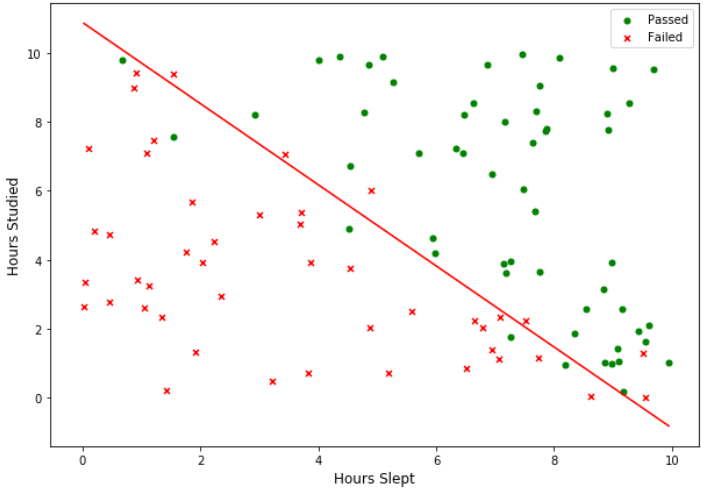
Math 7243

Machine Learning and Statistical Learning Theory

Section 7 Gaussian Discriminant Analysis

Instructor: He Wang
Department of Mathematics
Northeastern University

Review :



Today:

1. Gaussian Discriminant Analysis
2. Review Gaussian distribution
3. Linear Discriminant Analysis (LDA)
4. LDA v.s. Logistics regression

QDA

- Discriminative learning algorithms.

Data $(\vec{x}^{(i)}, y^{(i)})$ ⟶ classification.

Model: $P(\vec{y} \mid \vec{x}; \vec{\theta})$

- Generative learning algorithms.

using $P(\vec{x} \mid y=k)$ and $P(y)$

prior prob.

$P(\vec{x} \mid y=k)$ ← Assumption → { Gaussian $N(\vec{\mu}, \Sigma)$; Bernoulli (ϕ) ; or Categorical $(\phi_1, \ldots, \phi_k)$

$$P(y=k \mid \vec{x}) = \frac{P(y=k, \vec{x})}{P(\vec{x})}$$

$P(y=k \mid \vec{x})$: posterior prob.

$$= \frac{P(\vec{x} \mid y=k) P(y=k)}{\sum_{\text{all } i} P(\vec{x} \mid y=i) P(y=i)}$$

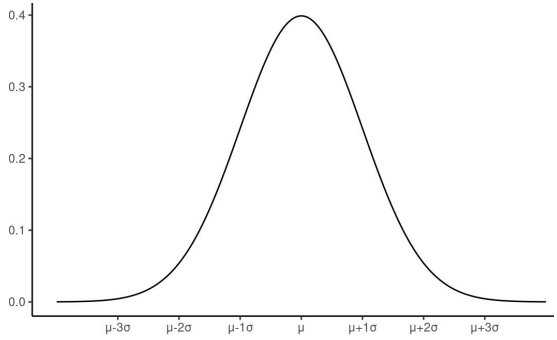
Bayes' Thm

$$\arg\max_{\vec{y}} P(\vec{y} \mid \vec{x}) = \arg\max_{\vec{y}} \left(P(\vec{x} \mid \vec{y}) P(\vec{y}) \right)$$

- Normal distribution.

pdf $f_X(x) = \frac{1}{\sqrt{2\pi}\,\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \quad \text{for} \quad -\infty < x < \infty.$

The mean is $E(X) = \mu$.
The variance is $\mathrm{Var}(X) = \sigma^2$.



- Multivariate normal distribution.

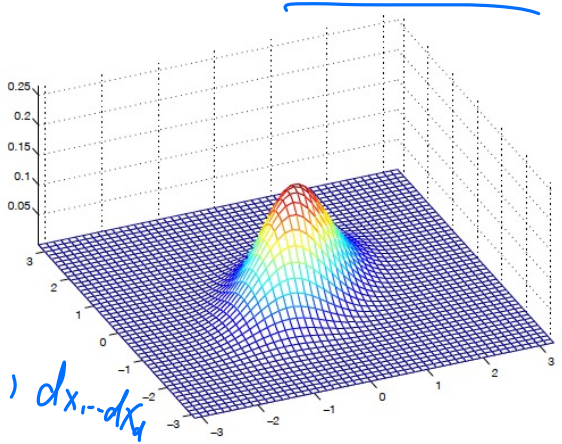
$\vec{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_d \end{bmatrix}$

pdf $f(\vec{x}; \mu, \Sigma) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\vec{\mu})^T \Sigma^{-1}(x-\vec{\mu})\right)$

$\Sigma \in \mathbb{R}^{d \times d}$
- symmetric
- positive definite

$\vec{x} \in \mathbb{R}^d \quad \vec{\mu} \in \mathbb{R}^d$

- $\int_{-\infty}^{\infty} \cdots \int f \, dx_1 \cdots dx_d = 1$



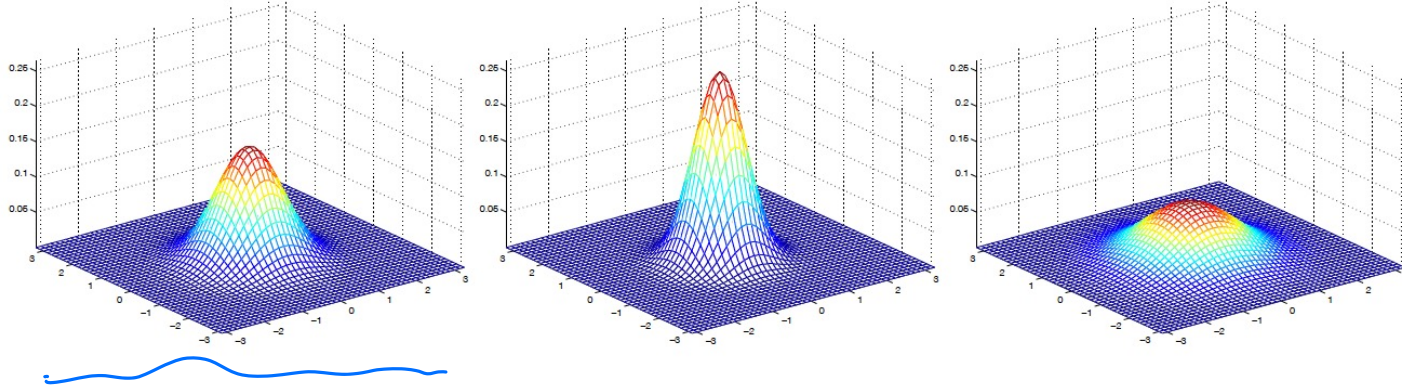
(1) $\vec{\mu} \in \mathbb{R}^d \qquad \vec{\mu} = E(\vec{X}) = \int_{-\infty}^{\infty} \cdots \int x f(\,) \, dx_1 \cdots dx_d$

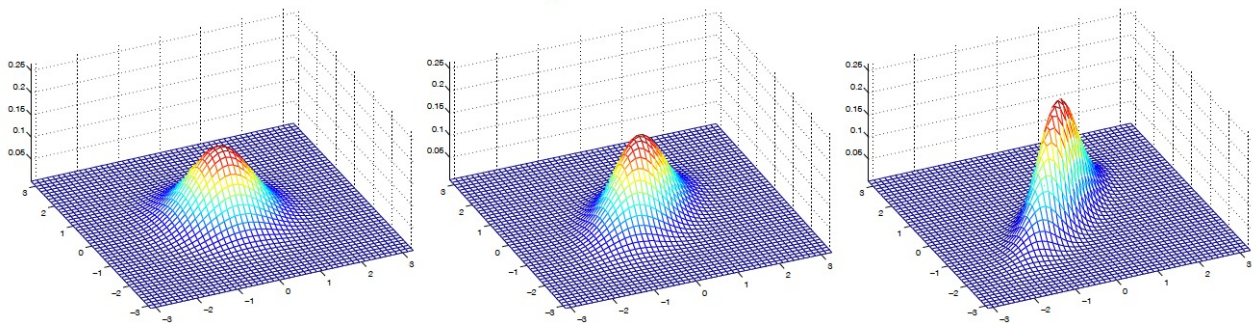
② $\Sigma = \mathrm{cov}(\vec{X}) \in \mathbb{R}^{d \times d}$ covariance matrix of $\vec{X}$

$$X \sim \mathrm{Normal}(\vec{\mu}, \Sigma)$$

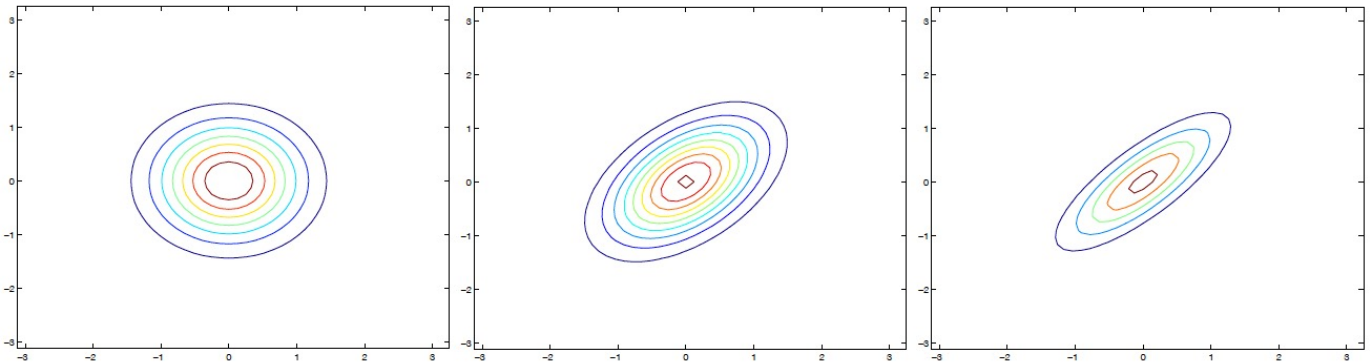
$$\mathrm{Cov}(X) = E\left(XX^T - E(X)E(X)^T\right)$$

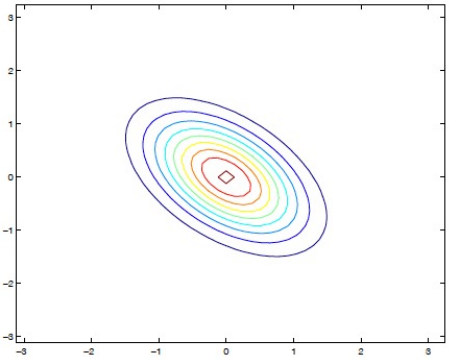
$\vec{\mu} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$ $\Sigma = \begin{bmatrix} 1 & \\ & 1 \end{bmatrix}$ $\Sigma = \begin{bmatrix} \sigma^2 & 0 \\ 0 & \sigma^2 \end{bmatrix}$ $\Sigma = 2\,I$

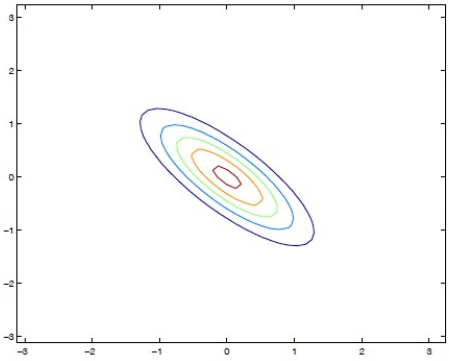


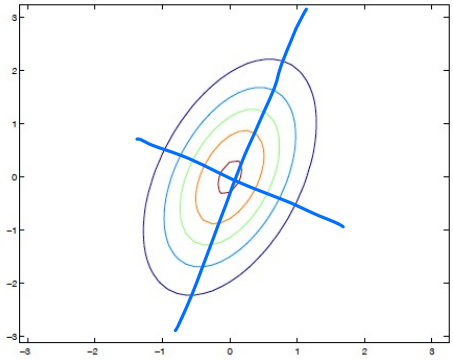


$$\Sigma = \begin{bmatrix} 1 & 0.5 \\ 0.5 & 1 \end{bmatrix} \qquad \Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$







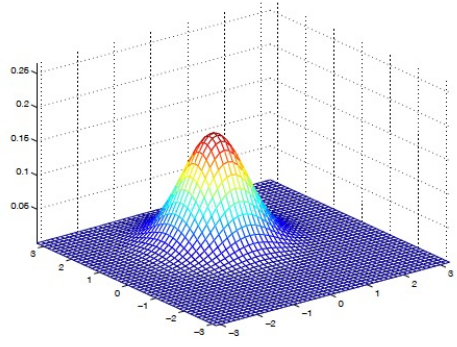


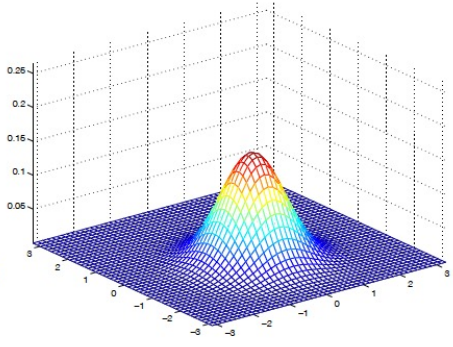
$\mathbb{R}[a_1, a_2]$

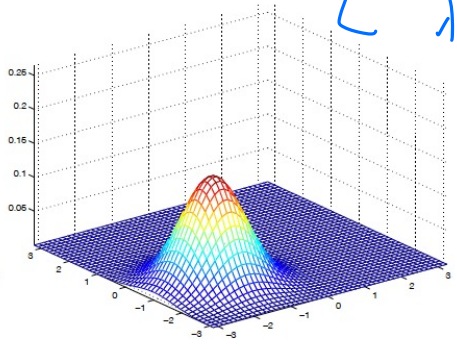
$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$ $\begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$

$\Sigma = PDP^T$

$D = \begin{bmatrix} \lambda_1 & \\ & \lambda_2 \end{bmatrix}$







Data $\{\vec{x}^{(i)}, y^{(i)}\}$ $y^{(i)} \in \{0, 1\}$ or $\{1, 2, \cdots k\}$

$\Downarrow$

- $P(y=j|\vec{x})$ → logistic reg. → model e.g. $P(y=j|\vec{x}) = \frac{1}{1+e^{-\theta^T x}}$

- $P(y=j|\vec{x})$ → L.DA. QDA → $P(y=j|\vec{x}) = \frac{P(\vec{x}|y=j)\,P(y=j)}{\sum_{i=1}^{k} P(\vec{x}|y=i)\,P(y=i)}$

Assumption: ① $(\vec{x}|y=j) \sim \text{Normal}(\vec{\mu}_j, \Sigma_j)$

② $y \sim \text{Bernoulli}(\phi)$

or $\text{Categorical}(\phi_1 \cdots \phi_k)$

$\phi_1 + \cdots + \phi_k = 1$

(LDA)

Assumption: $\Sigma_j = \Sigma$ for all j

- Gaussian Linear Discriminant Analysis(LDA).

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

$$P(y) = \begin{cases} \phi & \text{if } y=1 \\ 1-\phi & \text{if } y=0 \end{cases}$$

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

$$= \phi^{\mathbb{1}(y=1)}(1-\phi)^{\mathbb{1}(y=0)}$$

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

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

$$p(x \mid y=0) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_0)^T \Sigma_0^{-1}(x-\mu_0)\right)$$

$$p(x \mid y=1) = \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_1)^T \Sigma_1^{-1}(x-\mu_1)\right)$$

Maximize the likelihood func tn

$$L(\phi, \vec{\mu}_0, \vec{\mu}_1, \Sigma) \overset{\text{i.i.d.}}{=} \prod_{i=1}^{m} P(\vec{x}=\vec{x}^{(i)}, y=y^{(i)}) = \prod_{i=1}^{m} P(\vec{x}=\vec{x}^{(i)} \mid y=y^{(i)}) P(y=y^{(i)})$$

$$\begin{aligned}\ell(\phi, \mu_0, \mu_1, \Sigma) &= \log \prod_{i=1}^{m} p(x^{(i)}, y^{(i)}; \phi, \mu_0, \mu_1, \Sigma) \\ &= \log \prod_{i=1}^{m} p(x^{(i)} | y^{(i)}; \mu_0, \mu_1, \Sigma) p(y^{(i)}; \phi) = \sum_{i=1}^{m} \left(\log P(\vec{x}^{(i)} | y^{(i)}) + \log (P(y^{(i)})) \right)\end{aligned}$$

- $\frac{\partial \ell}{\partial \phi} = 0 \quad \frac{\partial \ell}{\partial \mu_1} = 0 \quad \frac{\partial \ell}{\partial \Sigma} = 0$ ⇘

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

$P(\vec{x} \mid y=0) = 90\%$

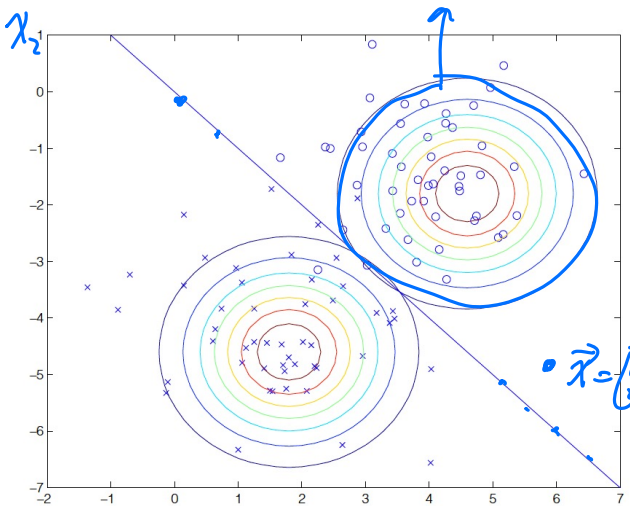
$$\phi = \frac{1}{m} \sum_{i=1}^{m} \mathbb{1}\{y^{(i)} = 1\} \quad \text{mean of } y$$

$$\mu_0 = \frac{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)} = 0\} \vec{x}^{(i)}}{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)} = 0\}} \quad \text{mean of } \vec{x}^{(i)}, 0$$

$$\mu_1 = \frac{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)} = 1\} \vec{x}^{(i)}}{\sum_{i=1}^{m} \mathbb{1}\{y^{(i)} = 1\}} \quad \text{mean of } \vec{x}^{(i)}, 1$$

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

m 代表 — Sample cov.



Data $(\vec{x}^{(i)}, y^{(i)}) \quad i = 1, \cdots, m$

$$0.5 = \cdot P(y=j \mid \vec{x}) = \frac{P(\vec{x} \mid y=j)\, P(y=j)}{\sum_{j=1}^{k} P(\vec{x} \mid y=j)\, P(y=j)}$$

$j=1$

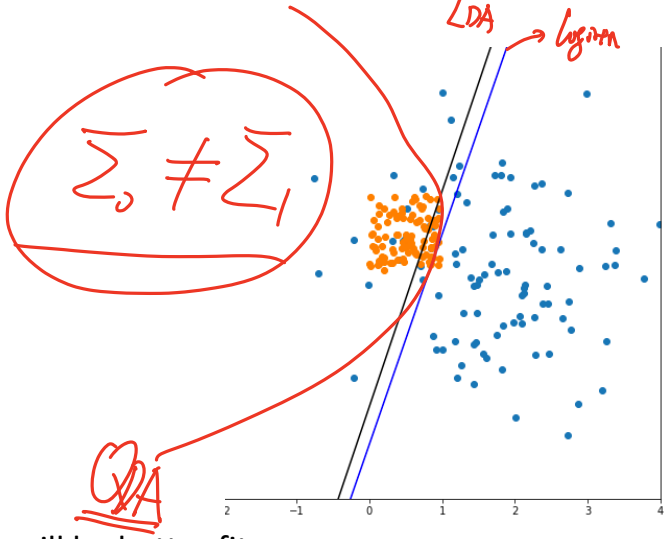
$$\Leftrightarrow P(y=0 \mid \vec{x}) = P(y=1 \mid \vec{x})$$

$$\Leftrightarrow \log \frac{P(y=0 \mid \vec{x})}{P(y=1 \mid \vec{x})} = 0$$

$$\Rightarrow \underline{\text{boundry}} \quad \boxed{f(\vec{x}) = 0} \quad \underline{\text{linear}}$$

{

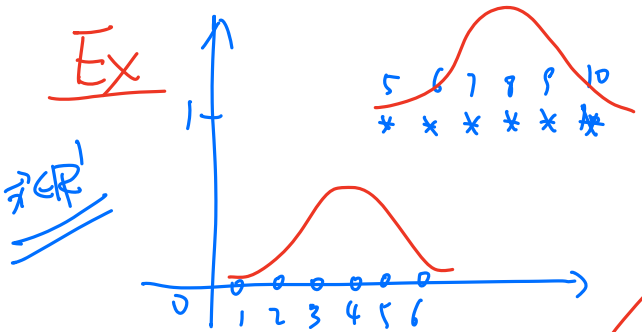
- Compare to Logistic regression



When these modeling assumptions are correct, then GDA will be better fits to the data.

GDA will be a better algorithm than logistic regression for small training set sizes.

logistic regression makes significantly weaker assumptions. So it is more robust and less sensitive to incorrect modeling assumptions.



$$p(x|y=i) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2}\left(\frac{x-\mu_i}{\sigma_i}\right)^2}$$

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

$$P(y=1|x) = \frac{P(x|y=1)P(y=1)}{P(x|y=0)P(y=0)+P(x|y=1)P(y=1)} = \frac{1}{1+e^{\frac{1}{2}\left(\frac{x-\mu_1}{\sigma_1}\right)^2 - \frac{1}{2}\left(\frac{x-\mu_0}{\sigma_0}\right)^2 + \log\left(\frac{1-\phi}{\phi}\right)}}$$

$$= \frac{1}{1+e^{-(\theta_0+\theta_1 x)}}$$

Remark 1

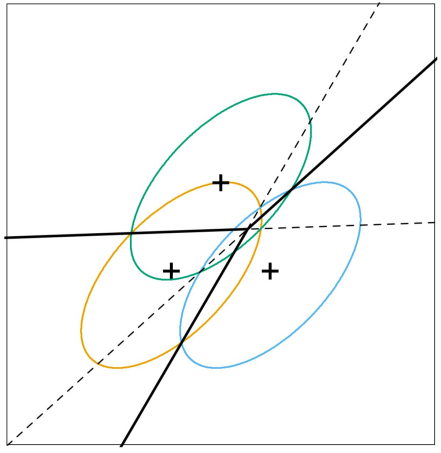
$y = \{-1, 1\}$

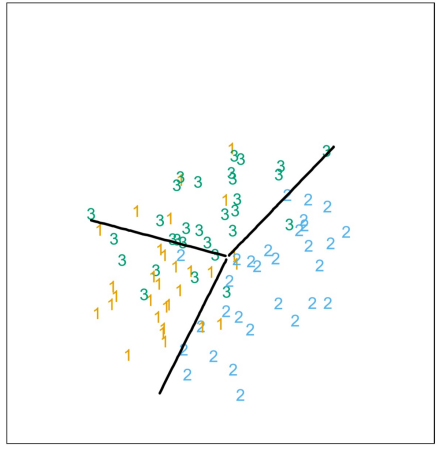
$$P(y=j) = \phi_1^{\mathbb{1}(y=1)} \phi_2^{\mathbb{1}(y=2)}$$

$\phi_1 + \phi_2 = 1$

$y \in \{1, 2, \cdots k\}$

$y \sim Cat(\phi_j)$





$$P(y) = \phi_1^{\mathbb{1}(y=1)} \phi_2^{\mathbb{1}(y=2)} \cdots \phi_k^{\mathbb{1}(y=k)} \qquad \phi_1 + \cdots + \phi_k = 1$$

$$\phi_j = \frac{\sum_{i=1}^{m} \mathbb{1}(y^{(i)} = j)}{m} = \frac{\#(\text{class-}j \text{ observation})}{\#(\text{observations})} =: \frac{N_j}{m}$$

$$\vec{\mu}_j = \frac{\sum_{i=1}^{m} \mathbb{1}(y^{(i)} = j) \cdot \vec{x}^{(i)}}{N_j} \qquad j = 1, 2, \cdots k \qquad \text{Sample mean}$$

$$\Sigma = \frac{1}{m-k} \sum_{i=1}^{m} (\vec{x}^{(i)} - \vec{\mu}_{y^{(i)}})(\vec{x}^{(i)} - \vec{\mu}_{y^{(i)}})^T \qquad \text{Sample Cov.}$$

Remark boundaries:

$$l_{ij} = \{ \vec{x} \in \mathbb{R}^d \mid P(y=i|\vec{x}) = P(y=j|\vec{x}) \}$$

$$= \{ \vec{x} \in \mathbb{R}^d \mid \log \frac{P(y=i|\vec{x})}{P(y=j|\vec{x})} = 0 \}$$

- Remark: Maximize

$$P(y=j|\vec{x}) = \frac{P(\vec{x}|y=j)P(y=j)}{P(\vec{x})}$$

$$\log\left(P(y=j \mid \vec{x})\right) = \log P(\vec{x} \mid y=j) + \log P(y=j) - \log\left(P(\vec{x})\right)$$

$$= \boxed{-\frac{1}{2}\log|\Sigma_j| - \frac{1}{2}(\vec{x}-\mu_j)^T \Sigma_j^{-1}(\vec{x}-\mu_j) + \log\phi_j} + \text{constant}$$

$\boxed{\quad} = \delta_j(\vec{x})$

Assume $\Sigma_1 = \cdots = \Sigma_k = \Sigma$

$$= \boxed{\vec{x}^T \Sigma^{-1}\mu_j - \frac{1}{2}\mu_j^T \Sigma^{-1}\mu_j + \log\phi_j} - \frac{1}{2}\log|\Sigma| + \text{constant}$$

$\boxed{\quad} = \delta_j(\vec{x})$ — Gaussian discriminant func.

- Class $G(\vec{x}) = \arg\max_j \delta_j(\vec{x})$

$$\frac{\partial \delta_j(\vec{x})}{\partial \phi_i} = 0$$

Remark: $x \in \mathbb{Z}^+$.

$$P(x \mid y=i) = \frac{\lambda_i^x e^{-\lambda_i}}{x!}$$

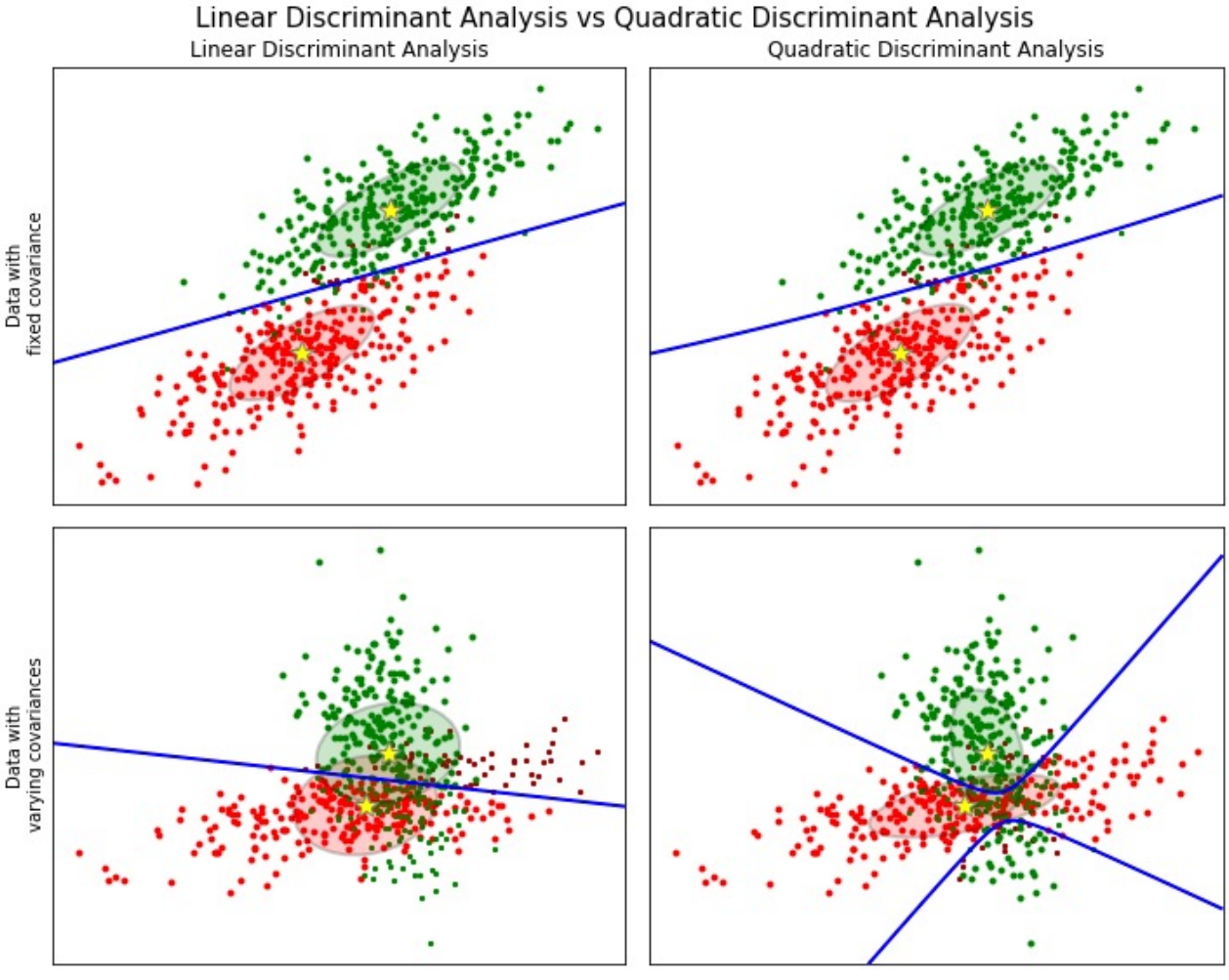
Assumption:

$$\begin{aligned} x \mid y=1 &\sim \text{Poisson}(\lambda_1) \\ x \mid y=0 &\sim \text{Poisson}(\lambda_0) \end{aligned} \rightsquigarrow$$

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

$z = f(x, \lambda_i, \phi)$

$$P(y=1 \mid \vec{x}) = \cdots\cdots = \frac{1}{1+e^{-z}}$$ Ex.


Linear Discriminant Analysis vs Quadratic Discriminant Analysis
Linear Discriminant Analysis
Quadratic Discriminant Analysis
Data with fixed covariance
Data with varying covariances