→ Focus on x

→ Model $\underbrace{P(x, y)}_{\text{Model}} = \underbrace{P(x|y)}_{\substack{\text{High} \\ \text{dimensional.}}} \cdot \underbrace{P(y)}_{\substack{\text{class} \\ \text{Prior}}}$

[Till now we have modeled $P(y|x)$]

$$\underbrace{P(y|x)}_{\substack{\text{Posterior} \\ \text{distribution.}}} = \frac{P(x|y) \cdot P(y)}{P(x)} = \frac{P(x|y) \cdot P(y)}{P(x|y=0) \cdot P(y=0) + P(x|y=1) \cdot P(y=1)}$$

[when y is binary]

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

$$= \arg\max_y \frac{P(x|y) \cdot P(y)}{P(x)}$$

$$= \arg\max_y P(x|y) \cdot P(y)$$

Two algorithms

GDA: Gauss Discriminant Analysis

NB: Naive Bayes

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

GDA — $x \in \mathbb{R}^d$ (continuous)

NB — x is discrete (text classification)

Model (Joint $p(x,y)$) $\Longleftrightarrow$ Data Generating Process

Hierarchy of steps

$\Downarrow$

Factorise our joint.

GDA

$y \sim \text{Bernoulli}(\phi)$

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

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

$$p(y) = \phi^{y}(1-\phi)^{1-y}$$

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

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

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

$x \in \mathbb{R}^d$

$$\underbrace{p(x,y)}_{\text{Generative}} = \underbrace{p(y|x)}_{\text{Discr.}} \cdot p(x)$$

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

Max. Likelihood to learn parameters

Log likelihood

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

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

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

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

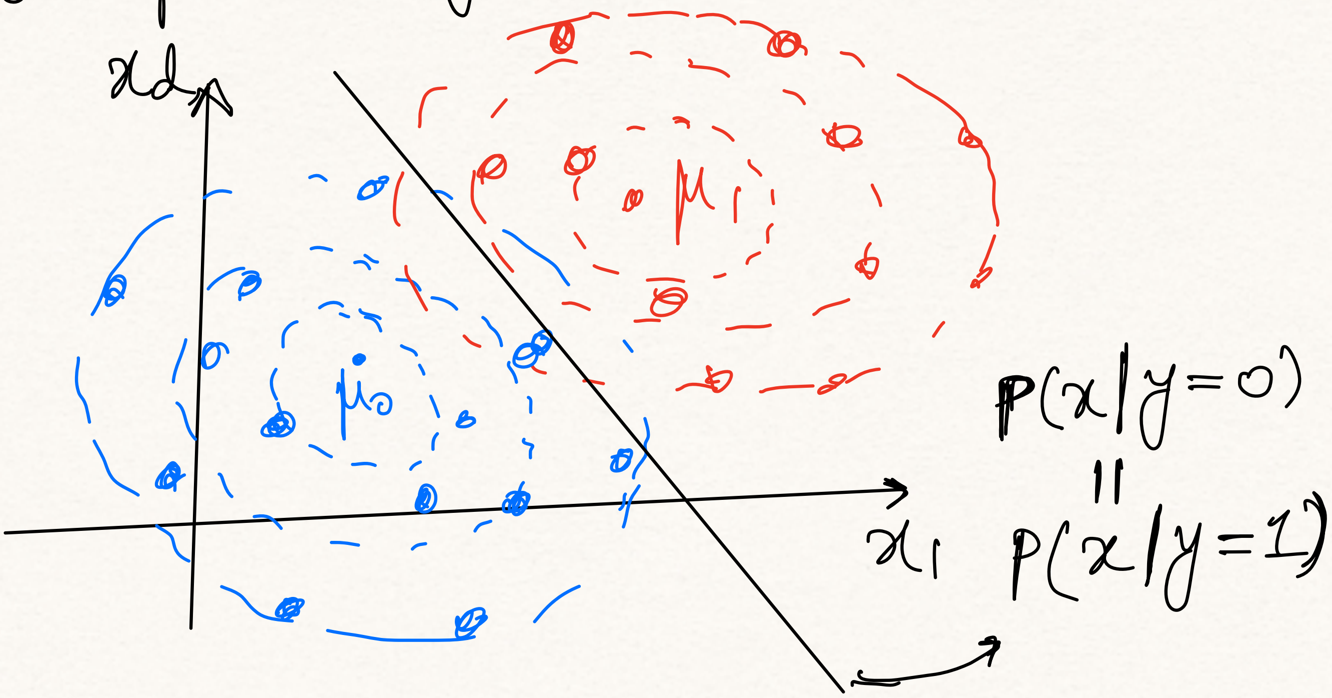
$$\hat{\mu}_0 = \frac{\sum_{i=1}^{n} 1\{y^{(i)} = 0\} . x^{(i)}}{\sum_{i=1}^{n} 1\{y^{(i)} = 0\}}$$

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

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

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

θ depends only on $\mu_0, \mu_1, \phi, \Sigma$



GDA $\Rightarrow$ Logistic Regression.

⊛ GDA more efficient than logistic reg.

If Σ was not same for both the distrib^n Then instead of str. line it could be a curve of degree 2/3/4 - - , etc.

Naive Bayes

x - discrete

Text classification (spam filters)

Conditional Independence

$$P(x_j | x_k) = P(x_j) \quad [\text{indep.}]$$

$$p(x_j | x_k, y) = p(x_j | y) \quad \begin{bmatrix} \text{conditional} \\ \text{indep.} \\ \text{on } y \end{bmatrix}$$

Bernoulli Event Model

$$\text{"Buy our lottery"} = \left.\begin{bmatrix} 0 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 1 \\ 0 \end{bmatrix} \underbrace{\begin{matrix} \text{a} \\ \text{aardvark} \\ \vdots \\ \text{buy} \\ \vdots \\ \\ \text{lottery} \\ \vdots \\ \\ \text{our} \\ \text{zygmorgy} \end{matrix}}_{\substack{\text{Vocabulary} \\ - d \text{ dim.}}} \right\} d$$

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

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

Model: $P(y=1) = \text{Bernoulli}(\phi_y)$ — 1

$P(\underset{''1}{x_j} | y=0) = \text{Bernoulli}(\phi_{j|y=0})$ — d

$$p(\underset{\overset{\shortparallel}{1}}{x_j} \mid y=1) = \text{Bernoulli}(\phi_{j|y=1}) - d$$

$$\ell(\underbrace{\phi_y}_{1}, \underbrace{\phi_{j|y=0}}_{d}, \underbrace{\phi_{j|y=1}}_{d}) = \log \prod_{i=1}^{n} p(x^{(i)}, y^{(i)}; \phi)$$

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

$$p(x_1, x_2, \dots, x_d \mid y) = p(x_1 \mid y) \cdot p(x_2 \mid x_1, y) \cdot p(x_3 \mid x_1, x_2, y) \dots$$

If conditionally independent $\Rightarrow p(x_1 \mid y) \cdot p(x_2 \mid y) \dots$

MLE

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

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

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

$$P(y|x) = \frac{P(x|y) \cdot P(y)}{P(x)} = \frac{P(x|y) \cdot P(y)}{P(x|y=0) \cdot P(y=0) + P(x|y=1) \cdot P(y=1)}$$

$$P(y=1|x) = \frac{\prod_{j=1}^{d} P(x_j^{(i)} | \underset{1}{\underset{\shortparallel}{y}}) \underset{1}{\underset{\shortparallel}{P(y)}}}{\text{denom.}}$$

PROBLEM !

For new words $\rightarrow P(y|x) = \frac{0}{0+0}$ $\rightarrow$ Thus, fails!!

$\rightarrow$ Can be corrected by Laplace smoothing.

Assume that each word has been seen once in each spam email as well as non-spam email.

Laplace smoothing ver.

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

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

Multinomial Event Model

$y \sim \text{Bernoulli}(\phi_y)$ — 1

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

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

$x^{(i)} \in \{1, \dots, |V|\}^{d_i}$

MLE

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

$$\phi_{k|y=0} = \underline{\qquad\qquad\qquad}$$

L.S ver

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