Pattern Recognition - 1

Bayes Classifier with $p \sim \mathcal{N}(\mu, \Sigma)$

Rudra Mukhopadhyay

December 28, 2021

Brief introduction

-Bayes theorem

$$P(\omega|x) = \frac{p(x|\omega)P(\omega)}{p(x)}$$

- Prior
- Likelihood (class-conditional)
- Evidence
- Posterior

Brief introduction

-Definitions: Model, feature, cost, risk

- State of nature & Action

  $$\mathcal{F} : \{\omega_i, x(\omega_i) | i = 1, \ldots, n\} \mapsto \{a_j | j = 1, \ldots, m\}$$

- Feature vector
  - Properties of *objects* to be classified
  - $\mathcal{F}(\omega_i, x_i) = a_i$
  - Feature vector for $\omega_i : x_i = \{x_{i1}, \ldots, x_{id}\}$
- Cost and risk
  - $\lambda_{ij} = \lambda(a_i | \omega_j)$
  - $R(a_i | x) = \sum_{j=1}^{n} \lambda_{ij} P(\omega_j | x)$

Brief introduction

-Zero-one loss function

- $\lambda_{ij} = \begin{cases} 1 \text{ if } i \neq j \\ 0 \text{ if } i = j \end{cases}$
- $R(a_i|x) = \sum_{j=1}^{n} \lambda_{ij} P(\omega_j|x)$

$$= \sum_{i \neq j} \lambda_{ij} P(\omega_j|x)$$

$$= 1 - P(\omega_i|x)$$

Brief introduction

-Bayes classifier

- $P(\omega_j|x) = \dfrac{p(x|\omega_j)P(\omega_j)}{p(x)}$
- For a given x, a_i is the optimum action if $R(a_i|x)$ is minimum $\forall i$
- Risk minimization and likelihood ratio:

$$R(a_2|x) > R(a_1|x)$$

$$\implies \frac{p(x|\omega_1)}{p(x|\omega_2)} > \frac{(\lambda_{12} - \lambda_{22})P(\omega_2)}{\lambda_{21} - \lambda_{11})P(\omega_1)}$$

$$\implies \text{decide } \omega_1$$

Discriminant function

- $\{g_k | k = 1, \ldots, n\} : x \mapsto \omega_i$ when $g_i(x) > g_j(x) \forall i \neq j$
- A natural choice: $g_i(x) = -R(a_i|x)$
- With zero-one loss function, $g_i(x) = P(\omega_i|x)$
- $\tilde{g}_i(x) := f(g_i(x))$, for any monotonically increasing f

$$\therefore g_i(x) = \ln P(\omega_i|x) = \ln p(x|\omega_i) + \ln P(\omega_i)$$

Decision region and boundary

- $\mathcal{R}_i, i = 1, \ldots, n$
- $g_i(x) > g_j(x) \forall i \neq j \implies x \in \mathcal{R}_i$
  - Dichotomizer: $g(x) := g_1(x) - g_2(x)$
- $\mathcal{B}_{ij}$ is given by the hyperplane $\{x | g_i(x) = g_j(x)\}$

A normal class-conditional

- $p(x|\omega_i) \sim \mathcal{N}(\mu, \Sigma)$
- $p(x) := \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[\frac{-(x-\mu)^t\Sigma^{-1}(x-\mu)}{2}\right]$
    - $\mu = E[x]$
    - $\Sigma = E[(x-\mu)(x-\mu)^t]$
- $r^2 := (x-\mu)^t\Sigma^{-1}(x-\mu)$

Digression

-Generating MVN samples

- Given μ, Σ
- Say, $\Sigma = LL^t$
- $x := \mu + Lu, u \sim \mathcal{N}(0, I)$
  - $E[x] = \mu + LE[u] = \mu$
  - $E[(x - \mu)(x - \mu)^t] = LE[uu^t]L^t = LL^t = \Sigma$

Constructing the discriminant function

$$
\begin{aligned}
&g_i(x) \\
&= \ln p(x|\omega_i) + \ln P(\omega_i) \\
&= -\frac{(x-\mu_i)^t \Sigma_i^{-1}(x-\mu_i)}{2} - \frac{d}{2}\ln 2\pi - \frac{1}{2}\ln|\Sigma_i| + \ln P(\omega_i)
\end{aligned}
$$

$\Sigma_i = \sigma^2 I_d$

-Constructing the discriminant function

Note that $\Sigma_i^{-1} = \frac{1}{\sigma^2} I_d$

$$g_i(x) = -\frac{||x - \mu_i||^2}{2\sigma^2} + \ln P(\omega_i)$$

$$= -\frac{1}{2\sigma^2}[x^t x - 2\mu_i^t x + \mu_i^t \mu_i] + \ln P(\omega_i)$$

$$= w_i^t x + w_{i0} \text{ (A linear machine)}$$

$\Sigma_i = \sigma^2 I_d$

-Estimating $\mathcal{B}_{ij}$

$$g_i(x) - g_j(x) = 0$$

$$\implies (w_i - w_j)^t x + w_{i0} - w_{j0} = 0$$

... some manipulation...

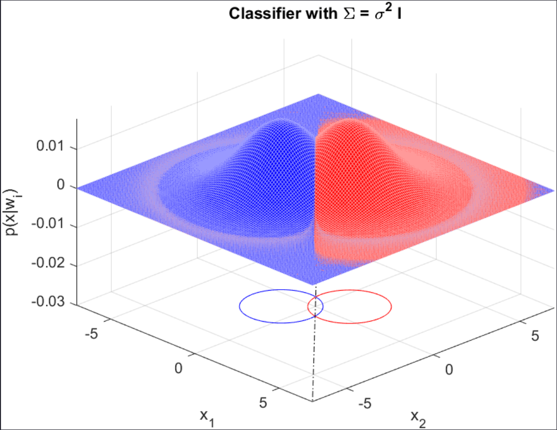
$$w^t(x - x_0) = 0$$

$$w = (\mu_i - \mu_j),$$

$$x_0 = \frac{(\mu_i + \mu_j)}{2} - \frac{\sigma^2}{||\mu_i - \mu_j||^2} \ln \frac{P(\omega_i)}{P(\omega_j)} (\mu_i - \mu_j)$$

$$\Sigma_i = \sigma^2 I_d$$

-Visualization



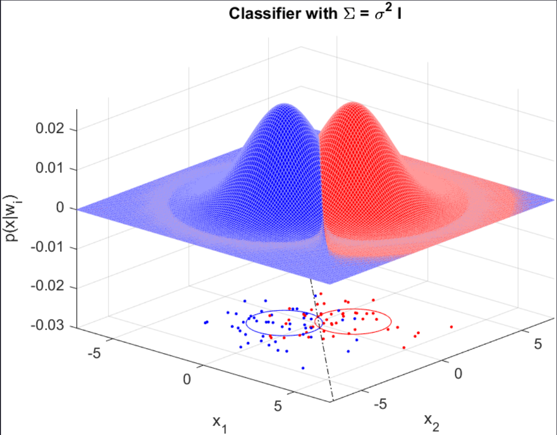
$\Sigma_i = \sigma^2 I_d$

-An example

- $\Sigma_i = 2.5 I_2; \mu_1 = (1, 1), \mu_2 = (-1, -1)$
- Generate data-set, $n = 100$
- First 50 used for training
- Tested for last 50
- Accuracy ≈ 0.84 (for one trial)

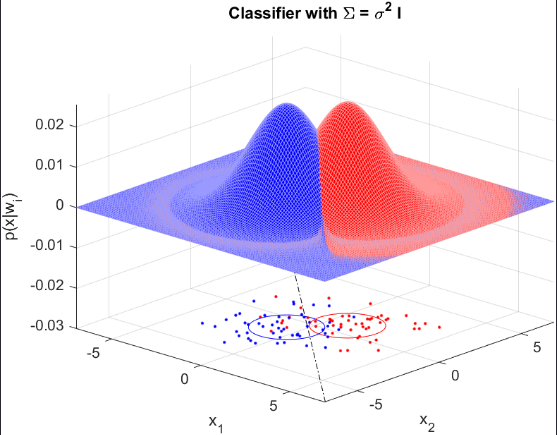
$\Sigma_i = \sigma^2 I_d$

-An example $P = (0.5, 0.5)$



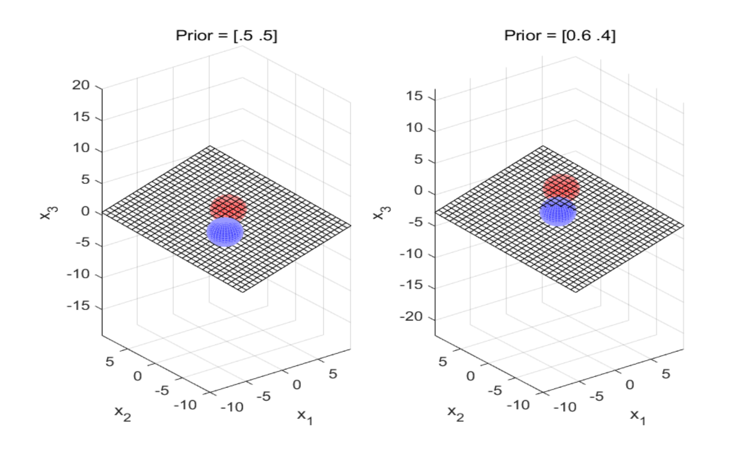
$\Sigma_i = \sigma^2 I_d$

-An example $P = (0.6, 0.4)$



$\Sigma_i = \sigma^2 I_d$

-An example in 3D



Constructing the discriminant function: Revisited

$$
\begin{aligned}
& g_i(x) \\
& = \ln p(x|\omega_i) + \ln P(\omega_i) \\
& = -\frac{(x-\mu_i)^t \Sigma_i^{-1} (x-\mu_i)}{2} + \ln P(\omega_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i|
\end{aligned}
$$

$\Sigma_i = \Sigma$

-Constructing the discriminant function

$$g_i(x) = \ln p(x|\omega_i) + \ln P(\omega_i)$$

$$= -\frac{(x-\mu_i)^t \Sigma^{-1} (x-\mu_i)}{2} + \ln P(\omega_i)$$

$= w_i^t x + w_{i0}$ (ignoring the quadratic term)

- Again a linear machine

$\Sigma_i = \Sigma$

-Estimating $\mathcal{B}_{ij}$

$$w^t(x - x_0) = 0$$

$$w = \Sigma^{-1}(\mu_i - \mu_j)$$

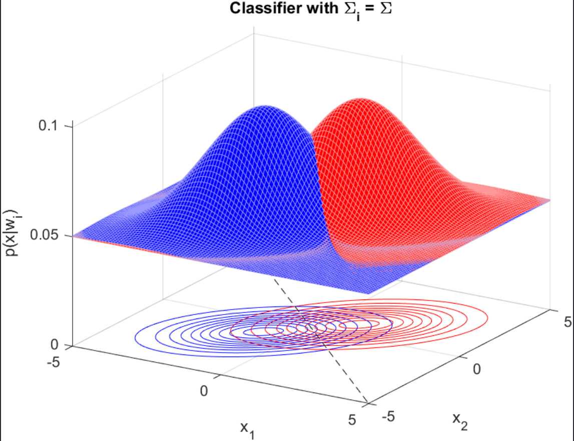
$$x_0 = \frac{(\mu_i + \mu_j)}{2} - \frac{1}{r^2} \ln \frac{P(\omega_i)}{P(\omega_j)}(\mu_i - \mu_j)$$

Note that x_0 depends on the Euclidean distance minimization, when the dimensions are not correlated.

It depends, however, on the Mahalanobis distance minimization in this case.

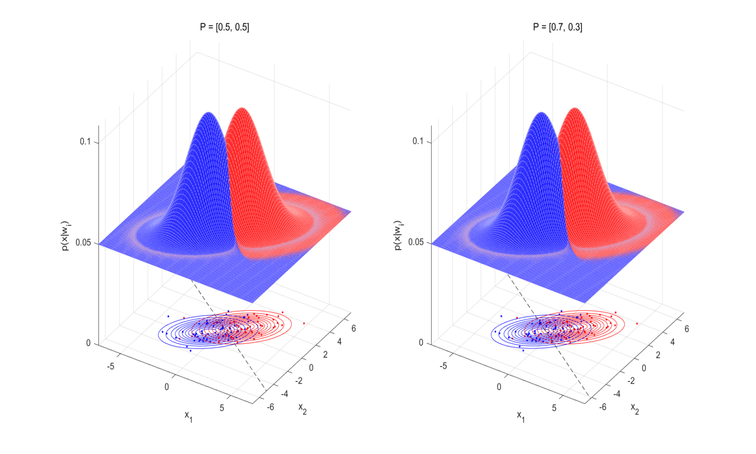
$\Sigma_i = \Sigma$

-Visualization



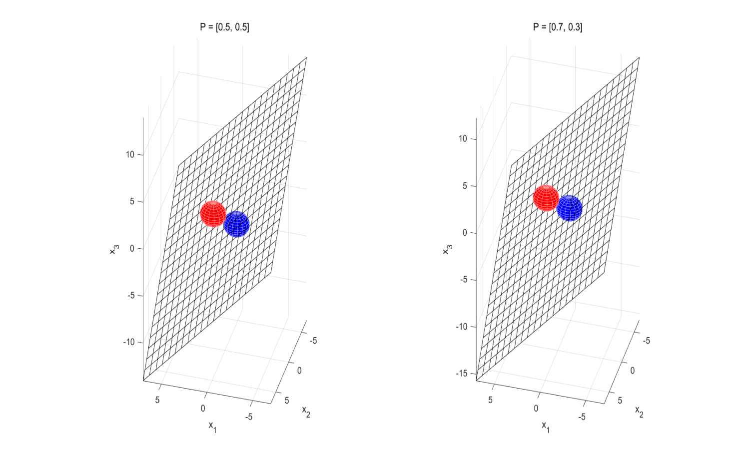
$\Sigma_i = \Sigma$

-An example



$\Sigma_i = \Sigma$

-An example in 3D



Constructing the discriminant function: Revisited

$$
\begin{aligned}
& g_i(x) \\
&= \ln p(x|\omega_i) + \ln P(\omega_i) \\
&= -\frac{(x-\mu_i)^t \Sigma_i^{-1} (x-\mu_i)}{2} + \ln P(\omega_i) - \frac{d}{2} \ln 2\pi - \frac{1}{2} \ln |\Sigma_i|
\end{aligned}
$$

Arbitrary Σ_i

-Constructing the discriminant function

$$(x-\mu_i)^t\Sigma_i^{-1}(x-\mu_i)$$

$$= (x^t-\mu_i^t)(\Sigma_i^{-1}x-\Sigma_i^{-1}\mu_i)$$

$$= x^t\Sigma_i^{-1}x-\mu_i^t\Sigma_i^{-1}x-x^t\Sigma_i^{-1}\mu_i+\mu_i^t\Sigma_i^{-1}\mu_i$$

$$\Sigma_i^t=\Sigma_i \implies \mu_i^t\Sigma_i^{-1}x=x^t\Sigma_i^{-1}\mu_i$$

$$\therefore g_i(x)=x^tW_ix+w_i^tx+w_{i0}$$

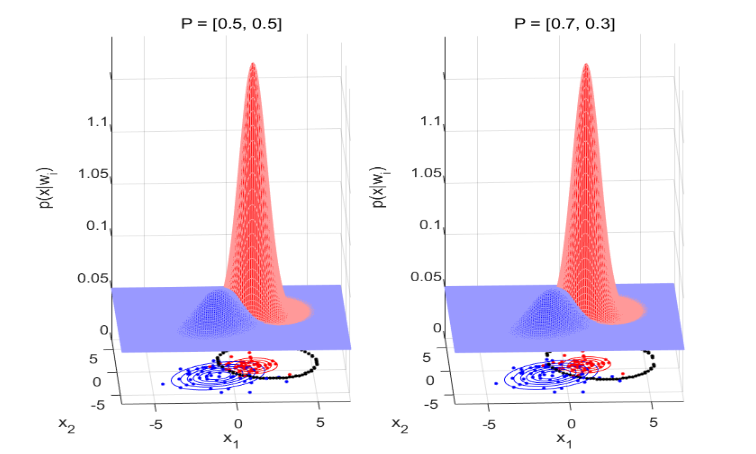
Arbitrary Σ_i

-An example

- $\mu_1 = (1, 1), \mu_2 = (-1, -1)$
- $\Sigma_1 = I_2, \Sigma_2 = 3I_2$
- $P = (.5, .5), P = (.7, .3)$
- 100 data points generated, last 50 tested (accuracy ≈ 0.89)

Arbitrary Σ_i

-An example



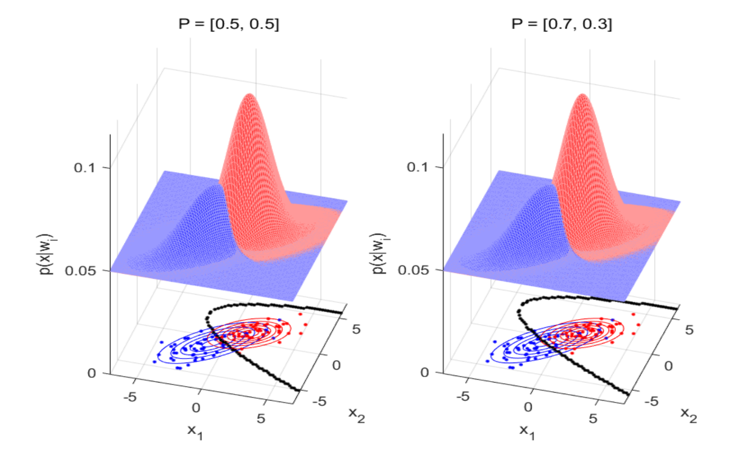
Arbitrary Σ_i

-An example

- $\mu_1 = (1, 1), \mu_2 = (-1, -1)$
- $\Sigma_1 = (2, 1; 1, 2), \Sigma_2 = (3, 2; 2, 3)$
- $P = (.5, .5), P = (.7, .3)$
- 100 data points generated, last 50 tested (accuracy ≈ 0.87)

Arbitrary Σ_i

-An example



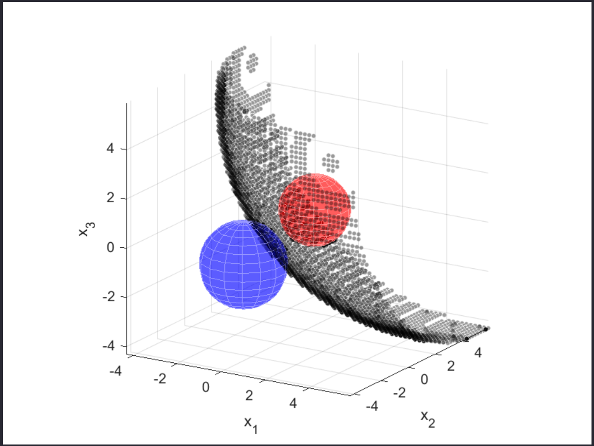
Arbitrary Σ_i

-An example in 3D

- $\mu_1 = (1, 1, 1), \mu_2 = (-1, -1, -1)$
- $\Sigma_1 = 2I_3, \Sigma_2 = 3I_3$
- $P = (.5, .5)$
- 100 data points generated, last 50 tested (accuracy ≈ 0.79)

Arbitrary Σ_i

-An example in 3D



Arbitrary Σ_i

-An example in 3D

- $\mu_1 = (1, 1, 1), \mu_2 = (-1, -1, -1)$
- $\Sigma_1 = (2, 1, 1; 1, 2, 1; 1, 1, 2)$
- $\Sigma_2 = (1, 0, 1; 0, 2, 1; 1, 1, 3)$
- $P = (.5, .5)$
- 100 data points generated, last 50 tested (accuracy ≈ 0.83)

Arbitrary Σ_i

-An example in 3D