TYPES OF MODELING

DISTRIBUTION D OVER $X \times Y$ ← $\{+1, -1\}$ (FEATURE SPACE)

$\{x_1, \ldots, x_n\}$
$y_1, \ldots, y_n$

UNKNOWN BUT FIXED

$(x_i, y_i) \sim D$ ← LINK BETWEEN TRAIN and TEST.

CLASSIFICATION

- GENERATIVE MODEL
- DISCRIMINATIVE MODEL

GENERATIVE MODEL

MODEL

- $P(x, y)$

↳ NEXT.

DISCRIMINATIVE MODEL

MODEL

- $P(y/x)$

Eg: K-NN
DECISION-TREES

↳ $P(y=1/x) = 1$ if decision tree for x says 1.

$P(y=1/x) = 1$ if majority of neighbours say 1
$= 0$ otherwise

DATA = $\{(x_1, y_1) \ldots (x_n, y_n)\}$

$x_i \in \{0,1\}^d$

$y_i \in \{0,1\}$

EXAMPLE: SPAM CLASSIFICATION

$x_i \in \{0,1\}^d$ ← # words in dictionary.

Eg: "Hello, how are you?"

$P(x, y)$ $\quad$ ("Hello, how are you?", spam) $= 0.01$

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

$P(x)$: $P(\text{email})$ $\quad$ $P(y/x)$: $P(\text{spam}/\text{email})$

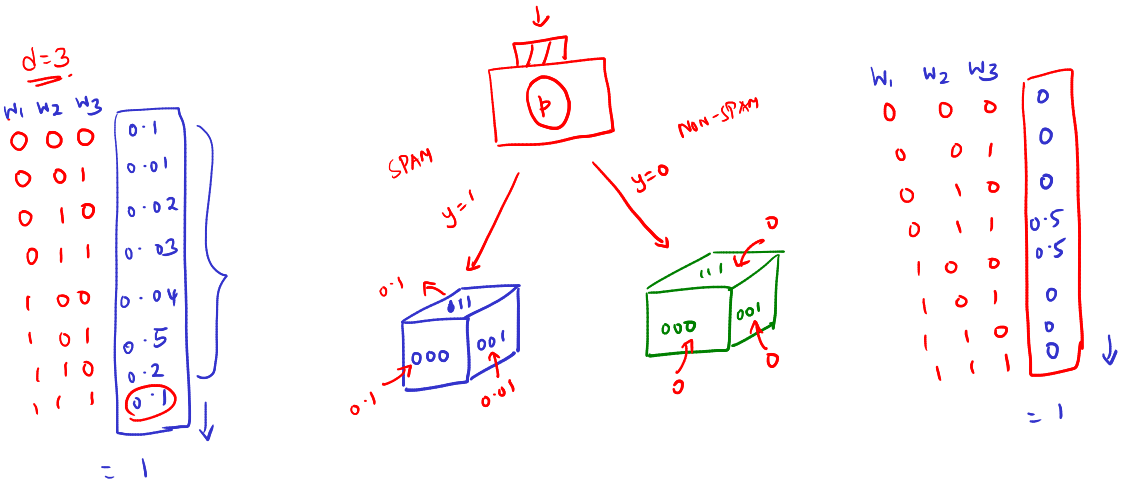
$P(y)$: $P(\text{spam})$ $\quad$ $P(x/y)$: $P(\text{email}/\text{spam})$

GENERATIVE STORY

STEP-1: DECIDE the Label by tossing a coin

$P(y_i = 1) = p$

STEP 2: DECIDE features using the label in step 1 by $P(x_i / y_i)$



parameters in the model $= 1 + (2^d - 1) + (2^d - 1)$

1: to decide label

$(2^d - 1)$: $P(x|y=1)$

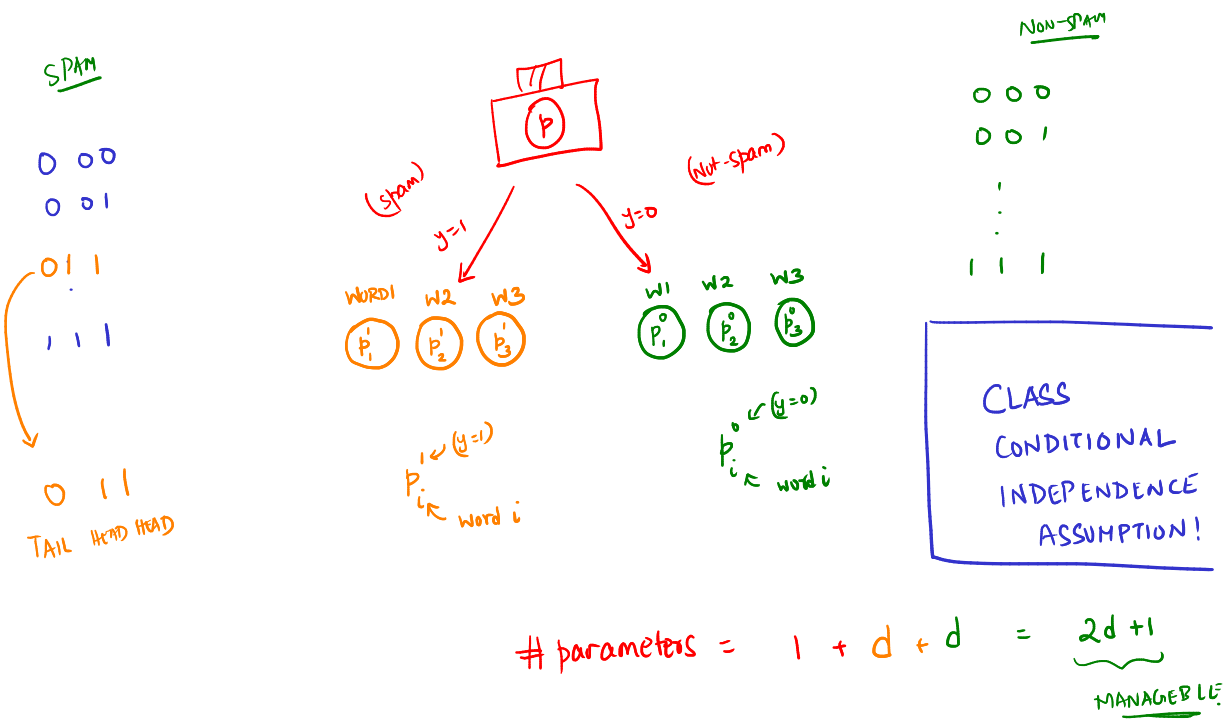
$(2^d - 1)$: $P(x|y=0)$

$= 1 + 2(2^d - 1)$

$= 2^{d+1} - 1$

ISSUE

- Too many parameters !
- Not a reasonable story.



parameters $= 1 + d + d = 2d + 1$

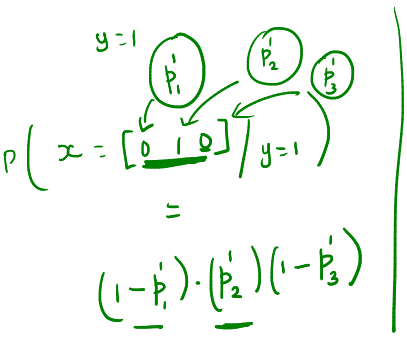
MANAGEBLE

STEP 1: $P(y=1) = p$

STEP 2: $P\left(x = [f_1 \; f_2 \; \cdots \; f_d] \,/\, y\right)$

$$= \prod_{i=1}^{d} \left(p_i^{y}\right)^{f_i} \left(1 - p_i^{y}\right)^{(1-f_i)}$$

FEATURES ARE CONDITIONALLY INDEPENDENT GIVEN LABEL



- How to estimate the parameters
- SOLUTION: MAXIMUM LIKELIHOOD!