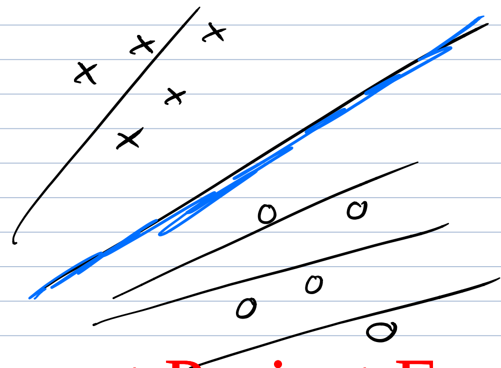
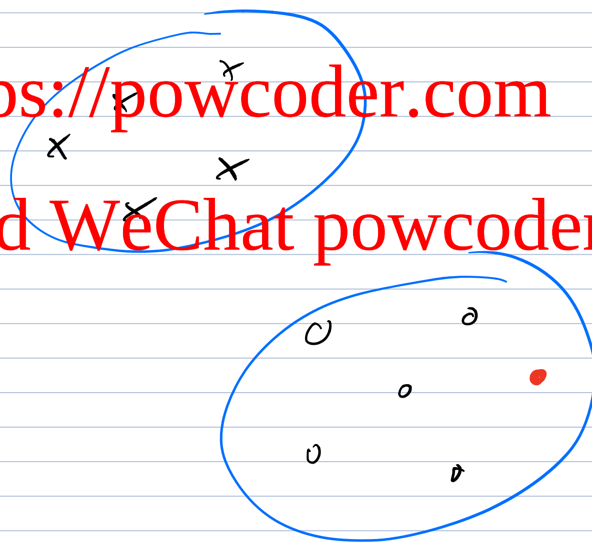
Generative Learning Algorithms

- **Gaussian Discriminant Analysis (GDA)**
- Generative & Discriminative Algorithms
- Naive Bayes





Discriminative

Learn $p(y|x)$ $\qquad x \rightarrow y$

or learn $h_\theta(x) = \begin{cases} 0 \\ 1 \end{cases}$

Generative Learning Algorithm

Learn $p(x \mid y)$ $p(y)$

feature   class

Bayes Rule:

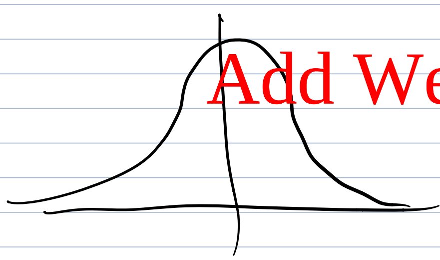
$$p(y=1 \mid x) = \frac{p(x \mid y=1) \cdot p(y=1)}{p(x)}$$

$$p(x) = p(x \mid y=1) \cdot p(y=1) + p(x \mid y=0) \cdot p(y=0)$$

Gaussian Discriminant Analysis (GDA)

Suppose $x \in \mathbb{R}^n$ (drop $x_0 = 1$ convention)

Assume $p(x \mid y)$ is Gaussian



$$z \sim \mathcal{N}(\underset{\mathbb{R}^n}{\vec{\mu}}, \underset{\mathbb{R}^{n \times n}}{\Sigma})$$

$z \in \mathbb{R}^n$ $(z_1, z_2 \ldots z_n)$

$$\mathbb{E}[z] = \mu$$

$$\mathrm{Cov}[z] = \mathbb{E}[(z-\mu)(z-\mu)^T]$$
$$= \mathbb{E}[zz^T] - (\mathbb{E}[z])(\mathbb{E}[z])^T$$
$$= \Sigma$$

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

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

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

Parameters: $\underset{\mathbb{R}^n}{\mu_0}, \underset{\mathbb{R}^n}{\mu_1}, \underset{\mathbb{R}^{n\times n}}{\Sigma}, \underset{\in[0,1]}{\phi}$

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

Training Set $\{(x^{(i)}, y^{(i)})\}_{i=1}^m$

Joint Likelihood

$$\mathcal{L}(\phi, \mu_0, \mu_1, \Sigma) = \prod_{i=1}^m p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma)$$

$$= \prod_{i=1}^m p(x^{(i)} \mid y^{(i)})\, p(y^{(i)})$$

Discriminative

$$\mathcal{L}(\theta) = \prod_{i=1}^m p(y^{(i)} \mid x^{(i)}; \theta)$$

(Conditional Likelihood)

Maximum Likelihood Estimation

$$\max_{\phi, \mu_0, \mu_1, \Sigma} \; \ell(\phi, \mu_0, \mu_1, \Sigma) = \log \mathcal{L}(\ldots)$$

$$\phi = \frac{\sum_{i=1}^{m} y^{(i)}}{m} = \frac{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)}=1\}}{m}$$

$\mathbb{1}\{\text{true}\} = 1 \qquad \mathbb{1}\{\text{false}\} = 0$

$$\mu_0 = \frac{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)}=0\}\, x^{(i)}}{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)}=0\}}$$

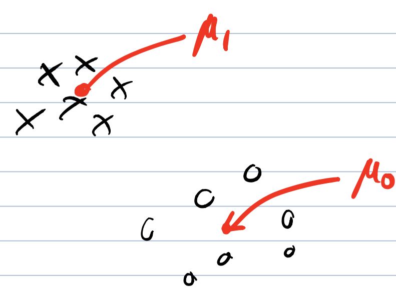
← sum of feature vectors for examples with $y=0$

← # examples with $y=0$

$$\mu_1 = \frac{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)}=1\}\, x^{(i)}}{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)}=1\}}$$

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

Covariance $= \mathbb{E}[(z-\mu)(z-\mu)^T]$

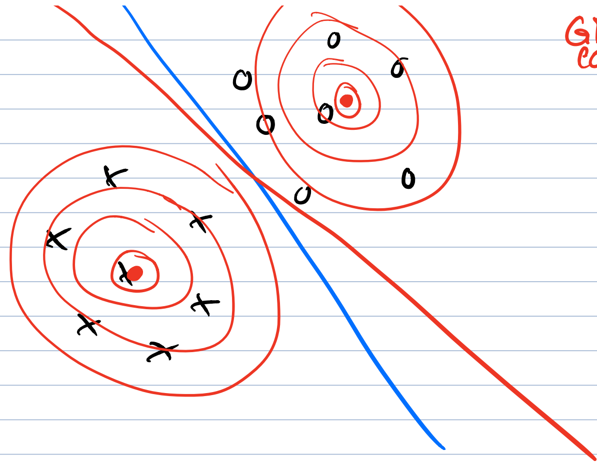


$(x^{(i)} - \mu_{y^{(i)}})$

Prediction:

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

eg. $\min_z (z-2)^2 = 0 \qquad \arg\min_z (z-2)^2 = 2$



GDA common covariance Σ

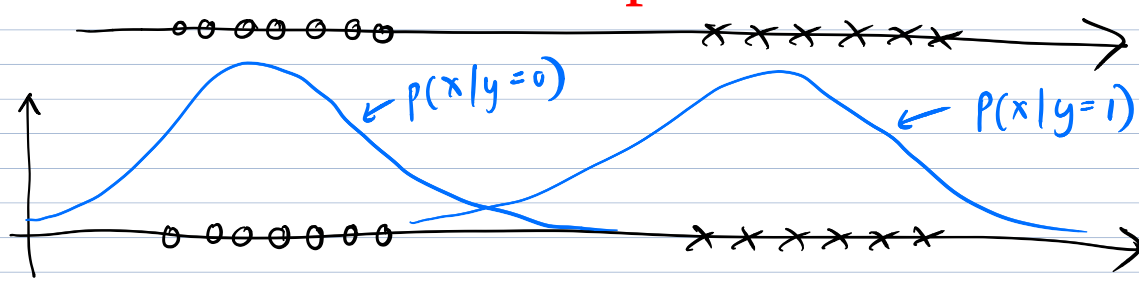
$\phi = \frac{1}{2}$

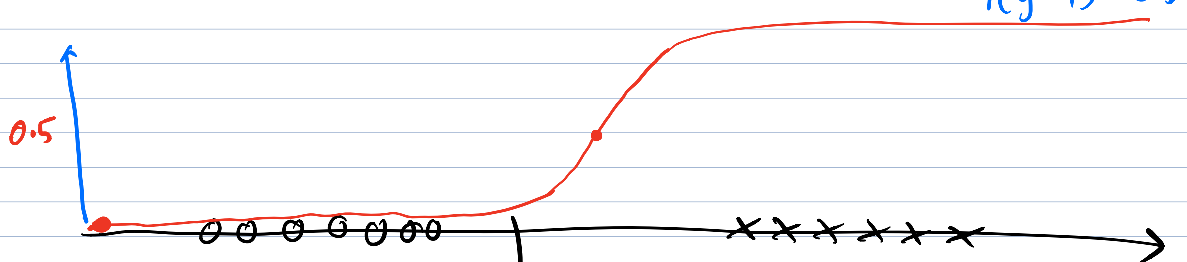
Comparison to Logistic Regression

For fixed $\phi, \mu_0, \mu_1, \Sigma$

plot $p(y=1 \mid x; \phi, \mu_0, \mu_1, \Sigma)$ as a f^n of x

$$= \frac{p(x \mid y=1; \mu_1, \Sigma)\, p(y=1; \phi)}{p(x; \phi, \mu_0, \mu_1, \Sigma)}$$





sigmoid f^n !

Generative	Discriminative
GDA assumes	Logistic Regression

$$\left.\begin{aligned} x \mid y=0 &\sim \mathcal{N}(\mu_0, \Sigma) \\ x \mid y=1 &\sim \mathcal{N}(\mu_1, \Sigma) \\ y &\sim \mathrm{Ber}(\phi) \end{aligned}\right\} \Longrightarrow \quad p(y=1 \mid x) = \frac{1}{1+e^{-\theta^T x}} \quad (\text{"}x_0 = 1\text{"})$$

logistic

$$\not\Longleftarrow$$

stronger assumption | weaker assumption

$$\left.\begin{aligned} x \mid y=1 &\sim \mathrm{Poisson}(\lambda_1) \\ x \mid y=0 &\sim \mathrm{Poisson}(\lambda_0) \\ y &\sim \mathrm{Ber}(\phi) \end{aligned}\right\} \longrightarrow p(y=1 \mid x) \text{ is logistic}$$

Naive Bayes

Feature vectors x ?

a
~~aardvark~~
~~aardwolf~~
⋮
buy
⋮
CS229
⋮
zymurgy

↑ top 10,000 ↓

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \vdots \\ 1 \\ \vdots \end{bmatrix} \begin{array}{l} \leftarrow \text{"a"} \\ \\ \\ \\ \leftarrow \text{"buy"} \\ \\ \end{array}$$

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

$x_i = 1$ if word i appears in email

Want to model $p(x|y)$, $p(y)$

$2^{10,000}$ possible values of x

$2^{10,000} - 1$

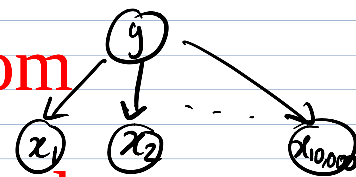
Assume x_i's are conditionally independent given y

$$P(x_1 \ldots x_{10,000}|y) = P(x_1|y) \cdot P(x_2 | x_1, y) \cdot P(x_3 | x_1, x_2, y) \cdots P(x_{10,000} | \ldots)$$

assume

$$= P(x_1|y) \, P(x_2|y) \, P(x_3|y) \cdots P(x_{10,000}|y)$$

CS228 Graphical models



$$= \prod_{i=1}^{n} P(x_i | y)$$

Parameters:

$\phi_{j|y=1} = P(x_j = 1 | y = 1)$

$\phi_{j|y=0} = P(x_j = 1 | y = 0)$

$\phi_y = P(y = 1)$

Joint Likelihood

$$\mathcal{L}(\phi_y, \phi_{i|y}) = \prod_{i=1}^{m} P(x^{(i)}, y^{(i)}; \phi_y \, \phi_{j|y})$$

MLE:
$$\phi_y = \frac{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)}=1\}}{m}$$

$$\phi_{j|y=1} = \frac{\sum_{i=1}^{m} \mathbb{1}\{x_j^{(i)}=1, y^{(i)}=1\}}{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)}=1\}}$$