Sid

## Spam Classifier

$\vec{x}$ : features of email, $\vec{y}$ = spam (1) or not spam (0)

$m$ = 100 words

$x \in \mathbb{R}^{100}$ $\quad x_j = \begin{cases} 1 & \text{word } j \text{ appears} \\ 0 & \text{otherwise} \end{cases}$

$$x \begin{bmatrix} 0 \\ 1 \\ 1 \\ 0 \\ \vdots \end{bmatrix}$$

## Error Analysis

- Start w/ simple algorithm & test on cross-validation data
- Plot learning curves
- Error analysis : manually examine errors (numerical value for error)

## Handling Skewed Data

Skewed clases → lot more/less examples for one class vs the other.

## Precision/Recall ($y = 1$ in presence of 'rare' class)

| Predicted \ Actual | 1 | 0 |
|---|---|---|
| 1 | True + | False + |
| 0 | False − | True − |

$$\text{Precision} : \frac{\text{True}^+}{\#\text{ predicted} +} = \frac{\text{True}^+}{\text{True}^+ + \text{False}^+}$$

$$\text{Recall} : \frac{\text{True}^+}{\#\text{ actual} +} = \frac{\text{True}^+}{\text{True}^+ + \text{False}^-}$$

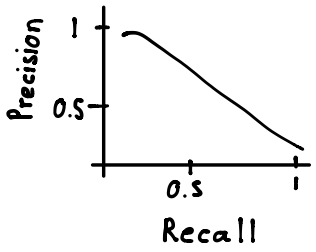
Logistic Regression : $0 \le h_\theta(x) \le 1$

Predict 1 if $h_\theta(x) \geq$ threshold
Predict 0 if $h_\theta(x) <$

---

threshold $\geq 0.5$
↳ ↑ Precision, ↓ Recall

threshold $< 0.5$
↳ ↓ Precision, ↑ Recall




F, Score : $2\frac{PR}{P+R}$
$[0, 1]$

Large Data :

- use algorithm w/ many parameters (low bias)
  → $J_{train}(\theta)$ small
- Use very large training set (unlikely to overfit - low variance)
  → $J_{train}(\theta) \approx J_{test}(\theta)$

$J_{test}(\theta)$ small