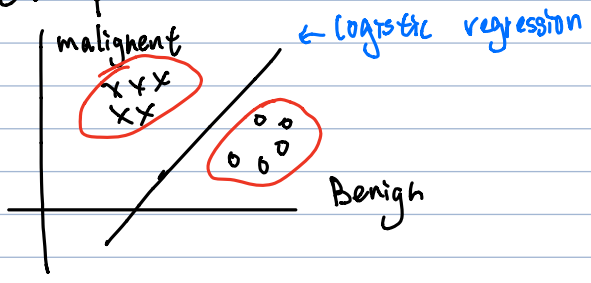
Generative Learning Algorithms

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

Example: tumor detection



Discriminative Learning Algorithm

learns $p(y|x)$

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

$x \rightarrow y$

Generative Learning Algorithm

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

$P(x|y)$: features | class

$p(y)$: class prior.

$P(x|y)$, $P(y)$ generative

$P(y|x)$, discriminative

Bayes Rule:

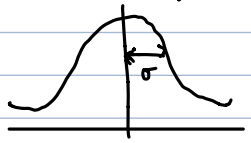
$$p(y=1|x) = \frac{P(x|y=1)\,p(y=1)}{p(x)}$$

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

Gaussian Discriminant Analysis (GDA)

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

Assume $p(x|y)$ is Gaussian



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

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

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

$\mu_i = E[z_i]$

$\Sigma_{ij} = E[(z_i - \mu_i)(z_j - \mu_j)] = E[z_i z_j] - E[z_i]E[z_j]$

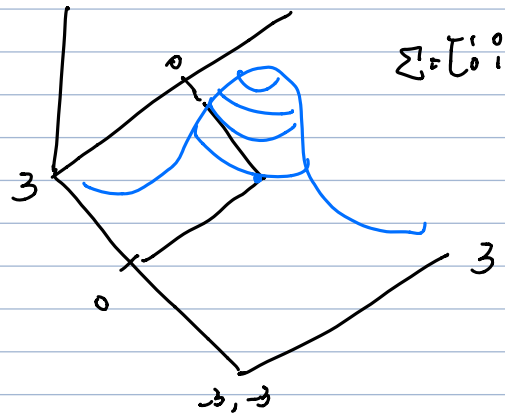
μ (covariance matrix)

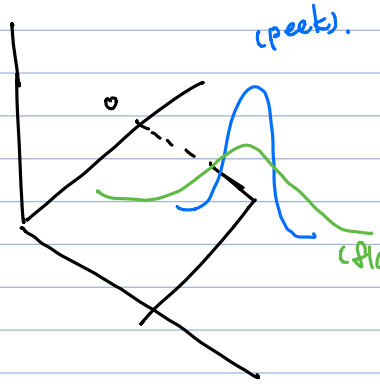
$$\mathbb{E}[Z] = \mu$$

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

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

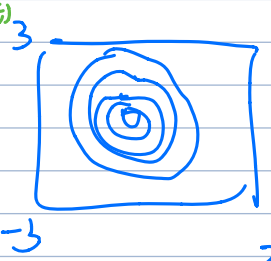
PDF. $p(Z) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)\right)$

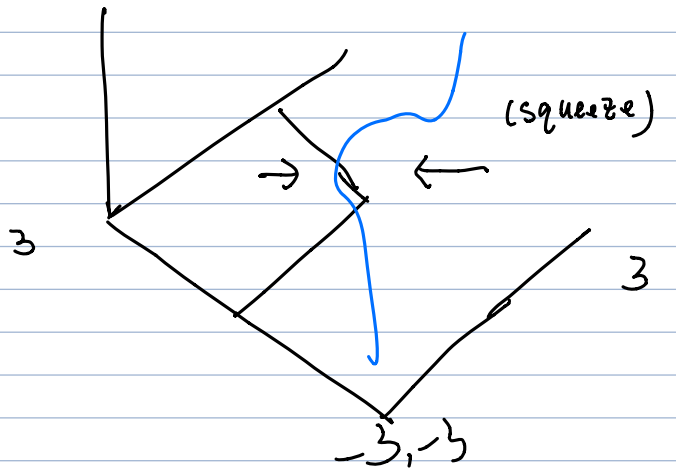




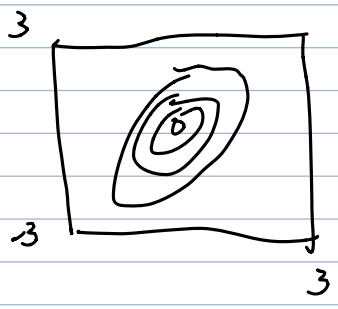
$\Sigma = \begin{bmatrix} 0.6 & 0 \\ 0 & 0.6 \end{bmatrix}$

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





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



3

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

GDA model

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

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

Parameters: μ_0, μ_1, Σ, ϕ (use same cov matrix)

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

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

Find $P(y=1|x)$, $P(y=0|x)$ by Bayes rule.

How to fit the parameters?

- Training set. $\{(x^{(i)}, y^{(i)})\}_{i=1}^{n}$
- Join likelihood func:

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

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

Cost fn. is joint fn. of x, y

Discriminative: $\mathcal{L}(\theta) = \prod_{i=1}^{n} P(y^{(i)}|x^{(i)}; \theta)$ (Diff. from Generative)

Maximum Likelihood Estimator

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

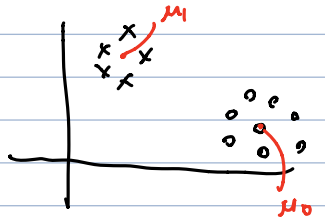
$$\Rightarrow \phi = \frac{\sum_{i=1}^{n} y^{(i)}}{n} = \frac{\sum_{i=1}^{n} \mathbb{1}\{y^{(i)} = 1\}}{n}$$

[Indicator fn. $\mathbb{1}\{\text{true}\} = 1$, $\mathbb{1}\{\text{false}\} = 0$]

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

(mean of all examples labeled 0)

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



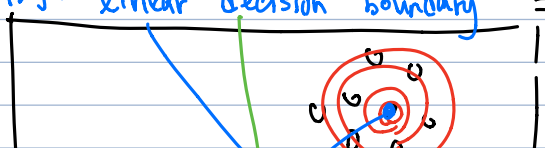
$$\Sigma = \frac{1}{n} \sum_{i=1}^{n} (x^{(i)} - \mu_{y^{(i)}})(x^{(i)} - \mu_{y^{(i)}})^T$$

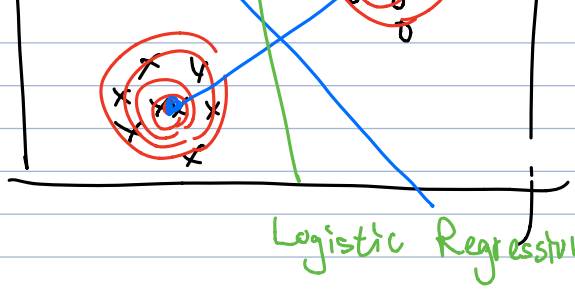
Prediction

$$\arg\max_y P(y|x) = \arg\max_y \frac{P(x|y)P(y)}{P(x)} \quad \text{(drop)}$$

$$= \arg\max P(x|y)P(y)$$

GDA: linear decision boundary $\Rightarrow$ Reason: Same Σ





Logistic Regression

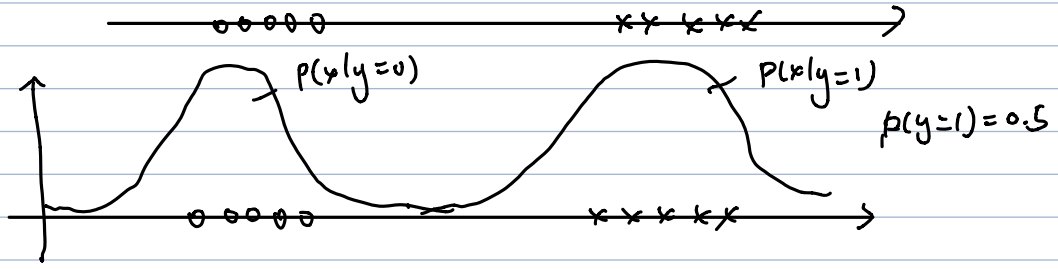
Comparison (w/ logistic regression)

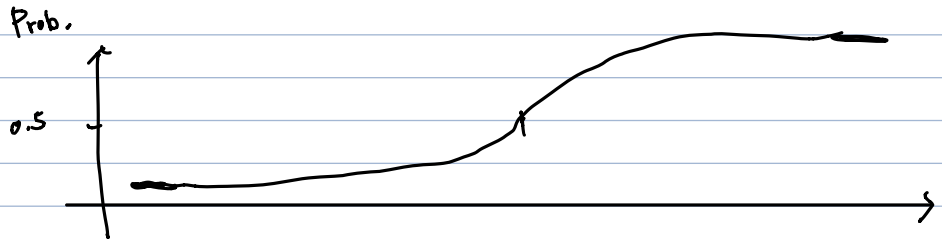
for fixed $\phi, \mu_0, \mu_1, \Sigma$

plot($y = 1 | x; \phi, \mu_0, \mu_1, \Sigma$) as a fn. of x

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

Suppose a 1-D example





sigmoid fn. !

generative (Does better if data IS multi-Gaussian) (Efficient)

GDA assumes

$$x|y=0 \sim N(\mu_0, \Sigma)$$
$$x|y=1 \sim N(\mu_1, \Sigma)$$
$$y \sim Ber(\phi)$$

$\Rightarrow$ $\not\Leftarrow$

discriminative (more data, use it! More robust)

Logistic Regression

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

($x_0 = 1$)

Stronger assumption (Performs less strong)

$$x|y=1 \sim Poisson(\lambda_1)$$
$$x|y=1 \sim Poisson(\lambda_2)$$
$$y \sim Ber(\phi)$$

$\Rightarrow$ $P(y=1|x)$ is logistic

Weaker assumption. (Do well!)

works for all generalized linear methods

Naive Bayes

Feature vector x?

English dict

$$x = \begin{bmatrix} a \\ apple \\ \vdots \\ zoo \end{bmatrix} = \begin{bmatrix} 1 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix} \quad d = 10{,}000$$

(one-hot vec)

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

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

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

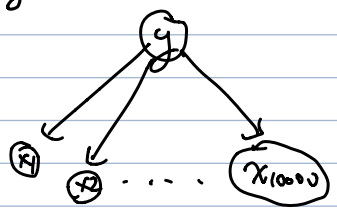
2^{10000} possible value of x.

Key assumption: x_i's are conditionally independent given y

$$p(x_1, \dots, x_d) = p(x_1|y)\, p(x_2|x_1, y) \cdots p(x_d|x_1, x_2 \dots y)$$

assume $p(x_1|y)\, p(x_2|y) \cdots p(x_d|y)$

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



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

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

n_y

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

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

Prove the GDA has linear decision boundary

Let $p \triangleq (\phi, \mu_0, \mu_1, \Sigma)$

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

$$= \frac{p(x|y=1;\mu_1,\Sigma)\phi}{p(x|y=1;\phi,\mu_1,\Sigma)\phi + p(x|y=0;\phi,\mu_2,\Sigma)(1-\phi)} = \frac{p(x|y=1,P)}{p(x|y=1,P)\phi + p(x|y=0,P)(1-\phi)}$$

$$= \frac{1}{1 + \left(\frac{1-\phi}{\phi}\right)\left(\frac{p(x|y=0,P)}{p(x|y=1,P)}\right)} \leftarrow r$$

$$r = \frac{\exp\left(-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0)\right)}{\exp\left(-\frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right)} = \exp\left[-\frac{1}{2}(x-\mu_0)^T\Sigma^{-1}(x-\mu_0) + \frac{1}{2}(x-\mu_1)^T\Sigma^{-1}(x-\mu_1)\right]$$

$$= \exp\left[\frac{1}{2}\left(-(x^T\Sigma^{-1} - \mu_0^T\Sigma^{-1})(x-\mu_0) + (x^T\Sigma^{-1} - \mu_1^T\Sigma^{-1})(x-\mu_1)\right)\right]$$

$$= \exp\left(\frac{1}{2}\left(-x^T\Sigma^{-1}x + x^T\Sigma^{-1}\mu_0 + \mu_0^T\Sigma^{-1}x - \mu_0^T\Sigma^{-1}\mu_0 + x^T\Sigma^{-1}x - x^T\Sigma^{-1}\mu_1 - \mu_1^T\Sigma^{-1}x + \mu_1^T\Sigma^{-1}\mu_1\right)\right)$$

$$= \exp\left(\frac{1}{2}\left((\mu_0^T - \mu_1^T)\Sigma^{-1}x + x^T\Sigma^{-1}(\mu_0-\mu_1) - \mu_0^T\Sigma^{-1}\mu_0 + \mu_1^T\Sigma^{-1}\mu_1\right)\right)$$

$$= \exp\left(\frac{1}{2}\left(2(\mu_0^T-\mu_1^T)\Sigma^{-1}x - \mu_0^T\Sigma^{-1}\mu_0 + \mu_1^T\Sigma^{-1}\mu_1\right)\right)$$

$$p(y=1|x,\phi,\mu_0,\mu_1,\Sigma) = \frac{1}{1+\exp\left(\log\left(\frac{1-\phi}{\phi}\right)\right)\exp\left((\mu_0-\mu_1)^T\Sigma^{-1}x - \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + \frac{1}{2}\mu_1^T\Sigma\mu_1\right)}$$

$$= \frac{1}{1+\exp\left((\mu_0-\mu_1)^T\Sigma^{-1}x - \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 + \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \log\left(\frac{1-\phi}{\phi}\right)\right)}$$

$$= \frac{1}{1+\exp\left(-\left((\mu_1-\mu_0)^T\Sigma^{-1}x + \frac{1}{2}\mu_0^T\Sigma^{-1}\mu_0 - \frac{1}{2}\mu_1^T\Sigma^{-1}\mu_1 + \log\left(\frac{\phi}{1-\phi}\right)\right)\right)}$$

let $\theta^T = (\mu_1 - \mu_0)^T \Sigma^{-1}$, $\theta_0 = \frac{1}{2}(\mu_0^T \Sigma^{-1} \mu_0 - \mu_1^T \Sigma^{-1} \mu_1) + \log\left(\frac{\phi}{1-\phi}\right)$

then

$$p(y=1 \mid x, \phi, \mu_0, \mu_1, \Sigma) = \frac{1}{1 + \exp(-(\theta^T x + \theta_0))}$$