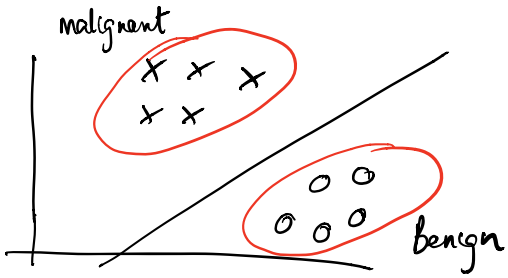
Generative Learning Algorithms

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



Discriminative Learning Algorithm

learns $p(y|x)$

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

$x \to y$

Generative Learning Algorithm

learns $p(x|y)$ $p(y)$

$p(x|y)$: features, class

$p(y)$: class prior

Bayes Rule:

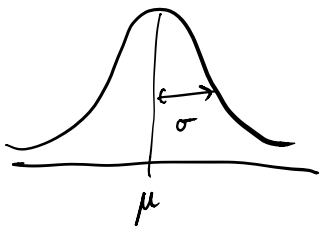
$$p(y=1|x) = \frac{p(x|y=1) \cdot p(y=1)}{p(x)}$$

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

Gaussian Discriminant Analysis (GDA)

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

Assume $p(x|y)$ is Gaussian



$Z \sim N(\vec{\mu}, \Sigma)$

$\vec{\mu} \in \mathbb{R}^d$, $\Sigma \in \mathbb{R}^{d \times d}$

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

$\mu_i = \mathbb{E}[Z_i]$

$\Sigma_{ij} = \mathbb{E}[Z_i Z_j] - \mathbb{E}[Z_i]\mathbb{E}[Z_j]$

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

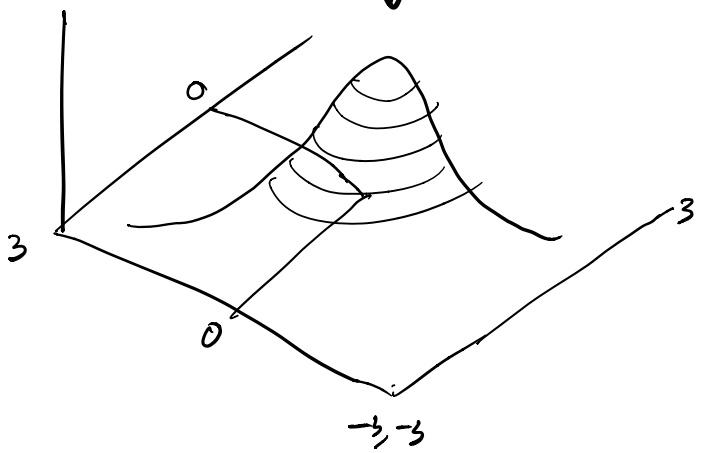
$\mathrm{Cov}(Z) = \mathbb{E}[(Z-\mu)(Z-\mu)^T]$

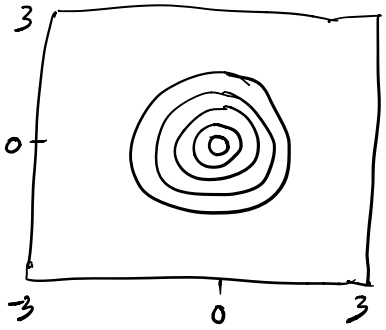
$= \mathbb{E}[ZZ^T] - (\mathbb{E}Z)(\mathbb{E}Z)^T$

$\mathbb{E}[Z] = \mathbb{E}Z$

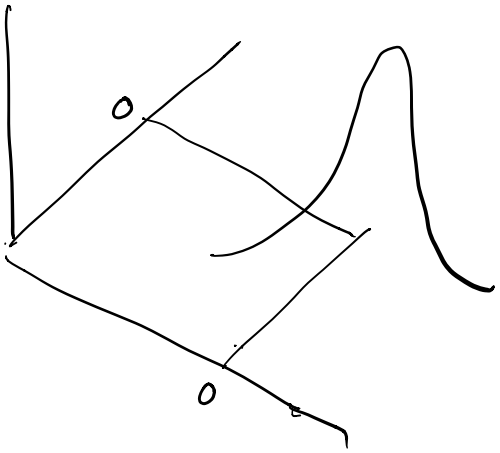
pdf $p(z) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$

$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

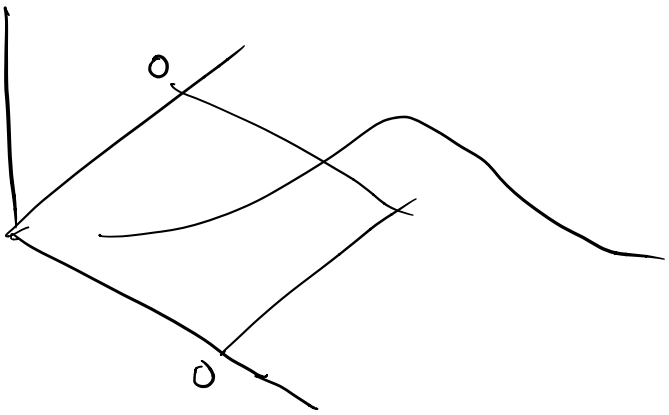




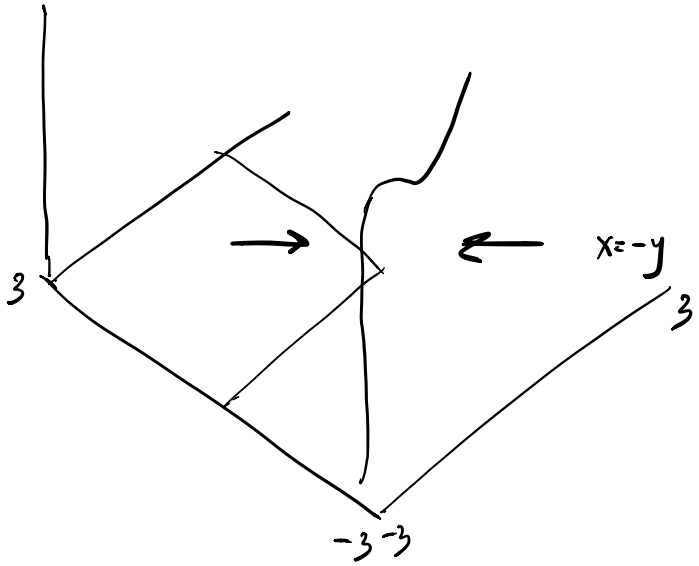
$\Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}$

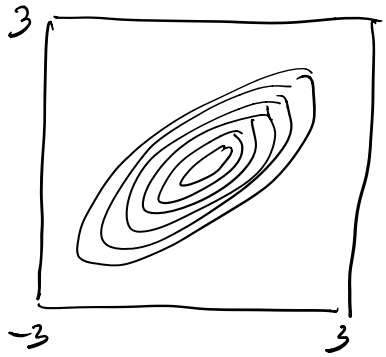


$$\Sigma = \begin{bmatrix} 2 & 0 \\ 0 & 2 \end{bmatrix}$$

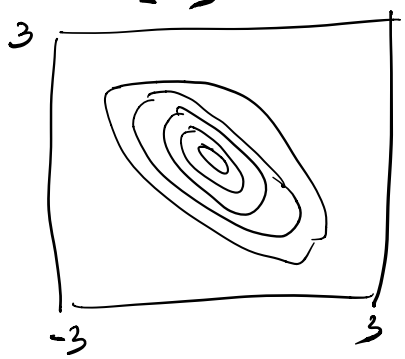


$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix}$$





$$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$



$$\mu = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \quad \begin{bmatrix} 1 \\ 2 \end{bmatrix} \quad \begin{bmatrix} -1 \\ -3 \end{bmatrix}$$

GDA model

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

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

Parameters: μ_0 μ_1 Σ, ϕ

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

μ_0, μ_1	Σ	ϕ
$\mathbb{R}^d$	$\mathbb{R}^{d \times d}$	$[0,1]$

$p(y=1 \mid x)$, $p(y=0 \mid x)$ via Bayes rule

Training set $\{(x^{(i)}, y^{(i)})\}_{i=1}^n$

Joint likelihood fn:

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

$$= \prod_{i=1}^n p(x^{(i)} \mid y^{(i)}) \cdot p(y^{(i)})$$

Cost fn is $\int$ fn of x, y

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

Maximum Likelihood Estimator

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

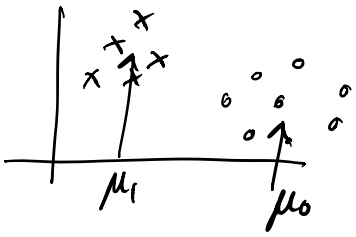
$$\phi = \frac{\sum_{i=1}^{n} y^{(i)}}{n}$$

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

$\mathbb{1}\{\text{true}\} = 1$

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

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



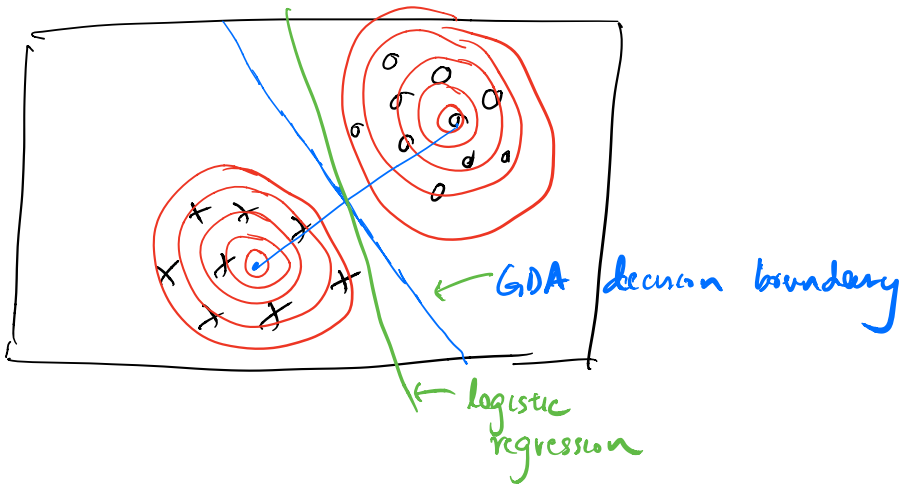
$$\mu_1 = \frac{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)} = 1\} x^{(i)}}{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)} = 1\}}$$

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

Prediction:

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

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

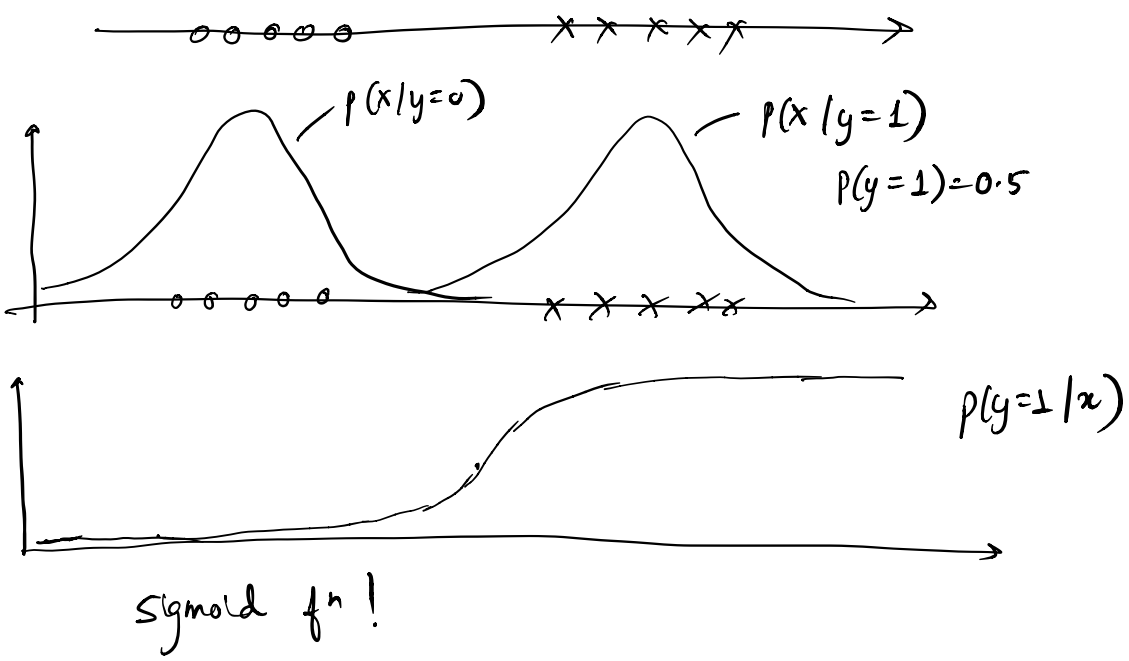


Companion to logistic regression

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

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

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



sigmoid fn !

generative

GDA assumes

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

$x \mid y=0 \sim N(\mu_1, \Sigma)$

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

$\Rightarrow$ $\not\Leftarrow$

Stronger assumption

discriminative

Logistic Regression

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

("$x_0 = 1$")

weaker assumption

$$\left.\begin{array}{l} x|y=1 \sim \text{Poisson}(\lambda_1) \\ x|y=0 \sim \text{Poisson}(\lambda_0) \\ y \sim \text{Ber}(\phi) \end{array}\right\} \Rightarrow \quad p(y=1|x) \text{ is logistic}$$

Naive Bayes

Feature vector x?

English dict

$$x = \begin{bmatrix} 1 \\ 0 \\ 0 \\ \\ \\ 1 \\ \\ \\ \end{bmatrix} \begin{array}{l} a \\ aardvark \\ aardwolf \\ \vdots \\ \\ buy \\ c5229 \\ \vdots \\ zymurgy \end{array}$$

top 10k

$x \in \{0,1\}^d \qquad d = 10{,}000$

$x_i = 1\{\text{word } i \text{ appears in email}\}$

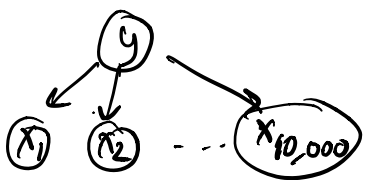
Want to model $p(x|y)$, $p(y)$

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

Assume x_i's are conditionally independent given y

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

$$\overset{\text{assume}}{=} p(x_1|y) \cdot p(x_2|y) \cdot p(x_3|y) \cdots p(x_{10,000}|y)$$



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

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

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

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

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