Error Analysis:

- Start with a simple algorithm
- Test with cross-validation data
- Plot learning curves
- Error Analysis
  - examine misclassified examples or residual for clues

——— Example From Spam Classifier ———

NLP - "Porter Stemmer"

↳ treat discount/discounted/discounting as a single word?

↳ treat Mom/mom as same word?

To evaluate, compute J_{cv} with and without stemming.

✳ Do error analysis on cross-validation set to preserve independence of J_{TEST} ✳

Prioritizing System Design

- Start with a simple algorithm
- Test with CV data
- Plot learning curves
- Error Analysis
- Make a list of options for:
  - features
  - data sources
  - requirements

Error Metrics For Skewed Classes

skewed classes: $p \approx 0$ or $p \approx 1$ for a given class

Precision/Recall:

$y=1$ in presence of rare class we want to detect ← this is a convention

PREDICTED \ ACTUAL	1	0
1	TRUE POSITIVE	FALSE POS
0	FALSE NEG	TRUE NEGATIVE

Precision: $\dfrac{\text{True Positives}}{\text{Predicted Positives}} = \dfrac{\text{TRUE POS}}{\text{TRUE POS + FALSE POS}}$

Recall: $\dfrac{\text{TRUE POSITIVES}}{\text{TRUE POS + FALSE NEG}}$

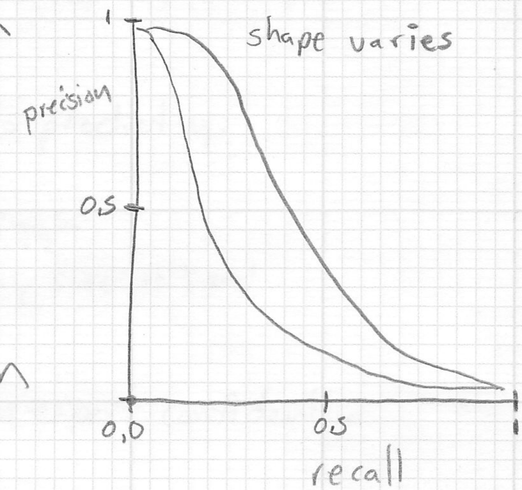
✳ Goal: High Precision & High Recall

Trading off Precision And Recall

Use $h_\theta(x) \geq \xi$ $\quad \xi > 0.5$ $\quad \xi$ = "threshold"

trade off: lower recall for higher precision

Use to avoid false positives,



Suppose we want to avoid false negatives:

$h_\theta(x) \geq \xi$ $\quad \xi < 0.5$

trade off: higher recall for lower precision

How to compare precision/recall numbers? "F-score"

Algorithm	P	R	$\frac{1}{2}(P+R)$	F_1 Score $= 2\frac{PR}{P+R}$
1	0.5	0.4	0.45	0.444
2	0.7	0.1	0.4	0.175
3	0.02	1.0	0.51	0.0392

not good:
predicts 3 as best
when it could just be
returning 1

$P=0 \;||\; R=0 \rightarrow F_1 = 0$

$P=1 \;\&\&\; R=1 \rightarrow F_1 = 1$

Use F_1 on the cross-validation set to adjust the threshold.

Data For Machine Learning

Banko & Brill, 2001

"It's not who has the best algorithm that wins. It's who has the most data."

- When does having a large data set help?

Assume feature $x \in \mathbb{R}^{n+1}$ has sufficient information to predict y.

Example: For breakfast I ate $\{$too$|$two$|$to$\}$ eggs.

Counter-Example: Predict housing price from only size(ft^2) and no other features.

*Useful test: Given the input x, can a human expert confidently predict y?

Large training sets help prevent over fitting - lower variance

Allows use of more features to reduce bias