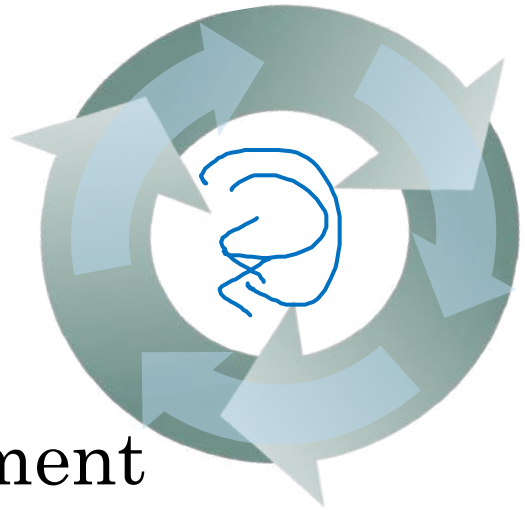
deeplearning.ai

Setting up your goal

Single number evaluation metric

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%

F1 score = "Average" of P and R.

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

Dev set + Single number evaluation metric

Speed up iterating

Another example

Algorithm	US	China	India	Other
A	3%	7%	5%	9%
B	5%	6%	5%	10%
C	2%	3%	4%	5%
D	5%	8%	7%	2%
E	4%	5%	2%	4%
F	7%	11%	8%	12%

deeplearning.ai

Setting up your goal

Satisficing and optimizing metrics

Another cat classification example

optimizing ↓ ↓ satisficing

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

Cost = accuracy − 0.5 * running Time

Maximize accuracy

subject to runningTime ≤ 100 ms.

N metrics: 1 optimizing

N−1 satisficing

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:

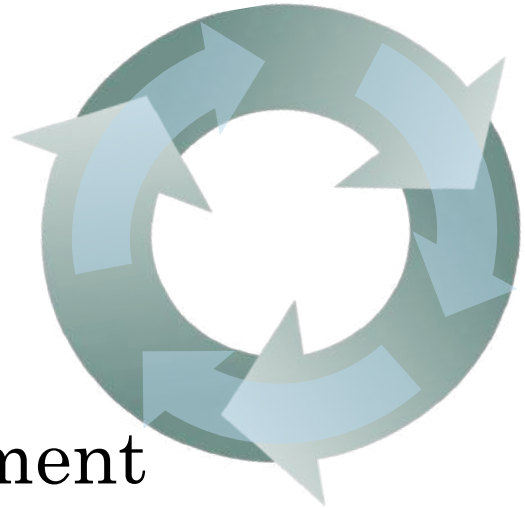
- US
- UK
- Other Europe
- South America
- India
- China
- Other Asia
- Australia

Dev

Test

Randomly shuffle into dev/test

dev set + metric



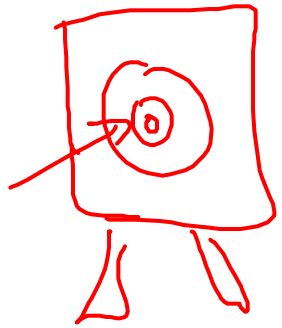
True story (details changed)

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

$x \rightarrow y$ (repay loan?)

Tested on low income zip codes

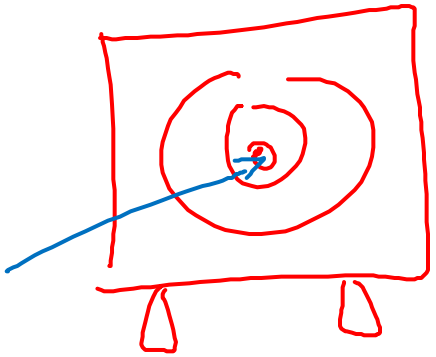
~3 month



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

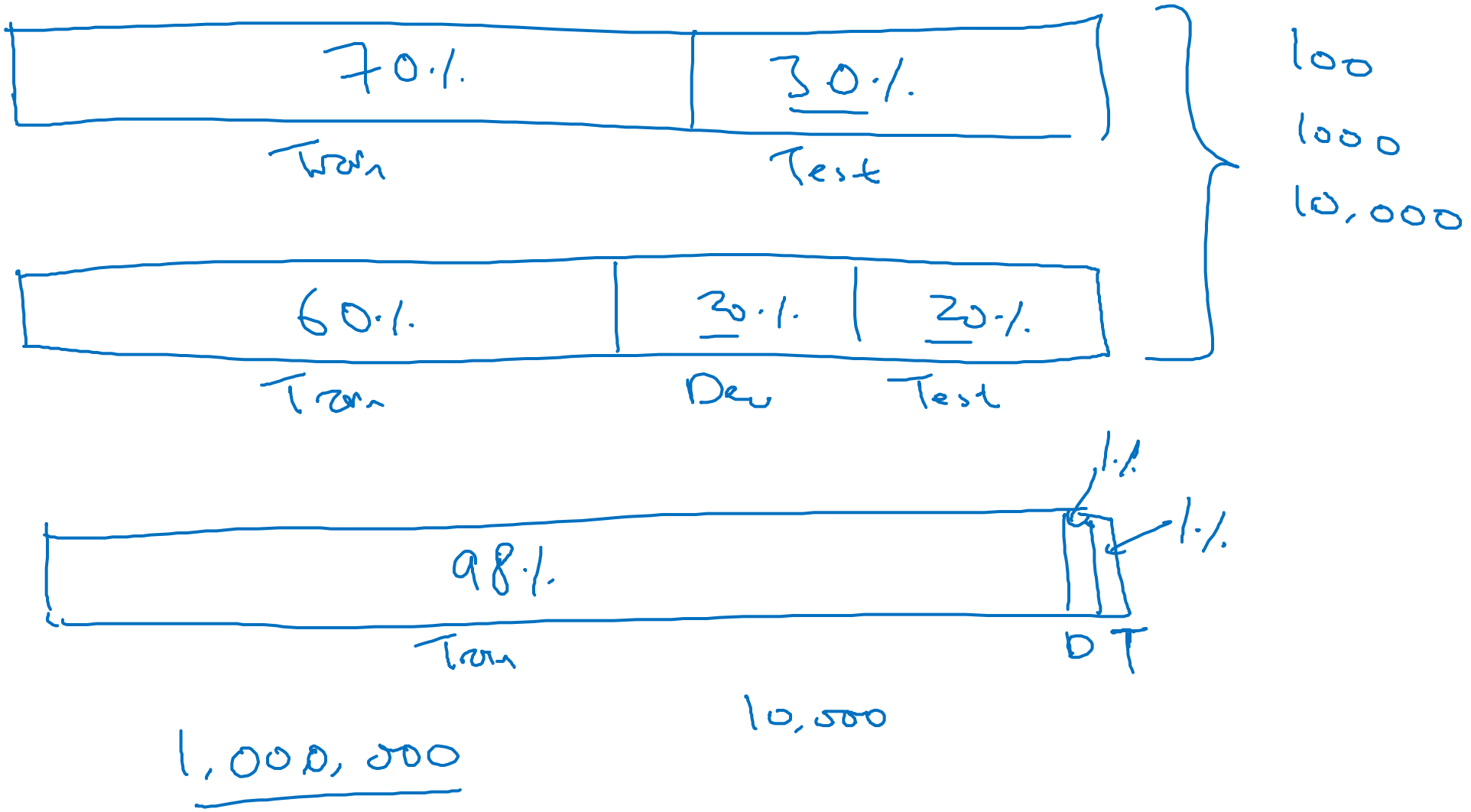
test

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.

A B

97% → 97.1%

0.1%

100: small

1%

1,000

10,000

100,000

0.01%

0.001%

Online advertising

Size of test set

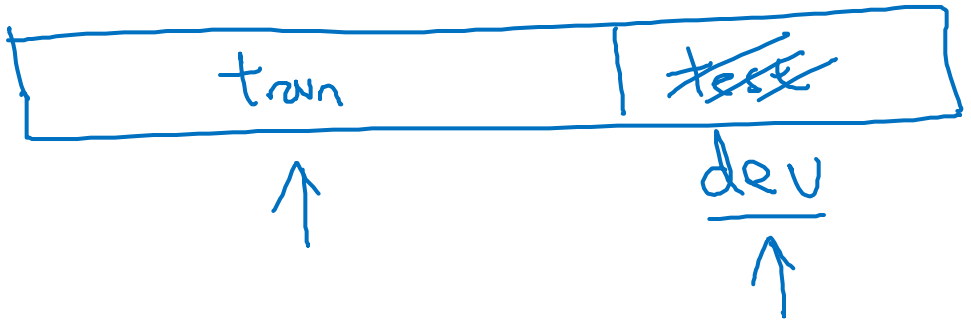
- Set your test set to be big enough to give high confidence in the overall performance of your system.

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

Your users : Prefer B.

→ Metric: classification error

Algorithm A: 3% error → pornographic

✓ Algorithm B: 5% error

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

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

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

Orthogonalization for cat pictures: anti-porn

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

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

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.