Copyright Notice

deeplearning.ai

Introduction to ML strategy

Why ML Strategy?

Motivating example

90%

Ideas:

- Collect more data ←
- Collect more diverse training set
- Train algorithm longer with gradient descent
- Try Adam instead of gradient descent
- Try bigger network
- Try smaller network
- Try dropout
- Add L_2 regularization
- Network architecture
  - Activation functions
  - # hidden units
  - …

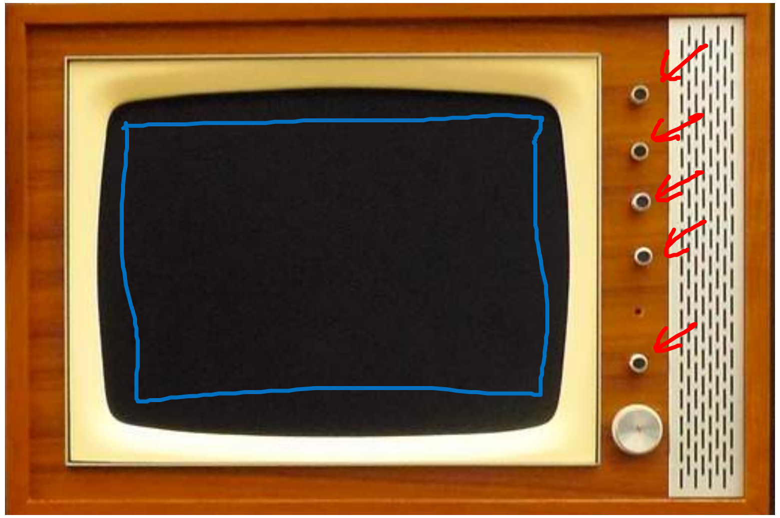
deeplearning.ai

Introduction to ML strategy

Orthogonalization

직교화

TV tuning example



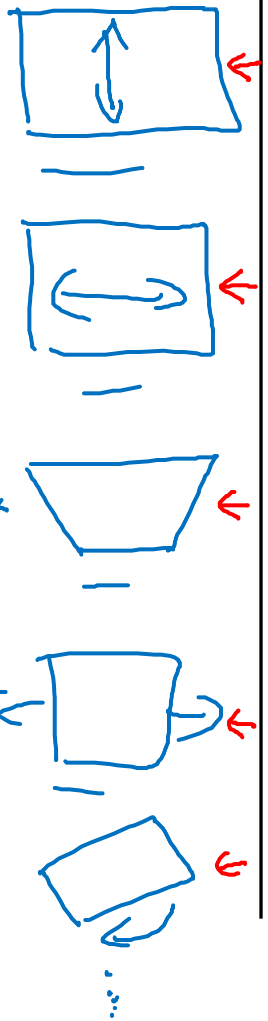
0.1 x

+0.3 x

-1.7 x

+0.8 x

+ ...



Orthogonalization

Car

→ Steering

→ Acceleration
Braking

→ 0.3 x angle − 0.8 speed

→ 2 x angle + 0.9 speed.

speed

angle

Chain of assumptions in ML

early stopping을 잘 사용하지 않는다. 왜냐하면 너무 일찍 학습을 멈추게 된다면 training set에서 잘 동작하지 않을 수 있는데, 동시에 early stopping은 dev set 성능 향상시키는데 사용되기 때문이다. 즉, 하나의 방법이 orthogonalization 하지 않고 동시에 2개 요소가 영향을 받게 만들었다.

→ Fit training set well on cost function

(≈ human-level performance)

width

- bigger network
- Adam
- ...

(교차성 깨진거면)

early stopping

→ Fit dev set well on cost function

height

- Regularization
- Bigger train set

→ Fit test set well on cost function

Bigger dev set

dev set에서 over fit 되어 있을 확률이 높기 때문에 이럴 때에는 더 큰 dev set을 사용해야 할 것이다.

→ Performs well in real world

(Happy cat pic app users.)

Change dev set or cost function

이 때에는 dev/test set 분포가 잘 반영되어 있지 않거나, cost function이 올바르게 계산되고 있지 않은 것으로 보아야 한다.

Andrew Ng

Setting up your goal

Single number evaluation metric

deeplearning.ai

정밀도란 모델이 True라고 분류한 것 중에서 실제 True인 것의 비율입니다.
즉, 아래와 같은 식으로 표현할 수 있습니다.

$$\text{precision} = \frac{TP}{TP + FP}$$

Recall (재현율)이란 실제 True인 것 중에서 모델이 True라고 예측한 것의 비율입니다.

$$(\text{Recall}) = \frac{TP}{TP + FN}$$

True Positive (TP) :
실제 True인 정답을 True라고 예측

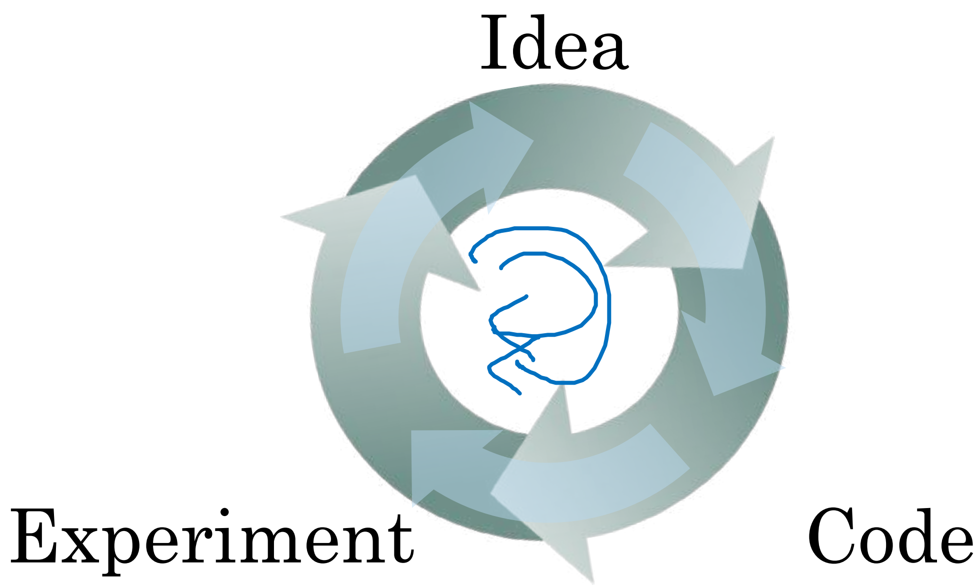
False Positive (FP) :
실제 False인 정답을 True라고 예측

False Negative (FN)
: 실제 True인 정답을 False라고 예측

True Negative (TN)
: 실제로 False인 정답을 False라고 예측

$$F_1 \text{ Score} = 2 \times \frac{\text{recall} \times \text{precision}}{\text{recall} + \text{precision}}$$

Using a single number evaluation metric



→ Of examples recognized as cat, what % actually are cats?

→ what % of actual cats are correctly recognized

Classifier	Precision	Recall
A	95%	90%
B	98%	85%

F_1 score = "Average" of P and R.

$\left(\frac{2}{\frac{1}{P} + \frac{1}{R}} . \right.$ "Harmonic mean")

Dev set + Single number evaluation metric

real

Speed up iterating

Another example

Algorithm	US	China	India	Other	Average
A	3%	7%	5%	9%	6%
B	5%	6%	5%	10%	6.5%
C	2%	3%	4%	5%	3.5%
D	5%	8%	7%	2%	5.25%
E	4%	5%	2%	4%	3.75%
F	7%	11%	8%	12%	9.5%

각각 대륙별(미국. 중국. 인도 그리고 나머지)로 다른 값의 에러를 가질 때,
A, B ... F 중 무엇이 나은지 결정하기 어렵다.
그러면 추천하는 방법은 대륙별 에러를 평균으로 계산하는 것이다.

deeplearning.ai

Setting up your goal

Satisficing and optimizing metrics

Another cat classification example

<정확도와 실행시간을 결합시켜 종합 평가 수치 만들기>

optimizing　　　　　　Satisficing

Classifier	Accuracy	Running time
A	90%	80ms
B	92%	95ms
C	95%	1,500ms

cost = accuracy − 0.5 × running Time

(정확도, 실행시간 두 가지를 선형 조합을 사용해서 결합)

maximize accuracy

<고려하고 있는 실행시간의 한계치를 만족하면서 정확도를 최대한으로 끌어올릴 수 있는 Classifier>

subject to running Time ≤ 100 ms.

ex) 실행시간 한계치는 100ms 이내로 설정하고, 이 수치 만족하는 분류기 중에서 가장 정확도가 높은 것을 선택

정확도 : optimizing metric　실행시간 : satisficing metric

N metrics : 1 optimizing

N−1 satisficing

일반적으로 N개의 고려해야 되는 metric가 있다면, 1개는 optimizing metric으로 선택하고, 나머지 N−1개를 satisficing 으로 선택하는게 합리적이다.

Wakewords / Trigger words

Alexa, OK Google, Hey Siri, nihaobaidu

你好百度

accuracy.

#false positive

maximize accuracy.

s.t. ≤ 1 false positive every 24 hours.

Andrew Ng

deeplearning.ai

Setting up your goal

Train/dev/test distributions

Cat classification dev/test sets

└ development set, hold out cross validation set

Regions:

- US
- UK
- Other Europe
- South America
- India
- China
- Other Asia
- Australia

Dev

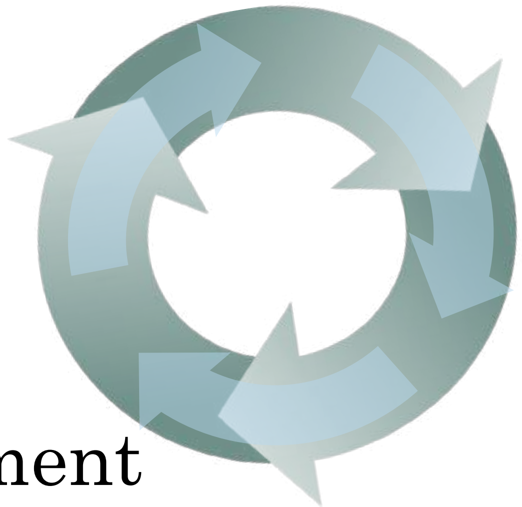
Test

이때! dev set과 test set는 다른 분포를 갖게 하는건 좋지 않다.

⇒ dev set 과 test set을 선택할때 같은 분포를 가지는 데이터에서 선택하는 것을 권장

Randomly shuffle into dev/test

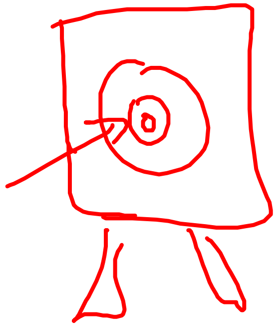
dev set + metric



True story (details changed)

Optimizing on dev set on loan approvals for medium income zip codes

$x \longrightarrow y$ (repay loan?)



Tested on low income zip codes

~3 month

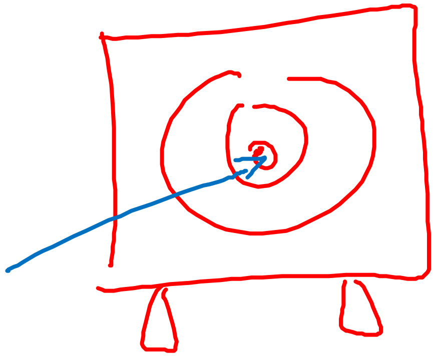
Andrew Ng

Guideline

Same distribution

Choose a dev set and test set to reflect data you expect to get in the future and consider important to do well on.

training



dev
metric

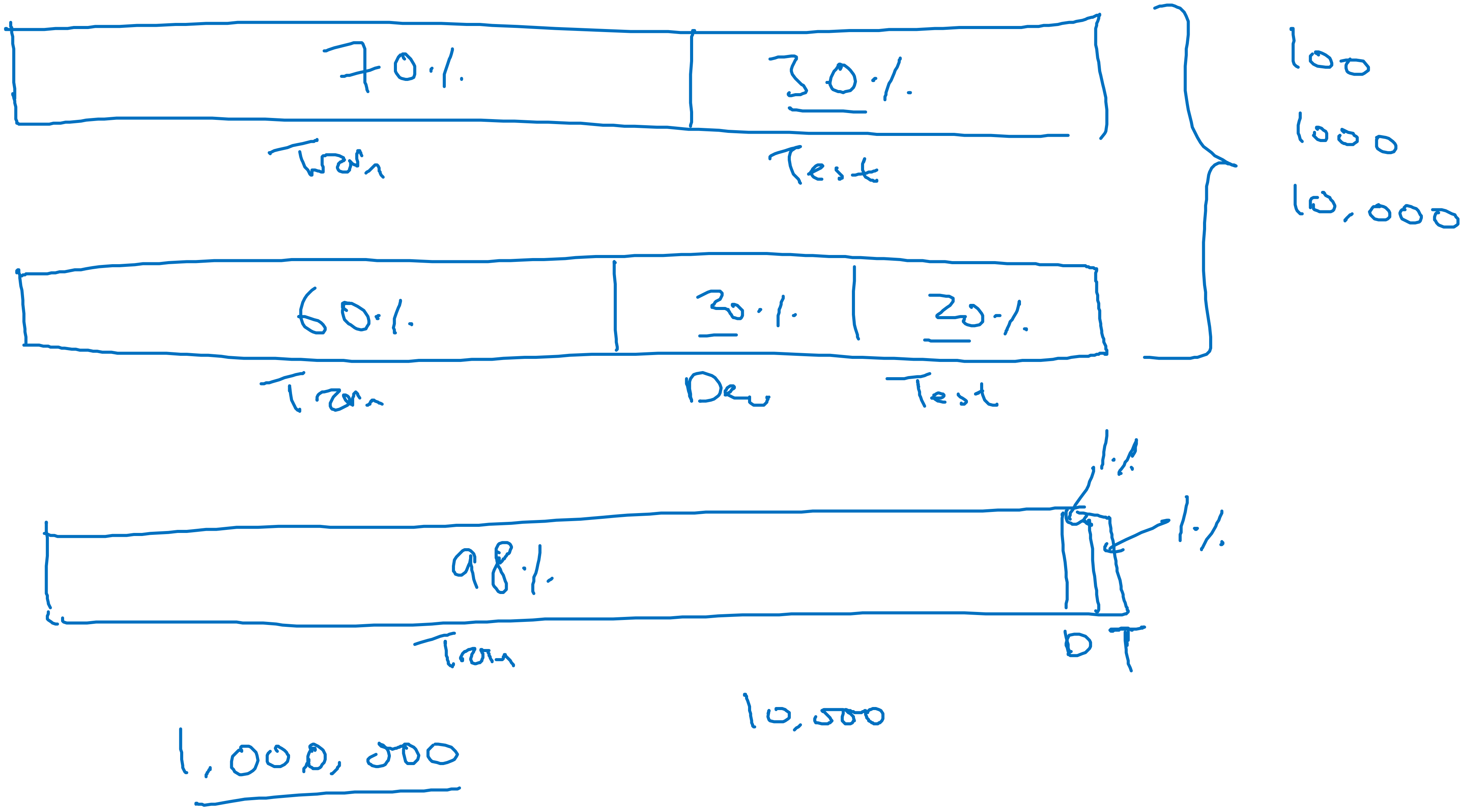
test

deeplearning.ai

Setting up your goal

Size of dev and test sets

Old way of splitting data



Size of dev set

A B

Set your dev set to be big enough to detect differences in algorithm/models you're trying out.

100 : small

1%

1,000

10,000

100,000

A 97% → B 97.1 %

0.1%

0.01%

0.001%

Online advertising

Andrew Ng

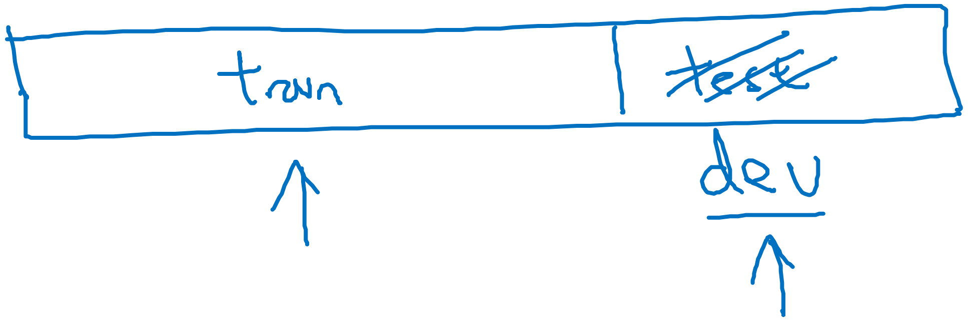
Size of test set

- Set your test set to be big enough to give high confidence in the overall performance of your system.

10,000

100,000

Train + Dev.   No test set.



deeplearning.ai

Setting up your goal

When to change dev/test sets and metrics

Cat dataset examples

Metric + Dev : Prefer A
You/users : Prefer B.

→ Metric: classification error

Algorithm A: 3% error → pornographic

✓ Algorithm B: 5% error

classification error:

$$\text{Error}: \frac{1}{\underbrace{\sum_i w^{(i)}}_{\text{normalization 위해 사용}}} \cancel{\frac{1}{m_{dev}}} \sum_{i=1}^{m_{dev}} \underbrace{w^{(i)}}_{\text{가중치}} \, I\{\underbrace{y_{pred}^{(i)} \neq y^{(i)}}\}$$

↑ predicted value (0/1)

$$\rightarrow w^{(i)} = \begin{cases} 1 & \text{if } x^{(i)} \text{ is non-porn} \\ 10 & \text{if } x^{(i)} \text{ is porn} \end{cases}$$

Andrew Ng

Orthogonalization for cat pictures: anti-porn

(직교화)

orthogonalization 적용 : target을 plan 하는 것과 tuning 단계는 분리해서 생각

1. So far we've only discussed how to define a metric to evaluate classifiers. ← Place target

2. Worry separately about how to do well on this metric.

Aim/shoot at target

평가지표 바꾸기

: 원하지 않는 결과에 가중치를 크게 줘서 손실값을 나쁘게 만든다.

$$J = \frac{1}{\cancel{m}} \sum_{i=1}^{m} w^{(i)} \mathcal{L}(\hat{y}^{(i)}, y^{(i)})$$

$\sum w^{(i)}$

(가중치 부여)

Andrew Ng

Another example

Algorithm A: 3% error

Algorithm B: 5% error

Dev/test

User images

If doing well on your metric + dev/test set does not correspond to doing well on your application, change your metric and/or dev/test set.

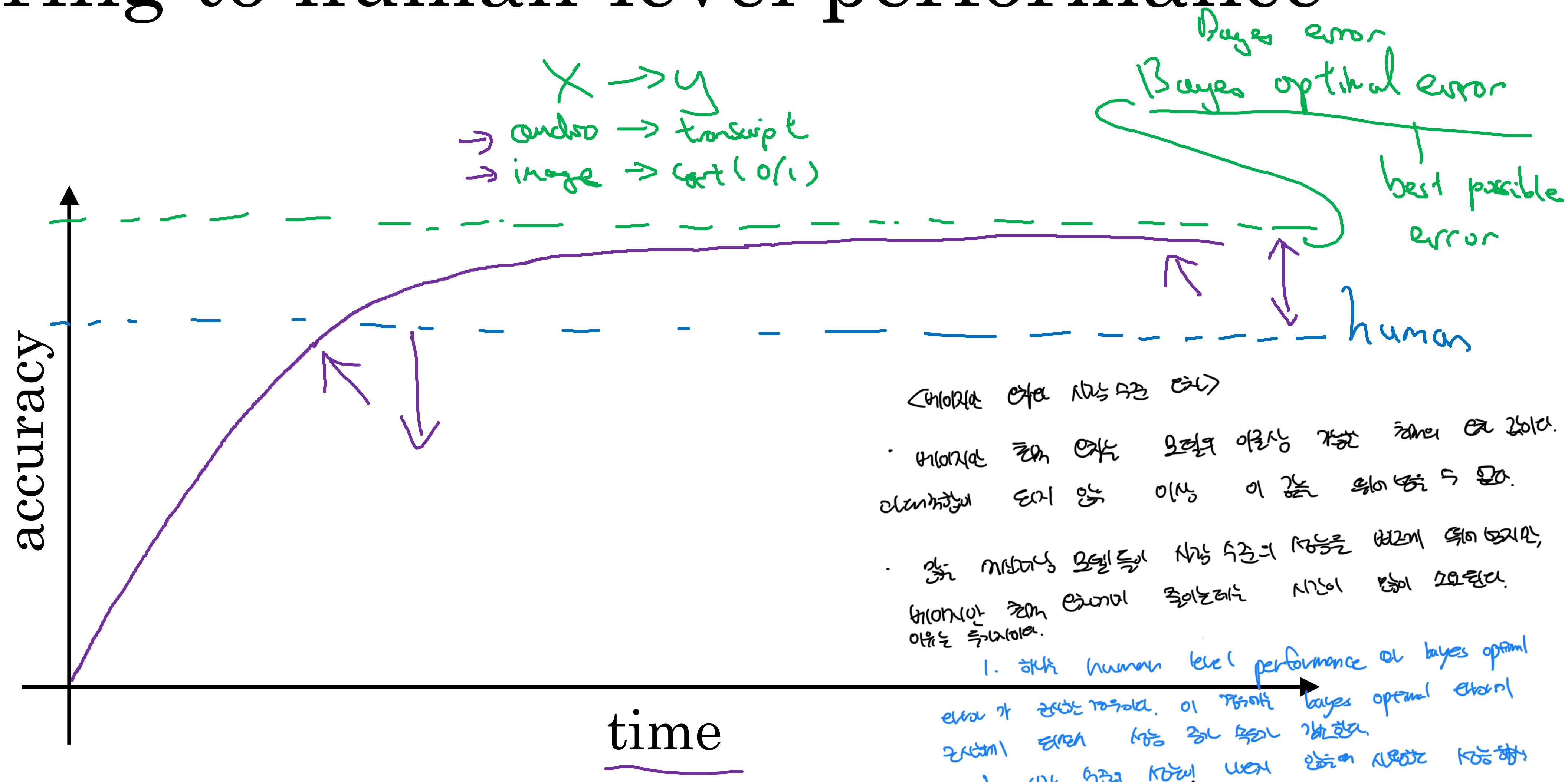
Andrew Ng

deeplearning.ai

Comparing to human-level performance

Why human-level performance?

Comparing to human-level performance



<베이지안 에러 성능 수준 한계>

· 베이지안 최적 에러는 모델의 이론상 가능한 최저의 에러 값이다. 과대적합이 되지 않는 이상 이 값을 넘어설 수 없다.

· 많은 머신러닝 모델들이 사람 수준의 성능을 빠르게 넘어 서지만, 베이지안 최적 에러에 다다르는데는 시간이 많이 소요된다. 이유는 두가지이다.

1. 하나는 human level performance 와 bayes optimal error 가 근접한 경우이다. 이 경우에는 bayes optimal error에 근접하게 되면 성능 향상 속도가 감소한다.

2. 사람 수준의 성능을 넘어서면 이전에 사용하던 성능 향상 기법을 쓸 수 없기 때문이다.

human level performance 에 도달하기 전까지 하이퍼파라미터 튜닝, 알고리즘 개선 등의 방법을 통해 성능 높이는 방법이 있다.

Andrew Ng

Why compare to human-level performance

Humans are quite good at a lot of tasks. So long as ML is worse than humans, you can:

- Get labeled data from humans. (x, y)
- Gain insight from manual error analysis:
  Why did a person get this right?
- Better analysis of bias/variance.

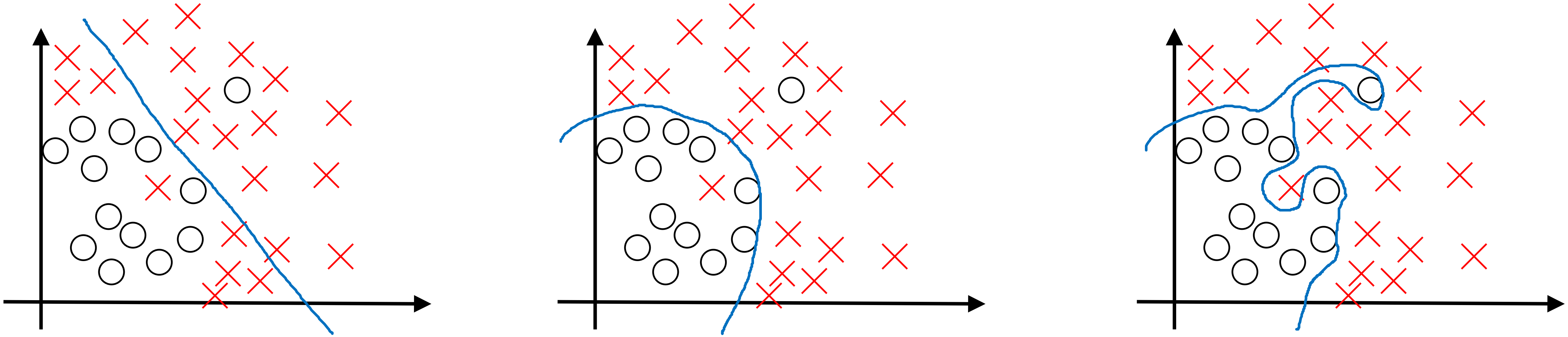
Andrew Ng

deeplearning.ai

Comparing to human-level performance

Avoidable bias

Bias and Variance



high bias

underfitting

"just right"

high variance

overfitting.

Bias and Variance

Cat classification

Human-level ≈ 0% ......

Training set error:

Dev set error:

high variance

high bias

high bias
high variance

low bias
low variance

Cat classification example

인간 레벨 성능에 비해 training error 가 높은 점은 train set 에서 훈련이 제대로 되지 않았음을 의미한다. 이 경우에는 데이터의 편향(bias)을 줄이는 방향으로 더 많은 train set 활용 및 더 오랜시간 동안 train 하는 방법을 사용할 수 있다.

인간 레벨 성능에 비해 training error 가 유사한 경우에는 training error 와 dev error의 줄여서 분산(variance)를 낮추고자 한다. 이 경우에는 정규화(regularization) 또는 데이터를 추가하는 방법들을 시도할 수 있다.

Humans (≈ Bayes)	1%	7.5%
Training error	8%	8 %
Dev error	10%	10 %

7%   2%   0.5%   2%   2%

Avoidable bias

Variance

Variance

Focus on bias

Focus on variance

bayes error와 training error 간의 오차(error) 차이를 avoidable bias 라고 한다. avoidable bias는 최소의 오류값 또는 특정 오류가 있다는 것을 인지하고, over-fitting 하지 않는 이상, bayes error 이하로 내려갈 수 없음을 의미한다.

Human-level error as a proxy for Bayes error.

Andrew Ng

deeplearning.ai

Comparing to human-level performance

Understanding human-level performance

Human-level error as a proxy for Bayes error

Medical image classification example:

Suppose:

(a) Typical human .... 3 % error

(b) Typical doctor .... 1 % error

(c) Experienced doctor .... 0.7 % error

(d) Team of experienced doctors .. 0.5 % error

Bayes error $\leq$ 0.5%

What is "human-level" error?

Andrew Ng

Error analysis example

Human (proxy for Bayes error)

Avoidable bias

Training error

Variance

Dev error

		Bias	Variance	
Human (proxy for Bayes error)		1% / 0.7% / 0.5%	1% / 0.7% / 0.5%	0.7% / 0.5% / 1%
Avoidable bias		4% / 4.5%	0% / 0.5%	0.2% / 0.0%
Training error		5%	1%	0.7%
Variance		1%	4%	0.1%
Dev error		6%	5%	0.8%

< 예시 머신러닝에서 인간 레벨 성능에 도달하는 시점에서 더 개선하기가 더 어려운 이유를 알려준다.> 마지막 예시에서 0.7%의 training error에 도달했을 때, bayes error를 정확하는 것이 명확히 설정되지 않는다면, training error가 bayes error에서 얼마나 차이가 있는지 알기는 어려울 것이고, 그래서 우리는 avoidable bias를 얼마나 줄여야 하는지도 모른다. 추가로, 머신러닝의 성능이 이에 근접하는 상황에서 발전하기 때문에 training set을 Fitting 하기에 어려워진다. 반면, 처음에는 human 레벨 training error가 human level performance와 차이가 많이 나는 상황에서 bias와 variance에 집중하기 쉽다.

이 예시에서는 training error가 0.7%이고 dev error가 0.8%인 경우이다. 이런 경우에는 bayes error(human level error)가 0.5%가 되도록 지정하는 게 중요하다. 이 경우에는 avoidable error가 0.2%이고, bias를 줄이는 것에 조금 더 중점을 둘 수 있다. 만약 0.7%를 bayes error로 추정한다면, avoidable error는 0%가 되고 variance를 줄이는 방향에 중점을 둘 수 있다.

Summary of bias/variance with human-level performance

0%

"Bias"

Human-level error

(proxy for Bayes error)

"Avoidable bias"

Training error

"Variance"

Dev error

deeplearning.ai

Comparing to human-level performance

Surpassing human-level performance

Surpassing human-level performance

Team of humans	0.5%	0.5%
One human	~~1%~~	1%
Training error	0.6%	0.3%
Dev error	0.8%	0.4%

0.1

0.2

0.1~

0.2%

0.3%

이런 경우 avoidable bias를 명확히 알 수 없다.

What is avoidable bias?

이 경우

bayes error : 0.5%

avoidable bias: 0.1%

variance : 0.2% 정도로 추정

⇒ 이런 예제는 training error가 human level performance를 능가했을때, 진행방향을 결정하고 개선할 방향이 명확하지 않다.

Problems where ML significantly surpasses human-level performance

- Online advertising
- Product recommendations
- Logistics (predicting transit time)
- Loan approvals

Structured data

Not natural perception

Lots of data

- Speech recognition
- Some image recognition
- Medical
  - ECG, Skin cancer, ...

deeplearning.ai

Comparing to human-level performance

Improving your model performance

The two fundamental assumptions of supervised learning

1. You can fit the training set pretty well.

   ~ Avoidable bias

   = 낮은 avoidable bias를 갖는다 란 의미

2. The training set performance generalizes pretty well to the dev/test set.

   ~ Variance

   = variance 가 너무 크지 않다는 의미

즉, 머신러닝 성능을 향상시키고자 한다면, training error와 bayes error의 차이를 보고, avoidable bias 얼마를 살펴보고 (training set가 얼마나 잘 fitting 했는지), 그 이후에 dev error와 training error의 차이를 살펴보면 된다.

Andrew Ng

Reducing (avoidable) bias and variance

Human-level

Avoidable bias

Training error

Variance

Dev error

Train bigger model

Train longer/better optimization algorithms

- momentum, RMSprop, Adam

NN architecture/hyperparameters search

RNN
CNN

더 좋은 구조 (NN) 더 좋은 hyperparameter 선택, activation function의 변경

layer의 개수, hidden unit의 개수를 변경하는 방법

More data

Regularization

- L_2, dropout, data augmentation

NN architecture/hyperparameters search

NN 구조 변경

Andrew Ng