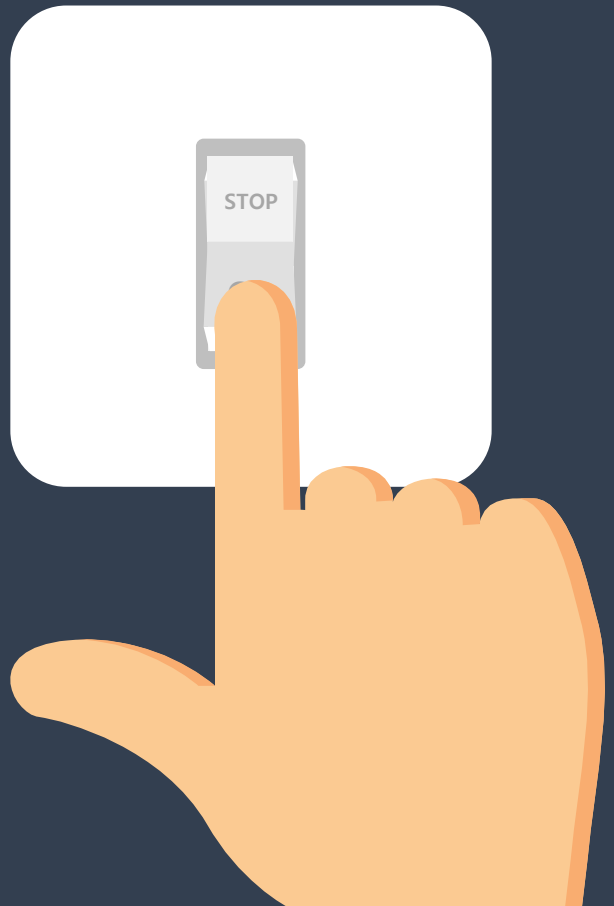
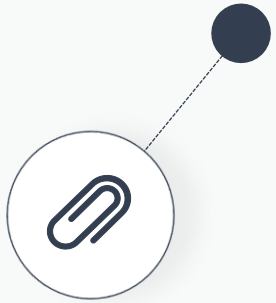
Machine Learning System Design

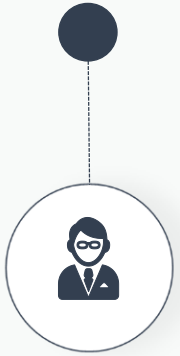


Machine Learning System Design

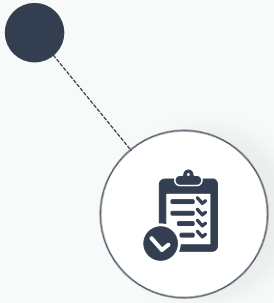
1770119 황서현



1. Building a Spam Classifier



2. Handling Skewed Data



3. Using Large Data Sets

building a Spam Classifier

스팸 분류 알고리즘 작성하기

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
Med1cine (any kind) - $50
Also low cost M0rgages
available.

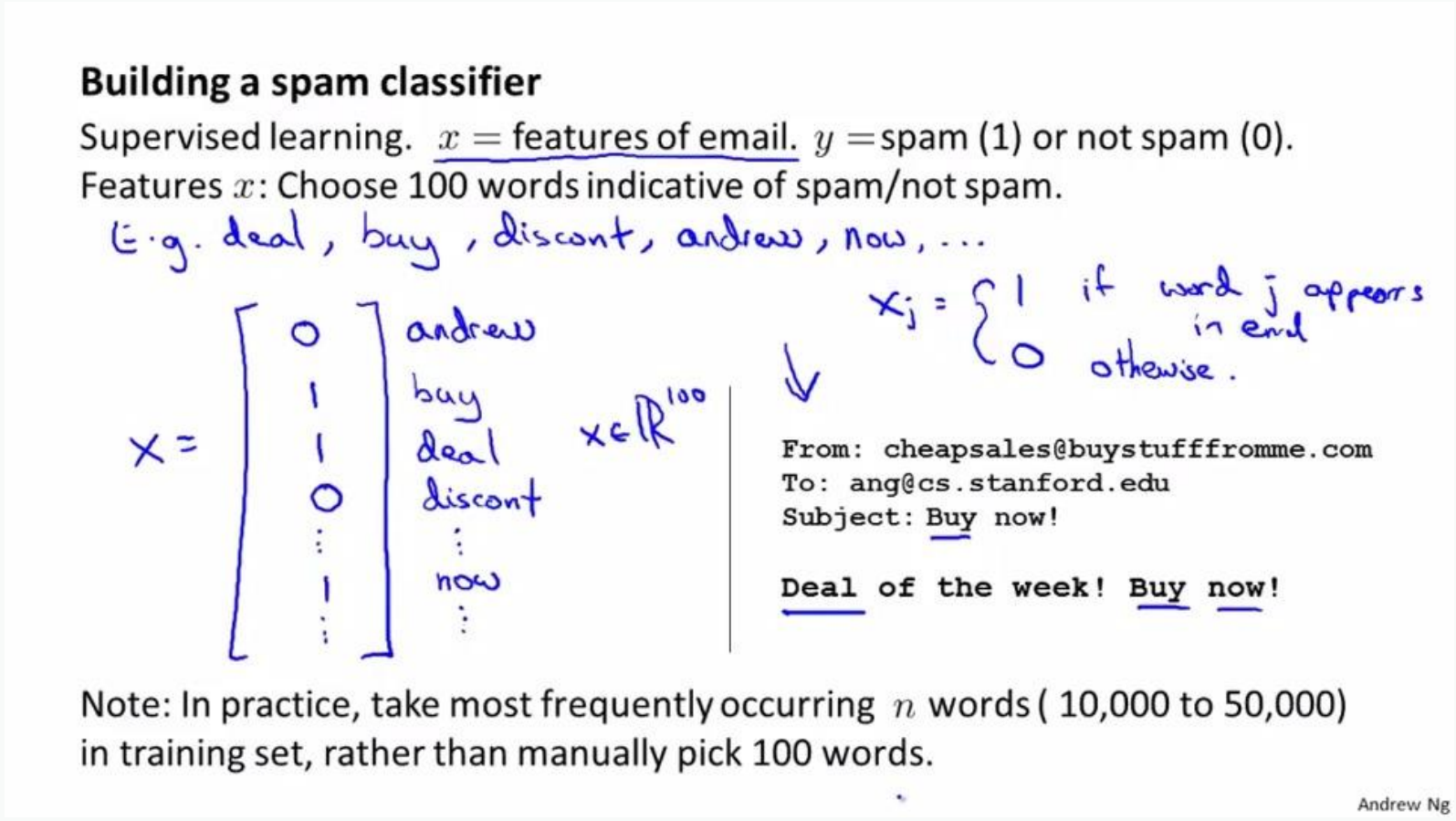
From: Alfred Ng
To: ang@cs.stanford.edu
Subject: Christmas dates?

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

Andrew Ng

building a Spam Classifier

스팸 분류 알고리즘 작성하기



1. Y값 지정하기

2. Feature 지정하기

building a Spam Classifier

스팸 분류 알고리즘 작성하기

<주의>

Data를 많이 모으기

Feature – 이메일 헤더

Feature – 비슷한 단어

오탈자 찾기

Return-Path : < example_from@dc.edu >
X-SpamCatcher- 점수 : 1 [X]
수신 : [136.167.40.119] (HELO dc.edu)
fe3.dc.edu (CommuniGate Pro SMTP 4.1.8)example_to@mail.dc.edu에
대한 ESMTP-TLS ID 61258719 ; 2004 년 8 월 23 일 월요일 11시 40 분 10 초 -0400
메시지 ID : < 4129F3CA.2020509@dc.edu >
날짜 : 2005 년 8 월 23 일 월요일 11시 40 분 36 초 -0400출처
: Taylor Evans < example_from@dc.edu >
사용자 에이전트 : Mozilla / 5.0 (Windows; U; Windows NT 5.1; en-US; rv : 1.0.1) Gecko / 20020823 Netscape / 7.0
X-Accept-Language : en-us, en
MIME 버전 : 1.0
받는 사람 : Jon Smith < example_to@mail.dc.edu >
제목 : 회의
Content-Type : text / plain; charset = us-ascii; format = flowed
콘텐츠 전송 인코딩 : 7 비트

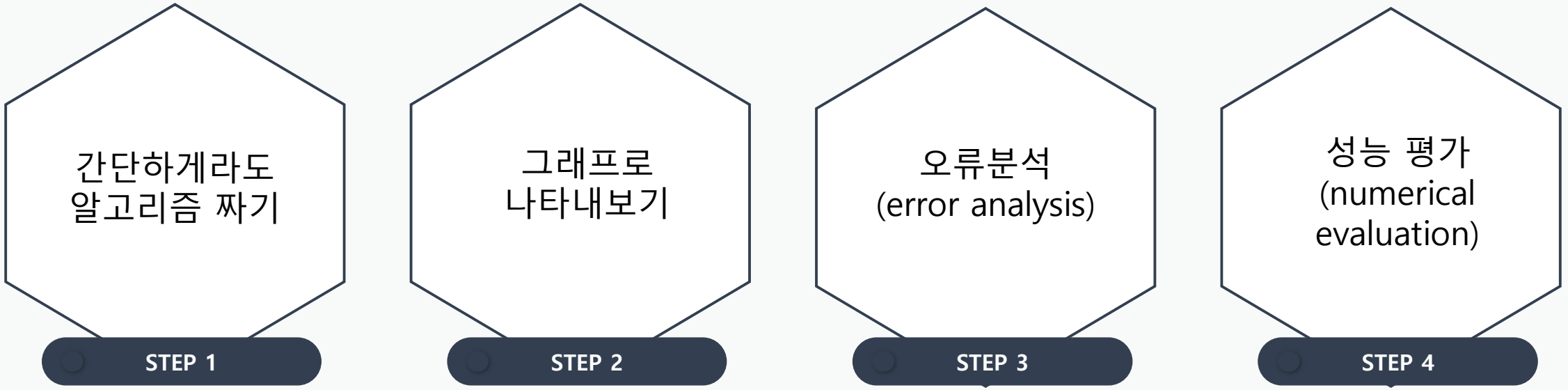
Discount
Discounts
Discounting
Medicine

W4tches

스팸 필터링에 안 걸리려고 고의로 넣은
오탈자도 구분할 수 있어야 함

building a Spam Classifier

스팸 분류 알고리즘 작성하기



<machine learning step>

building a Spam Classifier

스팸 분류 알고리즘 작성하기

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

Andrew Ng

SPAM 분류 예시

잘못 분류된
spam 메일들로
다시 분석해보기

Handling Skewed Data

스케위드 데이터 다루기

Skewed data
=skewed classes

치우친 데이터?

암 분류 예시

Y = 1 (암 0)
Y = 0 (암 X)

Function y = predictCancer(x)
Y = 0 ; % x가 무슨 값이든 무조건 y =0
return

error = 0 <-> accuracy = 100% ?????????

Handling Skewed Data

스케위드 데이터 다루기

Precision / recall

성능 평가
(numerical evaluation)

STEP 4

	Actual class		
Predicted Class		1 진짜 암	0 암 없음
	1 아마 암일거	Ture positive	False Positive
	0 아마 아닐거	False negative	True negative

Precision

$$\frac{True\ positive}{True\ positive + False\ positive}$$

Recall

$$\frac{True\ positive}{True\ positive + False\ negative}$$

Handling Skewed Data

스케위드 데이터 다루기

Trading off precision and recall

- 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

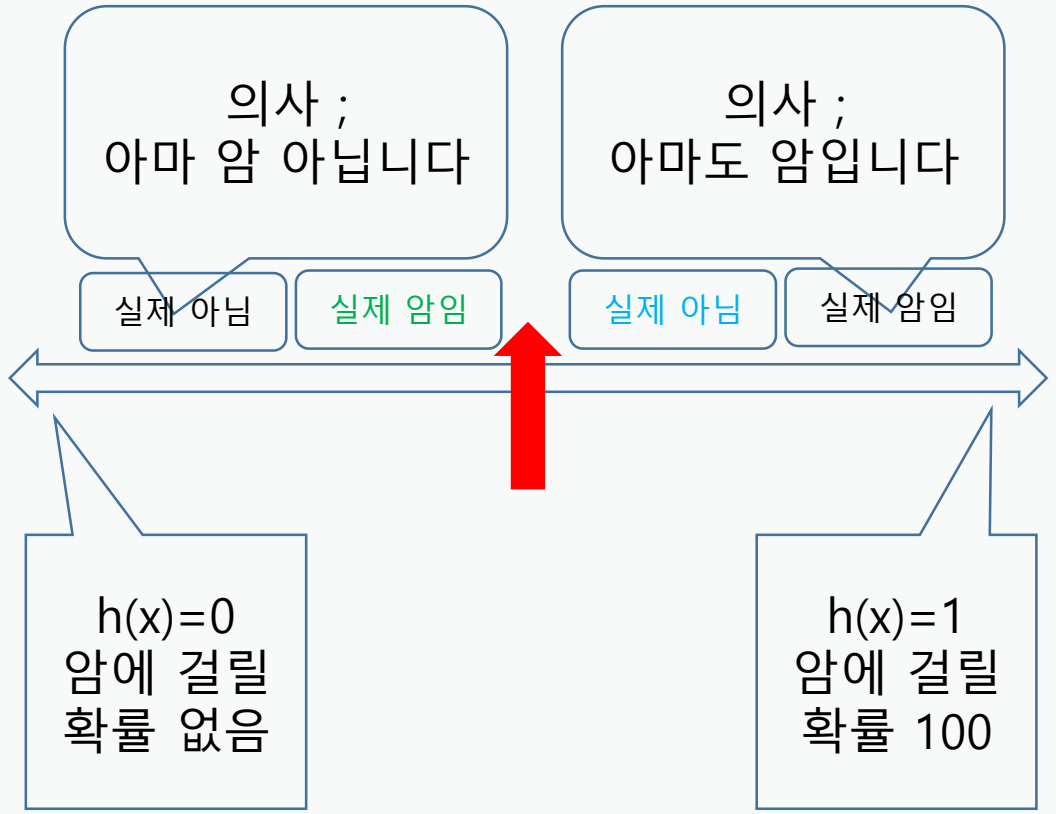
- Suppose we want to predict $y = 1$ (cancer) only if very confident.

  → 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 $h_\theta(x) \geq$ threshold.



Precision

$$\frac{True\ positive}{True\ positive + False\ positive}$$

Recall

$$\frac{True\ positive}{True\ positive + False\ negative}$$

Handling Skewed Data

스케위드 데이터 다루기

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 $= 1$

Andrew Ng

성능평가

Using Large Data Sets

엄청 큰 데이터 셋 이용하기

내가 feature x 를 받았을 때,
Y를 예측할 수 있는가?

Ex)
For breakfast I ate ____ eggs.
앞 뒤 단어 -> 빈칸 예측

Count ex)
only size feature(no other) ->
house price 예측

사이즈가 충분한가?

-parameter 가 많아야 한다

-training set이 많아야 한다

감사합니다

• • • • •