G.D.A.

$P(x, y) = P(x|y)\, P(y)$ ← fraction of egs belonging to a particular class

$P(x, y)$: Model

$P(x|y)$: High dimensional

$P(y)$: Class Prior

$$P(y|x) = \frac{P(x|y)\, P(y)}{P(x)} = \frac{P(x|y)\, P(y)}{\left[P(x|y=0)P(y=0) + P(x|y=1)P(y=1)\right]}$$

$P(y|x)$: Posterior

Prediction

$$\hat{y} = \arg\max_y p(y|x)$$

$$= \arg\max_y \frac{p(x|y)\, p(y)}{p(x)}$$

$$= \arg\max_y \frac{p(x|y)\, p(y)}{p(x)}$$

$p(x)$ → prob. that you are going to encounter it

Two Algorithms

Both: $y \in \{0, 1\}$

G.D.A → $x \in \mathbb{R}^q$

Naive Bayes → x is discrete (text classification)

Model ⟺ Data Generating Process

(Joint $p(x, y)$)

Hierarchy of steps

=

Factorize our joint

G.D.A.

$y \sim Bern(\phi)$

$x|y=0 \sim N(\mu_0, \Sigma)$

$x|y=1 \sim N(\mu_1, \Sigma)$

} Data Generating Process

$$P(y) = \phi^y (1-\phi)^y$$

$$P(x|y=0) = \frac{1}{(2\pi)^{d/2}(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma^{-1}(x-\mu_0)\right)$$

$$P(x|y=1) = \frac{1}{(2\pi)^{d/2}(\Sigma)^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma^{-1}(x-\mu_1)\right)$$

Parameters $\rightarrow \Sigma, \mu_1, \mu_0, \phi$

Likelihood over parameters

$$L(\phi, \mu_0, \mu_1, \Sigma) = \log \prod_{i=1}^{n} P(x^{(i)}, y^{(i)})$$

This part is p.d.f finding specific (x, y)

$$= \log \prod_{i=1}^{n} p(x^{(i)} | y^{(i)}) \, p(y^{(i)})$$

$$\nabla \ell(\,) = 0 \rightarrow \hat{\phi}, \hat{\mu}_0, \hat{\mu}_1, \hat{\Sigma}$$

$$\hat{\phi} = \frac{1}{n} \sum_{i=1}^{n} 1\{y^{(i)} = 1\}$$ ← what fraction of eqs are 1

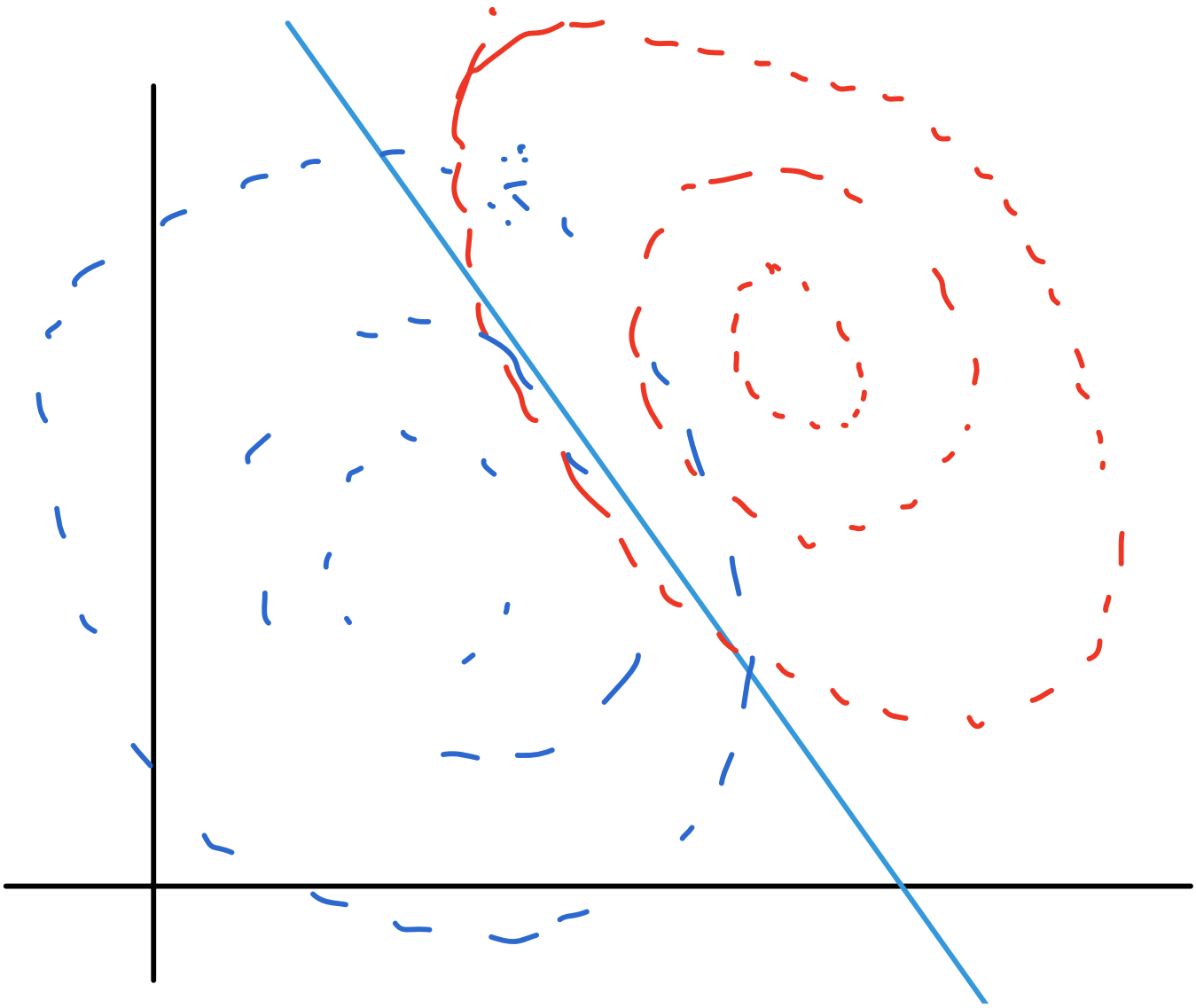
$$\hat{\mu}_0 = \frac{\sum_{i=1}^{n} 1\{y^{(i)} = 0\} \cdot x^{(i)}}{\sum_{i=1}^{n} 1\{y^{(i)} = 0\}}$$

$$\hat{\mu}_1 = \frac{\sum_{i=1}^{n} 1\{y^{(i)} = 1\} \cdot x^{(i)}}{\sum_{i=1}^{n} 1\{y^{(i)} = 1\}}$$

$$\Sigma = \frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$$

$$P(y=1|x) = \frac{1}{1+e^{-\theta^T x}}$$

$$\theta \rightarrow \mu_0, \mu_1, \phi, \Sigma$$



$P(x|y=0) = P(x|y=1)$

→ seperating hyperplane

GDA → Logistic Regression

(Σ same)

Principle axes of 2 ellipsoids are eigen vectors of covariance matrix

If unequal Σ, then nonlinear hyperplane (Logistic with non linear features)

When assumption of data
→ Gaussian, then G.D.A
works very well.
Logistic is more robust,
but when data is
less and we know
cond^n satisfied, GDA
better.

NAIVE BAYES

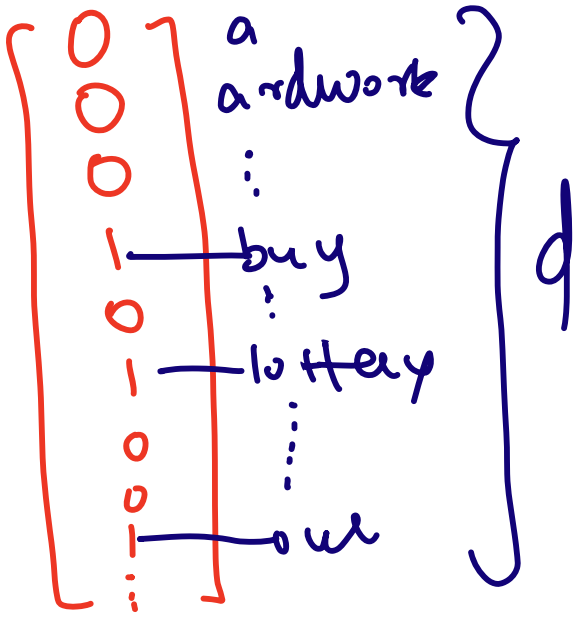
$x \rightarrow$ discrete

Text classification

$$P(x_j \mid x_k) = P(x_j)$$ Independence

$$P(x_j \mid x_k, y) = P(x_j \mid y)$$ [Condⁿ indep on y]

"Buy our lottery" =



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

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

Model: $p(y=1) \Rightarrow Bern(\phi_y)$ (1 parameter)

$p(x_j=1|y=0) \Rightarrow Bern(\phi_{j|y=0})$ (d parameters)

$p(x_j=1|y=1) \Rightarrow Bern(\phi_{j|y=1})$ (d parameters)

• $\rightarrow$ params

what fraction spammy/non spam

if email is spammy, x_j appears prob.

$$\ell(\phi_y, \phi_{j|y=0}, \phi_{j|y=1})$$

$$= \log \prod_{i=1}^{n} p(x^{(i)}, y^{(i)}, \phi)$$

$$[p(x,y) = p(y)p(x|y)$$

$$= p(y)p(x_1, x_2, x_3|y)$$

C.I.

$$= p(y)\,p(x_1|y)\,p(x_2|y)\,p(x_3|y)$$

$$= \log \prod_{i=1}^{n} p(y^{(i)}; \phi_y) \left(\prod_{j=1}^{d} p(x_j^{(i)} | y^{(i)}; \phi) \right)$$

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{n} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 1\}}{\sum_{i=1}^{n} 1\{y^{(i)} = 1\}}$$

↑ p that j^{th} word in vocab will show up in spammy email

$$\phi_{j|y=0} = \frac{\sum_{i=1}^{n} 1\{x_j^{(i)} = 1 \wedge y^{(i)} = 0\}}{\sum_{i=1}^{n} 1\{y^{(i)} = 0\}}$$

$$\phi_y = \frac{\sum_{i=1}^{n} 1\{y^{(i)} = 1\}}{n}$$

Prediction !!!

$$p(y=1|x) = \frac{p(x|y)\,p(y)}{p(x)}$$

$$= \frac{p(x|y)\,p(y)}{p(x|y=0)\,p(y=0) + p(x|y=1)\,p(y=1)}$$

$$= \prod_{j=1}^{d} \frac{p(x_j^{(i)}|y=1)\,p(y=1)}{\text{deno}}$$

$$p(y=0|x) = \frac{\text{"}\ y=0\ \text{"}\ y=0}{\text{deno}}$$

What if we encounter word not shown in training set?

Then $0/0$, $\Rightarrow$ indefinite

The method doesn't give prediction we use Laplace Smoothing

$x = \{0, 1\}$

$\phi = P(H) = \frac{10}{10}$

Count H	Count T
1	
1	
	0
1	
1	

$$\frac{\sum \text{Count } H}{\sum \text{Count } H + \sum \text{Count } T}$$

$P(H) = 1 \qquad P(T) = 0$

Laplace → add 1 to both initially

$$P(H) = \frac{10}{11} \qquad P(T) = \frac{1}{11}$$

Assume seen every word, once is spammy, once in non-spammy

Suppose $j = 2$

$x_2^{(i)} = 0 \quad \forall i$

$$\phi_{2|y=1} = \frac{0}{\#\text{positive examples}} = 0$$

$$\phi_{2|y=0} = \frac{0}{\#\text{negative ex}} = 0$$

$$p(x) = p(x|y=0)\,p(y=0) + p(x|y=1)\,p(y=1)$$

$$= \prod_{j=1}^{d} p(x_j|y=0)\; p(y=0)$$

$\leftarrow \prod_{j=1}^{d} p(x|y=1)p(y=1)$

$p(x_2=0|y=0)$

$\downarrow$

0

$p(x_2=1|y=0)$

$= \phi_{2|y=1} = 0$

$p(x) = 0$

But example does occur

$p(y=1|x) = \frac{p(x|y=1)p(y=1)}{p(x)}$

0

5

$z \sim \text{Bern}(\phi)$

n trials

$z^{(1)} = \text{tail}$

$\vdots$

$z^{(n)}$ } tail = 0

likelihood of ϕ

$$= (1-\phi)(1-\phi)\cdots\phi$$

$$= (1-\phi)^{\#\text{tails}}\,\phi^{\#\text{heads}}$$

$$\arg\max_{\phi} \ell(\phi) = \frac{\#\text{heads}}{\#\text{heads}+\#\text{tails}}$$

Suppose 3 draws, all tails

$$\phi = \frac{0}{0+3} = 0$$

So according to this, coin will never gives heads.

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{n} 1(x_j^{(i)}=1, y^{(i)}=1) + 1}{\sum 1(y^{(i)}=1) + 2}$$

$$\phi_{j|y=0} = \frac{1 + \sum_{j=1}^{n} 1\{x_j^{(i)}=1 | y^{(i)}=0\}}{2 + \sum 1\{y^{(i)}=0\}}$$

Multinomial Event Model

$y \sim Bern(\phi_y)$ — 1

$x|y=0 \sim Categorical(\phi_{k|y=0})$ $|V|-1$

$x_j \in \{1, \ldots, |V|\}$

MLE

$$\phi_{k|y=1} = \frac{\sum_{i=1}^{n} \sum_{j=1}^{d_i} 1\{x_j^{(i)} = k \wedge y^{(i)} = 1\}}{\sum_{i=1}^{n} 1\{y^{(i)} = 1\} \cdot d_i}$$

total words across messages in that class

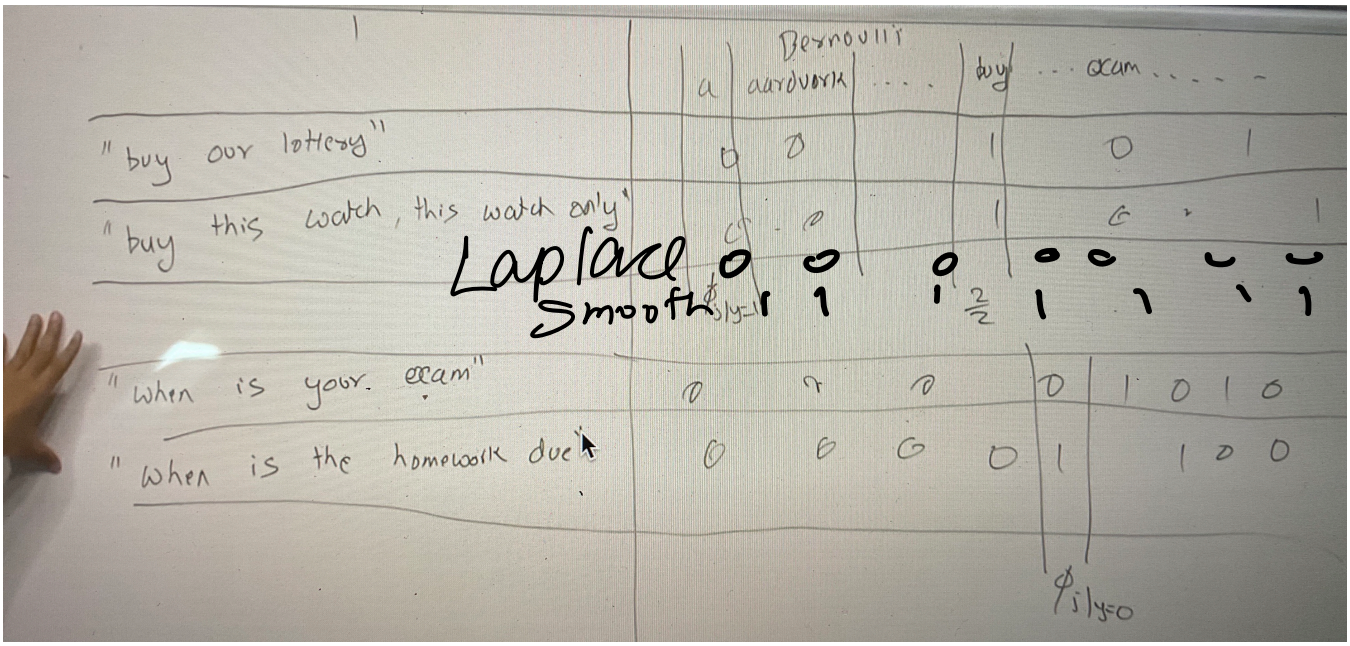
Sum over all words in all message and

count only those which are word & belong to our class

$\phi_{k|y=0}$ = replace above with $y=0$

Laplace Smooth

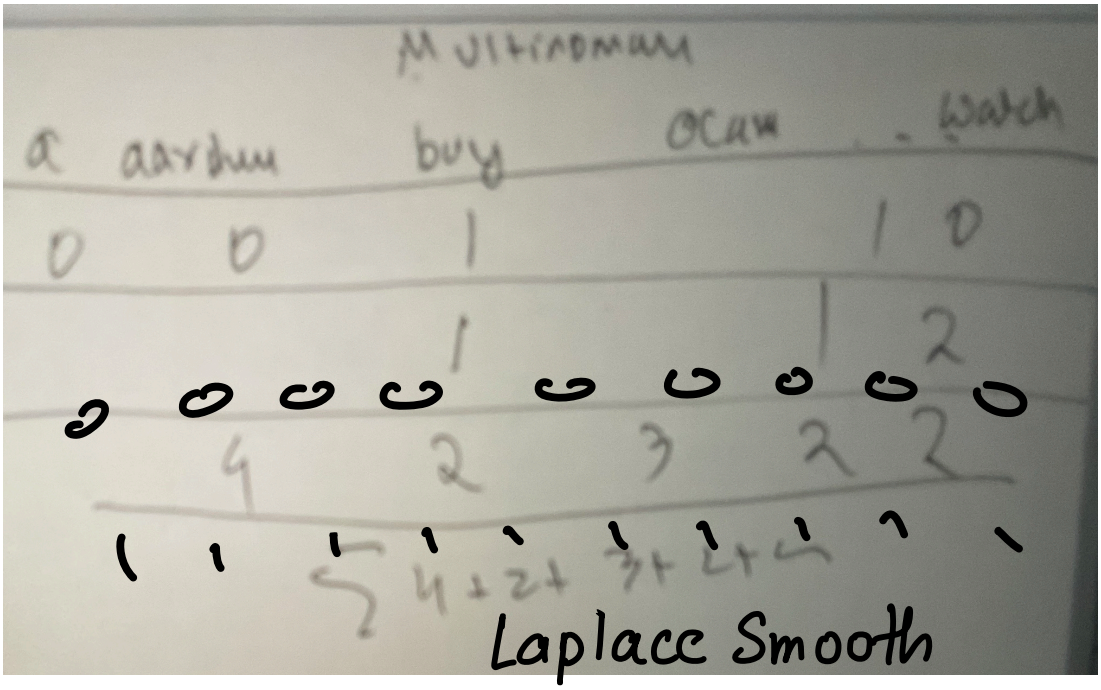
$$\frac{1 + \text{Num}}{|V| + \text{Deno}}$$



In Bernaulli, we normalize locally. We just see if it occurs / not.

Even if word occurs 10000 times or 1 time, it do the same.

In Multinomial, we normalize over the whole dataset & we count how many times each word appears



We can also think of it as concatenating over as a long sequence

Training $P(x, y)$

Prediction $P(y|x)$