COL774
Machine Learning
September 5 - 2021

Happy Teachers day.

GLMs:- Generalized Linear Models

Last class:- GDA I — Gaussian Discriminant Analysis

Generative story

$y^{(i)} | x^{(i)}; \theta \sim \text{Bernoulli}(\phi^{(i)})$

$\phi^{(i)} = \frac{1}{1+e^{-\theta^T x^{(i)}}}$

$(x^{(i)}, y^{(i)})$

$y^{(i)} \sim \text{Bernoulli}(\phi)$;

$x^{(i)} | y^{(i)}; \theta \sim \text{DIST}$ — Normal DIST

Discriminative

Generative Modelling — Both x & y

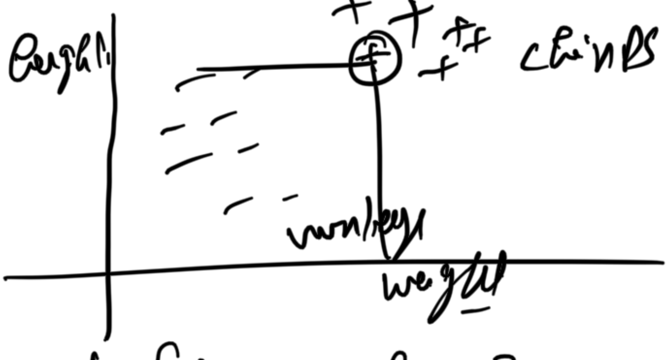
likelihood — examples are i.i.d.

$$\prod_{i=1}^{m} P(x^{(i)}, y^{(i)}; \theta) = \prod_{i=1}^{m} P(y^{(i)}; \theta)\, P(x^{(i)} | y^{(i)}; \theta)$$

(a) → (a) GLOB of (b) To generate attributes

log likelihood

$$\prod_{i=1}^{m} P(y^{(i)} | x^{(i)}; \theta)$$



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

$x|y \sim$? Normal(μ, Σ) — value, matrix

Background of Normal (multivariate) Distributions: I

X_1, X_2: Random Variables

$E[X_1]$ $E[X_2]$:- expected values

$= E[(X_1 - E[X_1])(X_2 - E[X_2])]$

$$\mathrm{Var}(X_1) := E[(X_1 - E[X])(X_1 - E[X])]$$

$$= E[X_1^2] - (E[X_1])^2 \geq 0$$

$$\mathrm{Cov}(X_1, X_2) = E[(X_1 - E[X_1])(X_2 - E[X_2])]$$

$$= E[X_1 X_2] - E[X_1]E[X_2]$$

$X_1 = X_2$:- $\mathrm{Cov}(X_1, X_2) \equiv \mathrm{Var}(X_1) = \mathrm{Var}(X_2)$

$X \in \mathbb{R}^n$ $\quad x = (x_1, \ldots, x_n)$

Generalized to Co-variance Matrix

$$\Sigma \in \mathbb{R}^{n \times n}$$

$$\Sigma_{jk} = \mathrm{Cov}(X_j, X_k)$$

H.W. to verify

$$\Sigma = E[(X - E[X])(X - E[X])^T]$$

$\Downarrow$ outer product of two vector

Multivariate Normal - Distribution:

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

$$x \sim N(\mu, \Sigma)$$

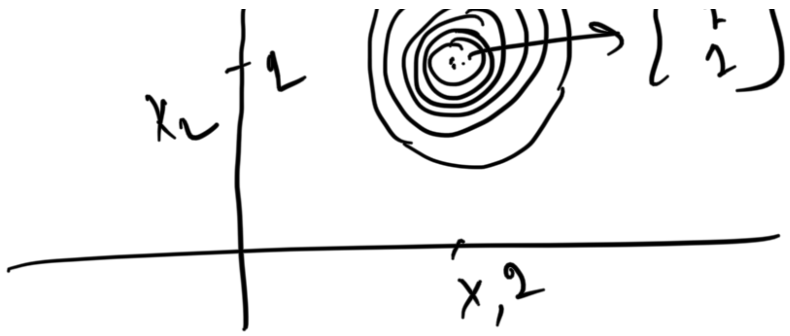
Σ: Covariance Matrix $\in \mathbb{R}^{n \times n}$

$\mu \Downarrow$ mean vector $\in \mathbb{R}^n$

$n = 1$ $\quad \sim N(\mu, \sigma^2)$

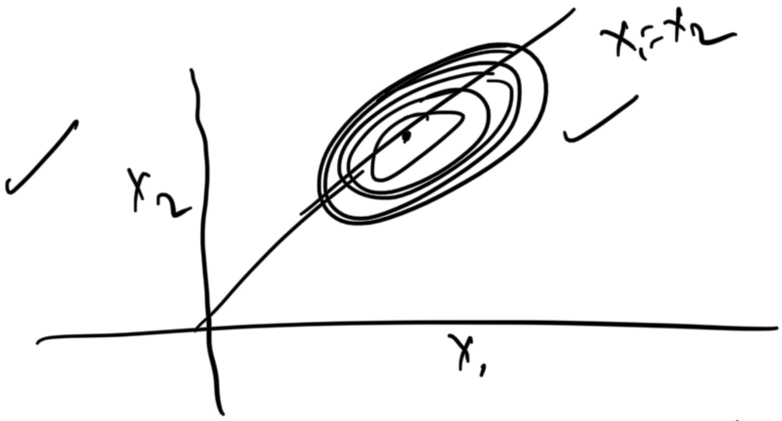
$X = (x_1, x_2)$

$X \sim N(\mu, \Sigma)$ $\qquad \Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$

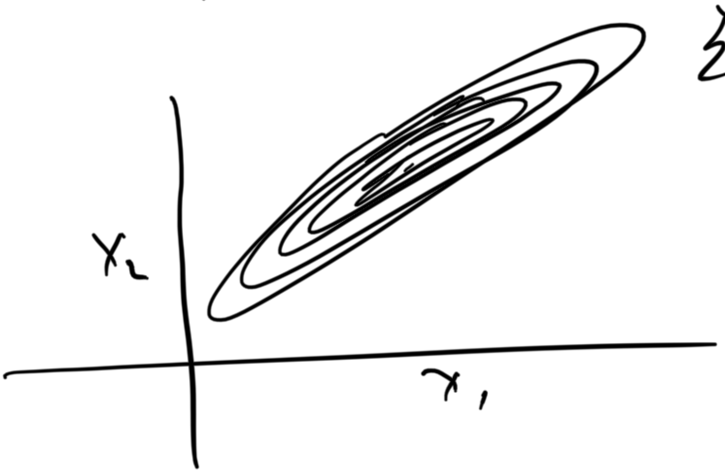


$\mu = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$\Rightarrow$ symmetric, the same dispersion



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



$\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$

$\Downarrow$

$\mu = \begin{bmatrix} 1 \\ 2 \end{bmatrix}$

$\Sigma = \begin{bmatrix} 3 & 0.5 \\ 0.5 & 1 \end{bmatrix}$

H.W.

$\Sigma = \begin{bmatrix} 3 & 0 \\ 0 & 1 \end{bmatrix}$

$\mathbb{R}^n$ $\quad x \in \mathbb{R}^n \quad \mathbb{R}^{n \times n}$

$x \sim N(\mu, \Sigma)$ Generalizes the expression for univariate normal distribution

$p(x) =$

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

Quadratic

univariate case

$$\rightarrow \quad p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

GDA:- Gaussian Discriminant Analysis } 2-class

↳ Generative Model

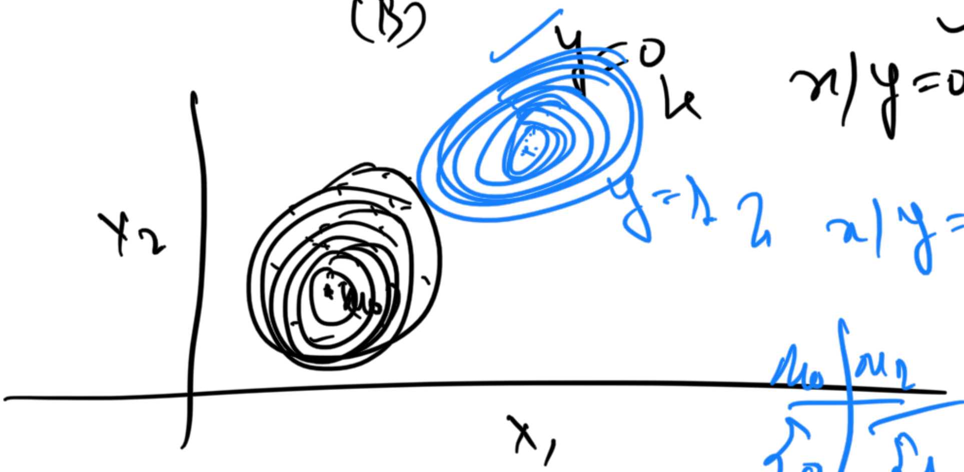
(x, y)

↳ (A) $y \sim$ Bernoulli(ϕ)

$y = 0$ (M)

$y = 1$ (C)

(B)



$x | y=0 \sim N(\mu_0, \Sigma_0)$

$x | y=1 \sim N(\mu_1, \Sigma_1)$

$\mu_0 \neq \mu_1$

μ_0	μ_1
Σ_0	Σ_1

→ Case:- $\Sigma_0 = \Sigma_1 = \Sigma$

⇓

Linear separator

Case 2:-

$\Sigma_0 \neq \Sigma_1$

↳ Quadratic sep.

class priors

$$\Theta = (\phi, \mu_0, \Sigma_0, \mu_1, \Sigma_1)$$

$\mu_0, \mu_1 \in R^n$, $\Sigma_0, \Sigma_1 \in R^{n \times n}$

Model parameters

$y \sim$ Bernoulli(ϕ)

GDA special case of
logistic reg
$\hat{\Sigma}_0 = \hat{\Sigma}_1 = \hat{\Sigma}$

Log likelihood

$$\underset{\Theta}{\arg\max}\ L(\theta) = \log\left[\prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \theta)\right] = \log \prod_{i=1}^{m} \left[p(y^{(i)}; \theta) \cdot p(x^{(i)} | y^{(i)}; \theta)\right]$$

$$= \sum_{i=1}^{m} \left[\log p(y^{(i)}; \theta) + \log p(x^{(i)} | y^{(i)}; \theta)\right]$$

$$p(y^{(i)}; \phi) = \begin{cases} \phi & \text{if } y^{(i)} = 1 \\ (1-\phi) & y^{(i)} = 0 \end{cases}$$

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

log

$$\theta = (\phi, \mu_0, \hat{\Sigma}_0, \mu_1, \hat{\Sigma}_1)$$

$$p(x^{(i)} | y^{(i)} = 1; \mu_1, \hat{\Sigma}_1) = \frac{1}{(2\pi)^{n/2} |\hat{\Sigma}_1|^{1/2}} e^{-\frac{(x^{(i)} - \mu_1)^T \hat{\Sigma}_1^{-1} (x^{(i)} - \mu_1)}{2}}$$

$$p(x^{(i)} | y^{(i)} = 0; \mu_0, \hat{\Sigma}_0) = \frac{1}{(2\pi)^{n/2} |\hat{\Sigma}_0|^{1/2}} e^{-\frac{(x^{(i)} - \mu_0)^T \hat{\Sigma}_0^{-1} (x^{(i)} - \mu_0)}{2}}$$

$$\Theta = (\phi, \mu_0, \hat{\Sigma}_0, \mu_1, \hat{\Sigma}_1)$$

ϕ

$\sum^{m} y^{(i)} \log p(y^{(i)} = 1)$ ... $\log p(y^{(i)} = 0)$

$$LL(\theta) = \sum_{i=1}^{m} \log [p(\underbrace{y^{(i)}}_{y^{(i)}}; \phi] \equiv \sum_{i=1} \; + (1-y^{(i)}) \log (1-\phi)$$

$$\theta$$

$$+ \sum_{i=1}^{m} \underbrace{1\{y^{(i)}=1\}} \log p(x^{(i)} | y^{(i)}=1; \mu_1, \Sigma_1)$$

$$\theta$$

$$+ \sum_{i=1}^{m} \underbrace{1\{y^{(i)}=0\}}_{(1-y^{(i)})} \log p(x^{(i)} | y^{(i)}=0; \mu_0, \Sigma_0)]$$

$$= \sum_{i=1}^{m} [y^{(i)} \log \phi + (1-y^{(i)}) \log (1-\phi)]$$

$$+ \sum_{i=1}^{m} (y^{(i)}) \log \frac{1}{(2\pi)^{n/2} |\Sigma_1|^{1/2}} e^{-\frac{(x^{(i)}-\mu_1)^T \Sigma_1^{-1} (x^{(i)}-\mu_1)}{2}}$$

$$+ \sum_{i=1}^{m} (1-y^{(i)}) \log \frac{1}{(2\pi)^{n/2} |\Sigma_0|^{1/2}} e^{-\frac{(x^{(i)}-\mu_0)^T \Sigma_0^{-1} (x^{(i)}-\mu_0)}{2}}$$

$$LL(\theta) = T1 \; \sum_{i=1}^{m} [y^{(i)} \log \phi + (1-y^{(i)}) \log (1-\phi)]$$

$$T2 + \sum_{i=1}^{m} y^{(i)} [\log \frac{1}{(2\pi)^{n/2} |\Sigma_1|^{1/2}} - \frac{(x^{(i)}-\mu_1)^T \Sigma_1^{-1} (x^{(i)}-\mu_1)}{2}]$$

$$T3 + \sum_{i=1}^{m} (1-y^{(i)}) [\underbrace{\log \frac{1}{(2\pi)^{n/2} |\Sigma_0|^{1/2}}} - \underbrace{\frac{(x^{(i)}-\mu_0)^T \Sigma_0^{-1} (x^{(i)}-\mu_0)}{2}}]$$

$\Rightarrow$ Maximize $LL(\theta)$ 고

$$\nabla_\theta LL(\theta) = 0$$

That parameter can be computed analytically.

$$\nabla_{\underbrace{\phi, (\mu_0, \Sigma_0), (\mu_1, \Sigma_1)}} [LL(\theta)] = 0$$

H.W. :-

$$\phi = \frac{\sum_{i=1}^{m} y^{(i)}}{m} \quad :- \quad \frac{\#\text{ of times } y^{(i)}=1 \text{ in the data}}{\#\text{ of examples}}$$

$1\{y^{(i)}=0\}$

$$\mu_0 = \frac{\sum_{i=1}^{m} (1-y^{(i)}) x^{(i)}}{\sum_{i=1}^{m} (1-y^{(i)})} \begin{array}{l} \Rightarrow \text{sample mean of examples belonging of class } \underline{0} \\ \Rightarrow \#\text{ of samples in class } \underline{0} \end{array}$$

$1\{y^{(i)}=0\}$

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

$n \times n$

$$\Sigma_0 = \frac{\sum_{i=1}^{m} 1\{y^{(i)}=0\} \underbrace{(x^{(i)}-\mu_0)(x^{(i)}-\mu_0)^T}_{\text{outer product}}}{\sum_{i}^{m} 1\{y^{(i)}=0\}}$$

$$\Sigma_1 = \sum^{m} 1\{y^{(i)}=1\} (x^{(i)}-\mu_1)(x^{(i)}-\mu_1)^T$$

$i=1$

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

$\boxed{\Sigma = \Sigma_0 = \Sigma_1}$??

$\Theta = \underbrace{(\phi^*, \mu_0^*, \Sigma^*, \mu_1^*)}_{\text{M2 parameters}}$

How do we make prediction.

$p(x, y; \theta) \rightarrow p(y|x, \theta)$

classification