Evaluation of a classifier

Accuracy = Proportion of examples w/ correct prediction

y_d gold, $\hat{y}_d$ prediction
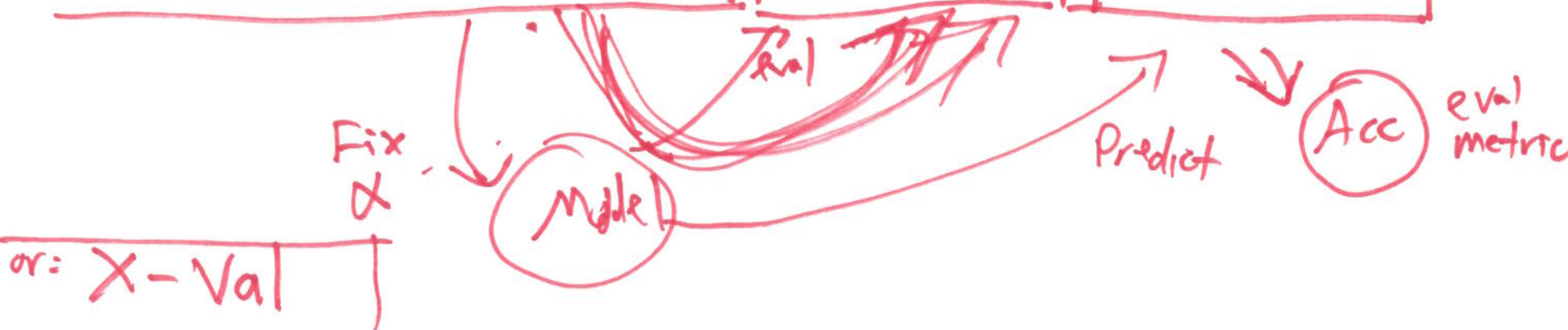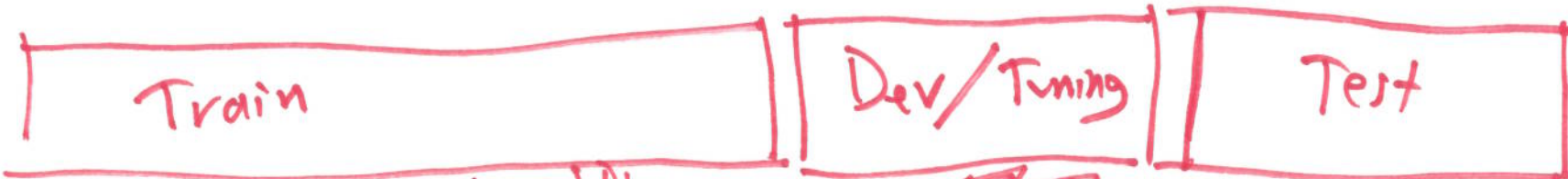
$$Acc = \frac{Count(y_d = \hat{y}_d)}{Num\ docs\ in\ test\ set}$$

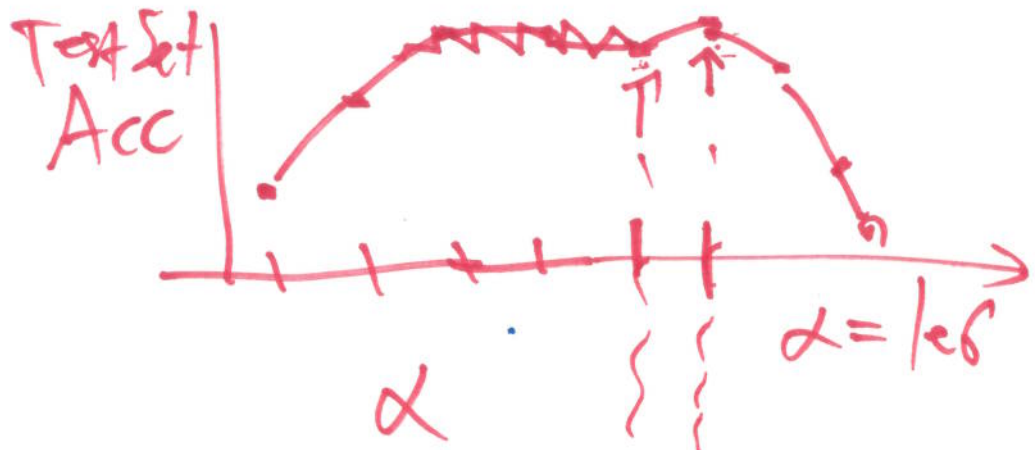$$= \frac{\sum_{d \in TestSet} \mathbb{1}\{y_d = \hat{y}_d\}}{|TestSet|}$$

$\mathbb{I}(y = \bar{y})$

int(y == yhat)

~~$f(\quad)$~~

| Train | Dev/Tuning | Test |

Eval

Fix α

Model

Predict

Acc — eval metric

or: X-Val

Test Set
Acc

α

$\alpha = 1e6$

Does test set acc
generalize?

Overusing test set $\Rightarrow$
Subtle overfitting

$P(w_d | d)$

$P(w_d)$

Metrics

doc
↓

True Neg
↓   ↓

Gold: ⊕ ⊕ ⊕ ⊖ ⊖ ⊖ ⊖ ⊖
Pred: ⊕ ⊕ ⊖ ⊖ ⊖ ⊕ ⊕

True Pos

error
False Negative

errors:
False Pos
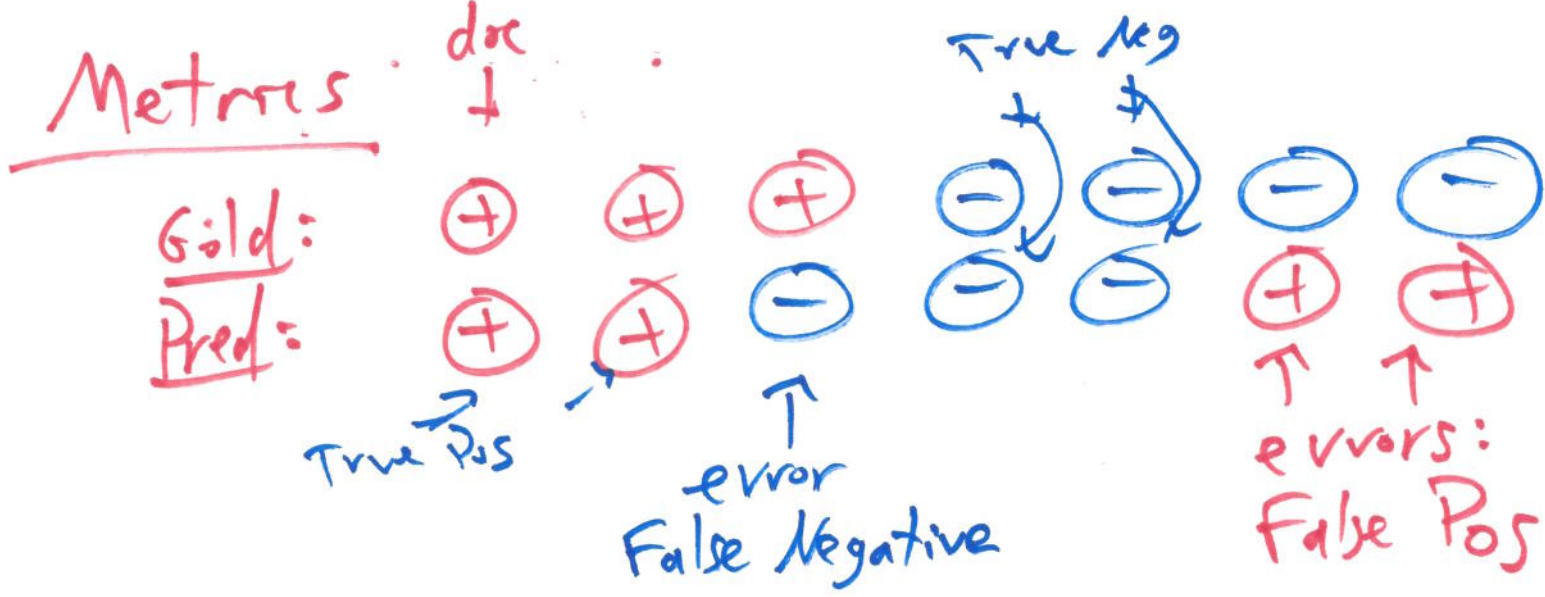
$$\text{Acc Rate} = 4/7$$

Confusion Matrix

Gold

Pred	⊕	⊖
⊕	2 TP	2 FP
⊖	1 FN	2 TN

Precision & Recall

$$\text{Prec} = \% \text{ correct of POS pred}$$

$$= \frac{TP}{TP+FP} = \frac{2}{4} = .5$$

$$\text{Recall} = \% \text{ correct of POS gold std}$$

$$= \frac{TP}{TP+FN} = \frac{2}{3}$$

Sarcasm Detection — UMass CS 585 in-class exercise, 9/19/17

These are Reddit comments (from https://arxiv.org/pdf/
1704.05579.pdf). Their authors tagged them as "sarcastic" or "not
sarcastic". These tags are hidden. The task is to predict whether the
message was tagged as sarcasm. Label them 1=sarcastic, 0=not
sarcastic.

1. but CNN told me the leaks are all faked by Russia!

2. Is this some sort of mug shot mash up?

3. there isn't... the website says specifically there is no way to change
it once placed, you must cancel the order and place a new one to
make any modifications.

4. All men are handsome!

5. Yeah but he paid his dues so it's his turn.

6. ....what?

7. Maybe you should stop reading it

8. It's okay, it's just his opinion!

9. I know that's what she's doing now, I'm saying this isn't the first
time I've seen her.

10. honestly i'd have turned around and sold it to buy the K. But at
the end of the day, if you aren't really one to overclock, then what
difference does it make to you anyway?

Handwritten annotations:

IAA   2 (# Human Pairs)

100%   2 4

90%   2(7) = 14

80%   More

Random Agreement?
50%