Motivating example

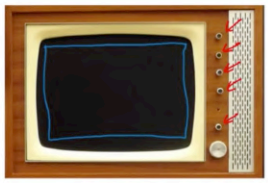
90%

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
  - …

Andrew Ng

TV tuning example



Orthogonalization

Each "knob" only does one thing

0.1 x

+0.5 x

-1.7 x

+0.8 x

+ …

Car

→ Steering

→ {Accelerator, Braking}

→ 0.3 x angle − 0.8 speed

→ 2 x angle + 0.9 speed

speed

angle

Andrew Ng

Chain of assumptions in ML

→ Fit training set well on cost function

(≈ human-level performance)

- bigger network
- Adam

→ Fit dev set well on cost function

- Regularization
- Bigger training set

early stopping

→ Fit test set well on cost function

- Bigger dev set

→ Performs well in real world

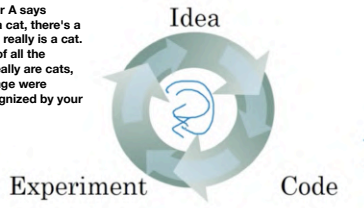
(Happy cat pic app users.)

- Change dev set or cost function

Andrew Ng

Using a single number evaluation metric

So if classifier A has 95% precision, this means that when classifier A says something is a cat, there's a 95% chance it really is a cat. And recall is, of all the images that really are cats, what percentage were correctly recognized by your classifier?



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

What % of actual cats are correctly recognized

Classifier	Precision	Recall	F1 Score
A	95%	90%	92.4%
B	98%	85%	91.0%

F1 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

Andrew Ng

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%

Error ↑

Andrew Ng

Another cat classification example

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

Somewhat artificial → Cost = accuracy − 0.5 × running time

"Better..." → Maximize accuracy subject to running time ≤ 100 ms.

N metrics: 1 optimizing, N−1 satisficing

Wakewords / Trigger words

Alexa, OK Google, Hey Siri, 你好百度

accuracy, #false positive

Maximize accuracy ← optimizing

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

Andrew Ng

Cat classification dev/test sets

Regions:

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

Dev

Test

Randomly shuffle into dev/test

Bad! different distribution

True story (details changed)

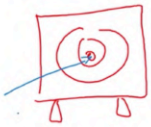
Optimizing on dev set on loan approvals for medium income zip codes

Tested on low income zip codes

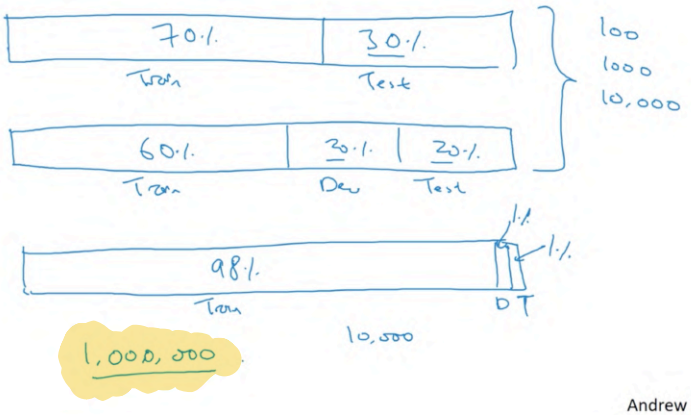
~3 month

Guideline

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



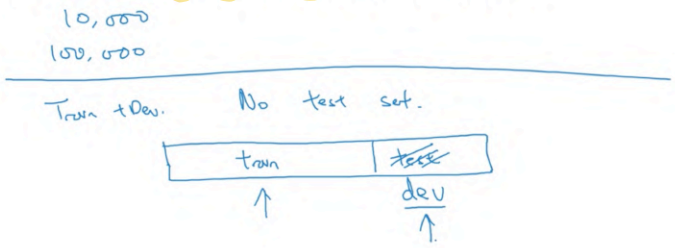
Old way of splitting data



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.



Andrew Ng

Cat dataset examples

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

Sign that you should Change the evaluation metric or dev/test set

→ Metric: classification error

Algorithm A: 3% error → pornographic

✓ Algorithm B: 5% error

$$\text{Error}: \frac{1}{\sum w^{(i)}} \frac{1}{m_{dev}} \sum_{i=1}^{m_{dev}} w^{(i)} \mathbb{1}\{y_{pred}^{(i)} \neq y^{(i)}\}$$

($y_{pred}^{(i)}$: predicted value (0/1))

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

high level of take away is, if you find that evaluation metric is not giving the correct rank order preference for what is actually better algorithm, then there's a time to think about defining a new evaluation metric.

Andrew Ng

Orthogonalization for cat pictures: anti-porn

1. So far we've only discussed how to define a metric to evaluate classifiers.
2. Worry separately about how to do well on this metric.

Andrew Ng

Another example

Algorithm A: 3% error

Algorithm B: 5% error

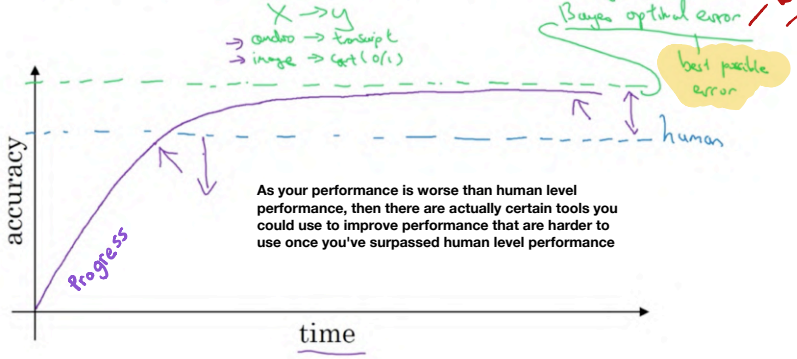
Dev/test

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

Comparing to human-level performance



As your performance is worse than human level performance, then there are actually certain tools you could use to improve performance that are harder to use once you've surpassed human level performance

Andrew Ng

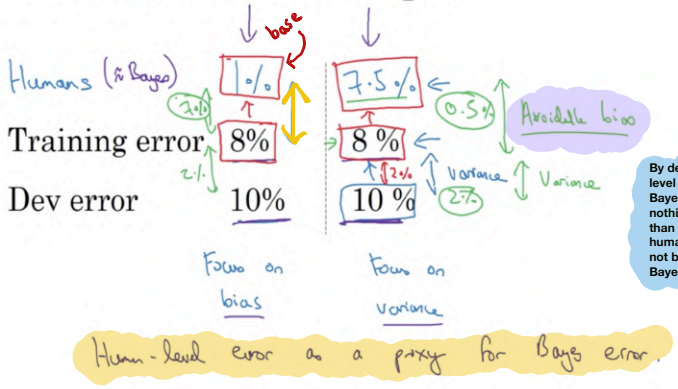
Why compare to human-level performance

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

Cat classification example



By definition, human level error is worse than Bayes error because nothing could be better than Bayes error but human level error might not be too far from Bayes error.

Andrew Ng

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

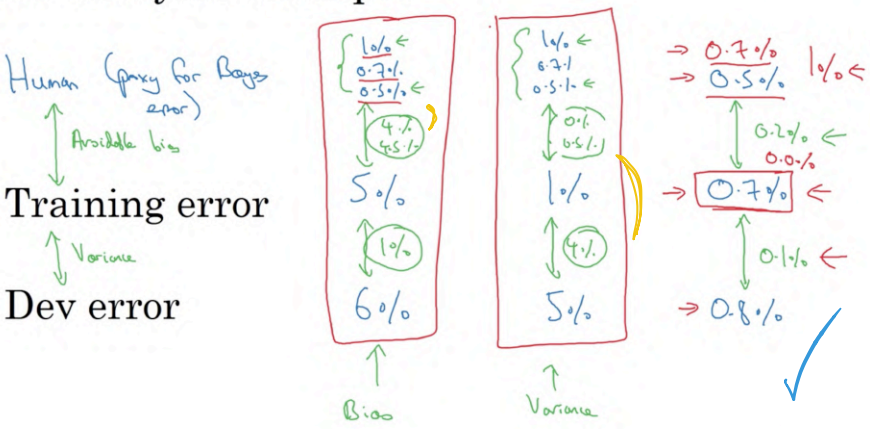
Bayes error $\leq$ 0.5%

What is "human-level" error?

Be clear when defining human level error

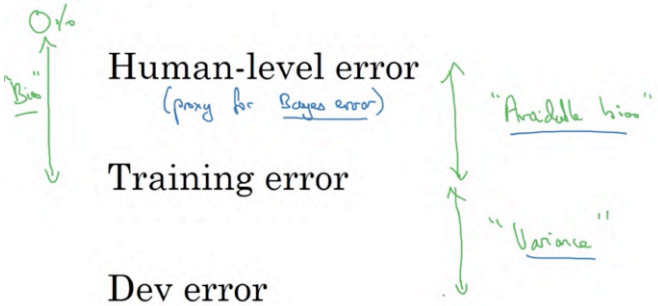
Andrew Ng

Error analysis example



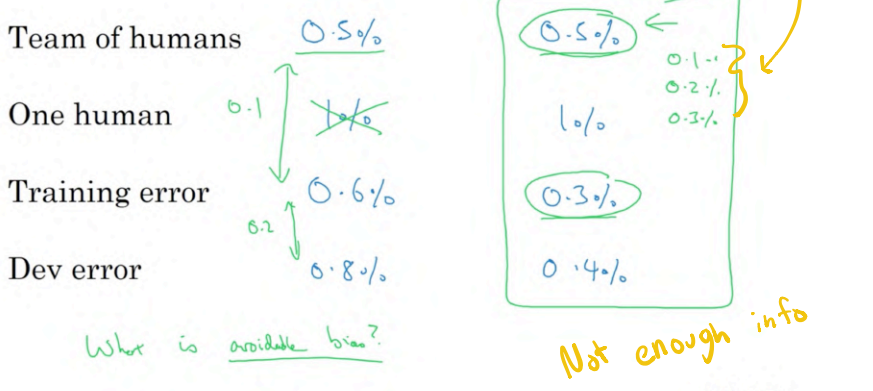
Andrew Ng

Summary of bias/variance with human-level performance



Andrew Ng

Surpassing human-level performance



Andrew Ng

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, ...

Andrew Ng

The two fundamental assumptions of supervised learning

1. You can fit the training set pretty well.

   low
   ~ Avoidable bias

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

   ~ Variance

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

More data
Regularization
- L_2, dropout, data augmentation
NN architecture/hyperparameters search

Andrew Ng