Perceptron

Distribution over $\mathbb{R}^d \times \{\pm 1\}$

$\mathcal{X} \times \mathcal{Y}$

Generative Model — $P(x,y)$ → Joint distribution of (feature, label)

Discriminative Model — $\boxed{P(y|x)}$ (or) some mapping from $\mathcal{X} \to \mathcal{Y}$

$\boxed{P(x,y)}$

$D = \{(x_1,y_1)(x_2,y_2)\ldots(x_n,y_n)\}$

Decision Rule:

$\boxed{P(y=1/x) > P(y=-1/x)}$ Predict $y=1$

o/w Predict $y=-1$

Perceptron:

$$\boxed{\begin{aligned} P(y=+1/x) &= 1 \quad \text{if } w^T x > 0 \\ &= 0 \quad \text{if } w^T x < 0 \end{aligned}}$$

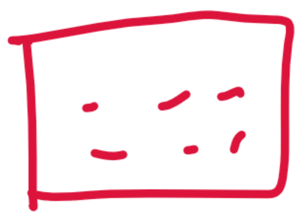
↓

Discriminative model

Generative model

$\{(x_i, y_i)\}_{i=1}^{n}$ $\quad y_i \in \{0,1\}$ $\quad x_i \in \{0,1\}^d$

Eg: spam classification

 $\longrightarrow \{0,1\}^d$

Dictionary of words $\quad$ $d \to$ # words.

$$\begin{array}{cccccccc} w_1 & w_2 & \cdot & \cdot & \cdot & \cdot & & w_d = 10{,}000 \\ \end{array}$$
$$[\,0 \quad 1 \quad 0 \quad \cdot \quad \cdot \quad \cdot \quad \cdot \quad]$$

Generative Story

$\boxed{P(x,y)}$

$P(y) \cdot P(x/y)$

- $y_i \sim$ Bernoulli (p)
- $P(\underbrace{x_i}_{\{0,1\}^d}/y_i)$ $\quad 2^d$ possible values for x_i

In the most general case, model $P(x_i/y_i)$ using a 2^d dimensional probability vector.

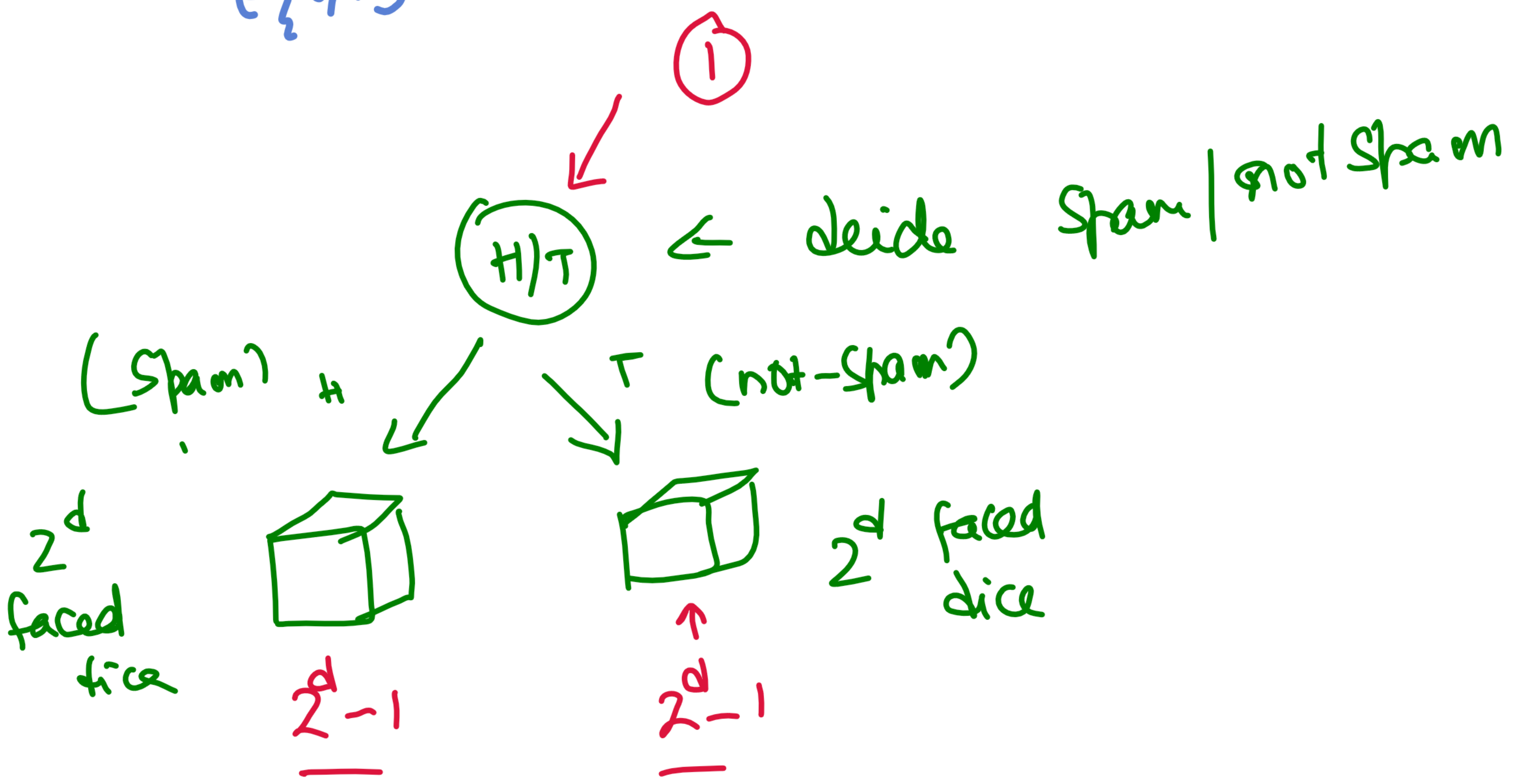
$\underline{d=3}$

$y_i = 1$ (spam)

$$x_i \quad 000 \quad 001 \quad 010 \quad 011 \quad 100 \quad 101 \quad 110 \quad 111$$

$$\left[\ p^1_{000} \quad p^1_{001} \quad \cdots \quad p^{1}_{111}\ \right] \quad \leftarrow y_i = 1$$

$$\sum_{\substack{b_1, b_2, b_3 \\ \in \{0,1\}}} p_{b_1 b_2 b_3} = 1 \quad \Leftarrow$$



$\underline{\#\text{ parameters?}}$ $\quad (2^d - 1) \times 2 + 1 = \underline{\underline{2^{d+1} - 1}}$

Too large!

$$P(x_i / y_i) = P\left(\underline{[f^i_1 \ f^i_2 \ \cdots \ f^i_d]} / y_i\right)$$

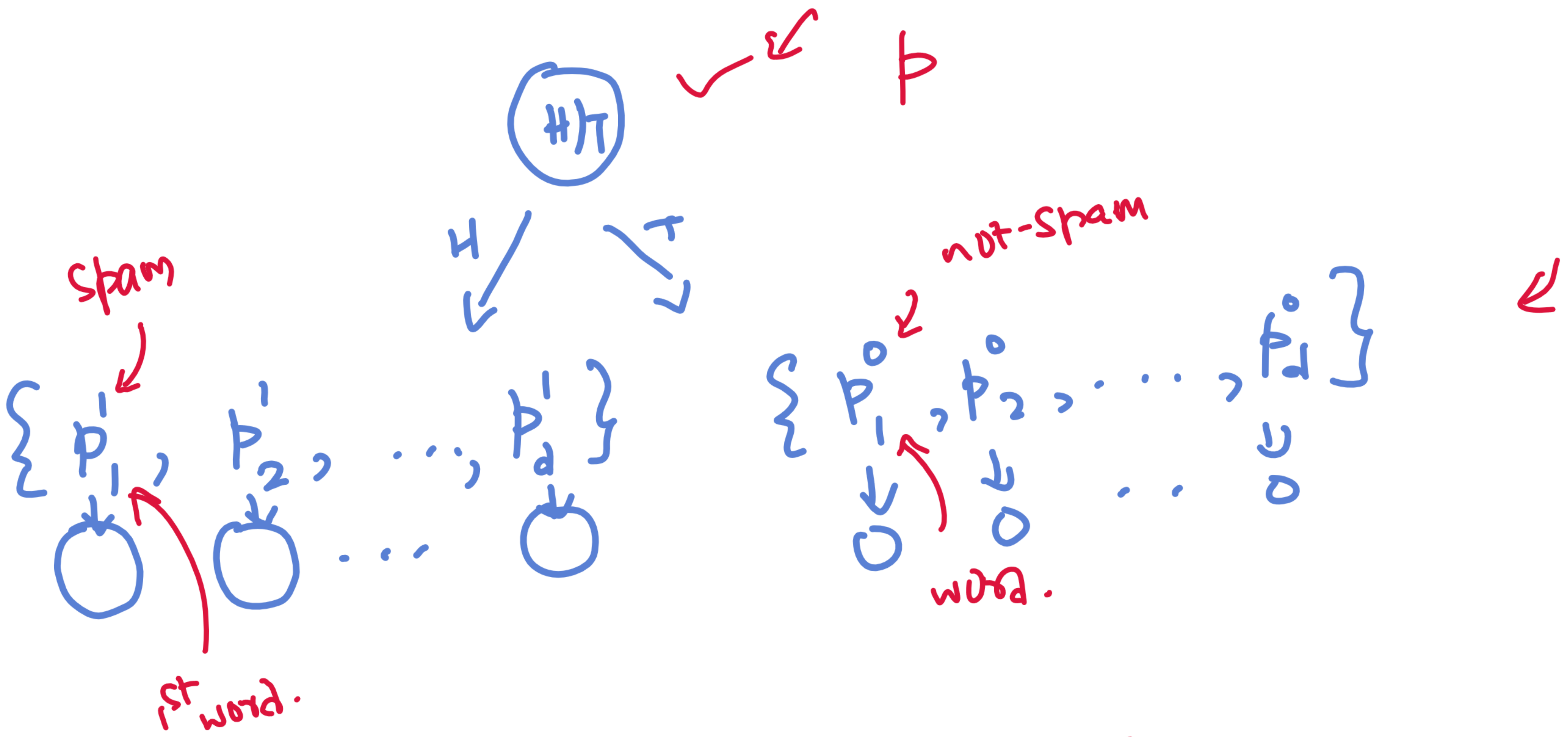
$$f^i_j \in \{0,1\} \qquad \forall i = 1, \ldots n \quad \forall j = 1, \ldots d.$$

$\underline{\text{ASSUMPTION}}$

"Features are Conditionally independent given the label/class"

$$P(x_i / y_i) = \prod_{j=1}^{d} \underline{P(f^i_j / y_i)}$$

$$f^i_j / y_i \sim \text{Bernoulli}\left(p^{y_i}_j\right)$$



$$\# \text{ parameters } = \boxed{2d+1}$$

Estimate

p, $[p_1^0, \ldots, p_d^0]$, $[p_1^1, \ldots p_d^1]$

$$\hat{p} = \frac{1}{n} \sum_{i=1}^{n} y_i$$

Fraction of spam mails!

non-spam

$$\hat{p}_j^0 = \frac{\sum_{i=1}^{n} \mathbb{1}(f_j^i = 1, y_i = 0)}{\sum_{i=1}^{n} \mathbb{1}(y_i = 0)}$$

j^{th} word

Exercise

Verify these are indeed MLE estimates!

$$\hat{p}_j^1 = \frac{\sum_{i=1}^{n} \mathbb{1}(f_j^i = 1, y_i = 1)}{\sum_{i=1}^{n} \mathbb{1}(y_i = 1)}$$

Prediction

$x^{test} \in \{0,1\}^d$

$P(y^{test}=1 / x^{test}) > P(y^{test}=0 / x^{test})$

Predict spam

o/w predict not-spam.

$$P(y^{test}=1/x^{test}) = \frac{\boxed{P(x^{test}/y^{test}) \cdot P(y^{test}=1)}}{P(x^{test})}$$

$P(x^{test}) \leftarrow$ Evidence.

$$\propto \prod_{j=1}^{d} P(f_j^{test} / y^{test}=1) \cdot \hat{p}$$

$$\propto \prod_{j=1}^{d} (\hat{p}_j^1)^{f_j^{test}} (1-\hat{p}_j^1)^{(1-f_j^{test})} \cdot \hat{p}$$

NAIVE - BAYES ALGORITHM.

NAIVE $\downarrow$ Class-conditional independence

BAYES $\rightarrow$ Bayes theorem to predict.