Lecture 21

- why does the cross entropy loss make sense?

Imagine this scenario

coin, $p_H = ? = \theta$

$p_T = 1 - p_H = 1 - ? = 1 - \theta$

Sample Size $= 10$

$\{0, 0, 0, 0, 0, 0, 1, 1, 1, 1\}$

$\hookrightarrow$ $n = 10$

$n_H = 6$

$n_T = 4$

- Our best guess is $\hat{\theta} = 0.6$

$\downarrow$

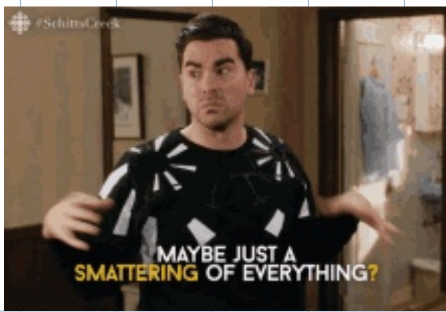
Likelihood

Probability that I see this result given an estimate for θ

$P(Y_{1-6} = 0, Y_{7-10} = 1 \mid \theta = 0.6)$ High

$P(Y_{1-6} = 0, Y_{7-10} = 1 \mid \theta = 0.3)$ Low

Ok... cool but what does this have to do w/ logistic regression.



Our data is a sample; the coin is the world model we try to predict.

$$f_\theta(x) = \sigma(x^T\theta) = p(Y=1 \mid x) = p$$

Our goal maximize likelihood

for $\sigma(x^T\theta) = f_\theta(x) = p$

Back to the coin

$$p(Y_{1-6} = \emptyset,\ Y_{7-10} = 1 \mid \theta = p)$$

Bernoulli(p)

$p^y (1-p)^{(1-y)}$

$$\prod_{i=1}^{10} p^{y_i} (1-p)^{(1-y_i)}$$

generalize

for n samples of data

and our model $f_\theta(x) = \sigma(x^T\theta)$

$\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ " p

$$\prod_{i=1}^{n} p^{y_i} (1-p)^{(1-y_i)}$$

y_i : observed data

p : predicted probability, as output by $\sigma(x^T\theta)$

$$\prod_{i=1}^{n} (\sigma(x^T\theta))^{y_i} (1 - \sigma(x^T\theta))^{(1-y_i)}$$

Our goal for $\hat{\theta}$:

find best $\hat{\theta}$ parameters that maximize the likelihood of observing our data samples $\{(x_1, y_1) \ldots (x_n, y_n)\}$

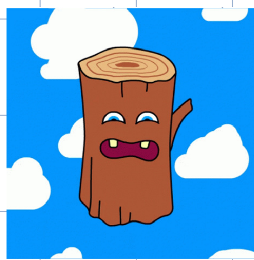
$$\hat{\theta} = \underset{\theta}{\arg\max} \prod_{i=1} \sigma(x^T\theta)^{y_i} (1 - \sigma(x^T\theta))^{(1-y_i)}$$

<u>Trivia</u> log is monotonically increasing;

therefore $\underset{\theta}{\arg\max}\, f(x) = \underset{\theta}{\arg\max}\, \log(f(x))$

iff $f(x) > \emptyset$

$$\hat{\theta} = \underset{\theta}{\arg\max} \; \log\left(\prod_{i=1}^{n} \sigma(x^T\theta)^{y_i} \left(1 - \sigma(x^T\theta)\right)^{(1-y_i)}\right)$$



$$= \underset{\theta}{\arg\max} \sum_{i=1}^{n} \log\left[\sigma(x^T\theta)^{y_i}\right] + \log\left(\left(1 - \sigma(x^T\theta)\right)^{(1-y_i)}\right)$$

$$= \underset{\theta}{\arg\max} \sum_{i=1}^{n} y_i \log(\sigma(x^T\theta)) + (1-y_i)\log(1 - \sigma(x^T\theta))$$

$$= \underset{\theta}{\arg\min} \; -\sum_{i=1}^{n} y_i \log(\sigma(x^T\theta)) + (1-y_i)\log(1 - \sigma(x^T\theta))$$

CROSS entropy loss