deeplearning.ai

## Setting up your goal

# When to change dev/test sets and metrics

① When

# Cat dataset examples

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

→ Metric: classification error

Algorithm A: 3% error 将一些色情图片一并推送 → pornographic

✓ Algorithm B: 5% error

find a new evaluation metric

$$\text{Error}: \frac{1}{\sum_i w^{(i)}} \cancel{\frac{1}{m_{dev}}} \sum_{i=1}^{m_{dev}} w^{(i)} \mathcal{I}\{\underbrace{y_{pred}^{(i)}}_{\text{predicted value (0/1)}} \neq y^{(i)}\}$$

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

# 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

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

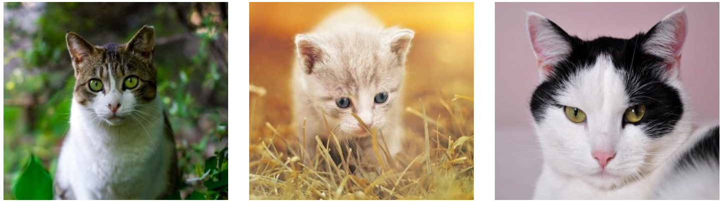
② when

# Another example

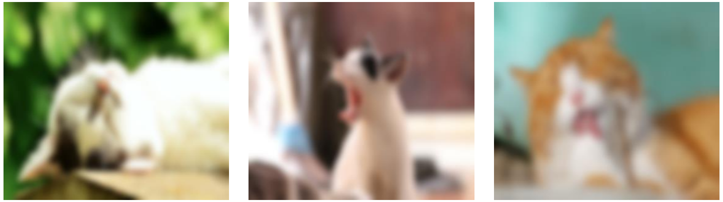
Algorithm A: 3% error

✓ 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.

# When to change development/test sets and metrics

## Example: Cat vs Non-cat

A cat classifier tries to find a great amount of cat images to show to cat loving users. The evaluation metric used is a classification error.

| Algorithm | Classification error [%] |
|---|---|
| A | 3% |
| B | 5% |

It seems that Algorithm A is better than Algorithm B since there is only a 3% error, however for some reason, Algorithm A is letting through a lot of the pornographic images.

Algorithm B has 5% error thus it classifies fewer images but it doesn't have pornographic images. From a company's point of view, as well as from a user acceptance point of view, Algorithm B is actually a better algorithm. The evaluation metric fails to correctly rank order preferences between algorithms. The evaluation metric or the development set or test set should be changed.

The misclassification error metric can be written as a function as follow:

$$Error: \frac{1}{m_{dev}} \sum_{i=1}^{m_{dev}} \mathcal{L}\left\{(\hat{y}^{(i)} \neq y^{(i)}\right\}$$

This function counts up the number of misclassified examples.

The problem with this evaluation metric is that it treats pornographic vs non-pornographic images equally. On way to change this evaluation metric is to add the weight term $w^{(i)}$.

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

The function becomes:

$$Error: \frac{1}{\sum w^{(i)}} \sum_{i=1}^{m_{dev}} w^{(i)} \mathcal{L}\left\{(\hat{y}^{(i)} \neq y^{(i)}\right\}$$

Guideline

1. Define correctly an evaluation metric that helps better rank order classifiers
2. Optimize the evaluation metric