能否自动选取临界值?*

Andrew Ng

	Actual class	
	1	0
Predicted class	1 True positive	0 False positive
	0 False negative	1 True negative

F₁ Score (F score)

How to compare precision/recall numbers?

平均那80分

有两个评价指标, 如何判断哪种更好?

	Precision (P)	Recall (R)	Average	F ₁ Score	
→ Algorithm 1	<u>0.5</u>	<u>0.4</u>	0.45	0.444	←
→ Algorithm 2	<u>0.7</u>	<u>0.1</u>	0.4	0.175	←
Algorithm 3	<u>0.02</u>	1.0	0.51	0.0392	←

平均: $\frac{P+R}{2}$
 但是这种极端情况
 但平均值最高 故平均值不可作为评价指标

Predict y=1 all the time

F₁ Score: $2 \frac{PR}{P+R}$

当 P=0 或 R=0, F₁ Score = 0
 当 P=1 且 R=1, F₁ Score = 1
 一般较大的 F₁ 值, 其 P, R 也较大

若 P, R 平均

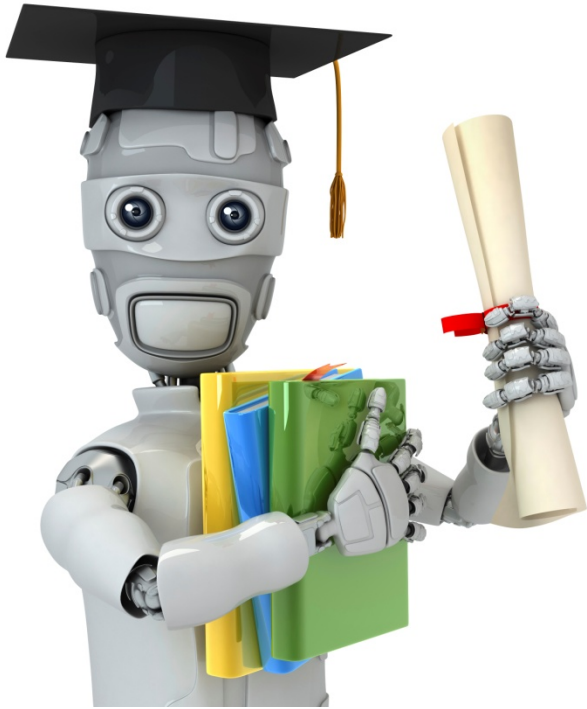
则 P, R 中较低者更重要

Precision = $\frac{TP}{\text{predicted}}$ | Recall = $\frac{TP}{\text{actual}}$

Andrew Ng

Predict y=0 all the time

Precision = $\frac{TP \downarrow}{\text{predicted} \downarrow}$ | Recall = $\frac{TP \downarrow}{\text{actual} \downarrow}$



Machine Learning

Machine learning system design

Data for machine learning

用于机器学习的训练数据

研究使用不同算法去分类, 并实际使用到不同训练数据上

Designing a high accuracy learning system

如何分类混淆词?

E.g. Classify between confusable words.

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

→ For breakfast I ate two eggs.

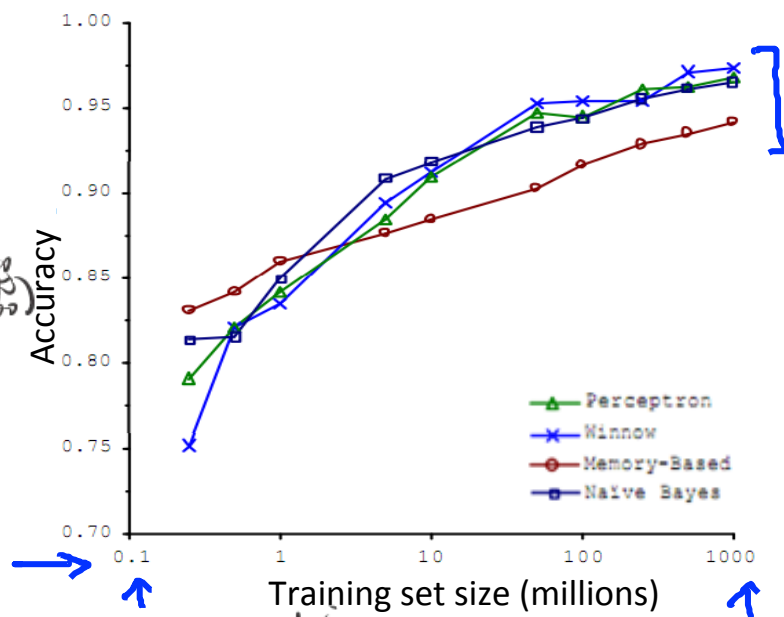
Algorithms

→ - Perceptron (Logistic regression)

→ - Winnow 类似于回归, 但方法不同

→ - Memory-based 基于内存

→ - Naïve Bayes 朴素贝叶斯



“It’s not who has the best algorithm that wins.”

It’s who has the most data.”

改变训练数据大小

将不同算法用于不同大小训练集中

大量数据重要性

[Banko and Brill, 2001]

Large data rationale 大量数据训练集

→ Assume feature $x \in \mathbb{R}^{n+1}$ has sufficient information to predict y accurately. 特征 x 包含足够信息 → 用于准确预测 y ↖

例 Example: For breakfast I ate ~~two~~ eggs. ↖

例 Counterexample: Predict housing price from only size (feet²) and no other features. 仅有一个特征: 房大小 → 很难准确预测其价格 ↖

给定一个特征 x 一个人是否能够准确预测出 y 值?
Useful test: Given the input x , can a human expert confidently predict y ? ↖

问题: 假设是包含足够信息 → 足够多的数据

Large data rationale

使用时需要大量参数 (特征多, 隐藏层多) 的算法
→ 可拟合很多复杂的函数

→ Use a learning algorithm with many parameters (e.g. logistic regression/linear regression with many features; neural network with many hidden units). 低偏差 low bias algorithms. ←

→ $J_{\text{train}}(\theta)$ will be small 训练误差会很低

Use a very large training set (unlikely to overfit)

low variance ← 低方差

→ $J_{\text{train}}(\theta) \approx J_{\text{test}}(\theta)$ 训练集跟测试集数据量更多, 不易过拟合

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