TRAINING DATA : $\{(x_i, t_i)\}_{i=1}^{N}$, $x_i, t_i \in \mathbb{R}$

MAPPING MODEL :

$$y = f(\phi(x), w)$$

$$= w_0 \cdot x^0 + w_1 \cdot x^1$$

$$= w_0 + w_1 \cdot x$$

UNKNOWN PARAMETERS : $w = \begin{bmatrix} w_0 \\ w_1 \end{bmatrix}$

TRUE VALUES : $w = \begin{bmatrix} -0.3 \\ 0.5 \end{bmatrix}$

The data x_i will include

additive Gaussian noise

$$t_i = -0.3 + 0.5 \cdot x_n + \varepsilon$$

where $\varepsilon \sim G(\vec{0}, \Sigma)$

↑ vector of zeros $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ ↑ covariance matrix

ASSUME covariance is ISOTROPIC

$\Sigma = \beta \cdot I$, β is a constant

Bayesian Inferencing

DATA likelihood on the ERRORS $\varepsilon_n = t_n - y_n$

$$P(\varepsilon_n \mid w)$$

$$G(0, \beta I)$$

$\times$

PRIOR PROBABILITY on the parameters w

$$P(w)$$

$$G(0, \Sigma_0) = G(0, \alpha^{-1} I)$$

$\propto$

POSTERIOR PROBABILITY on the parameters w

$$P(w \mid \varepsilon)$$

$$G(\mu_N, \Sigma_N)$$

Let's also assume Σ_0 to be isotropic:

$$\boxed{\Sigma_0 = \frac{1}{\alpha} \cdot I = \alpha^{-1} I}$$

$$P(t \mid w) \times P(w) \propto P(w \mid t)$$

$$G(y, \beta I) \times G(0, \alpha^{-1} I) \propto G(\mu_N, \Sigma_N)$$

$$J(w) = \frac{1}{2} \sum_{i=1}^{N} (t_i - y_i)^2$$

$$\arg\min_w J(w)$$

$$= \arg\min_w \prod_{i=1}^{N} \exp\left(-\frac{1}{2} \underbrace{(t_i - y_i)^2}_{\varepsilon_i}\right)$$

$$\mu_N = \Sigma_N \left(\Sigma_0^{-1} \cdot \mu_0 + \beta \cdot X^T \cdot t \right)$$

$$= \beta \cdot \Sigma_N \cdot X^T \cdot t$$

↙ Because we assumed Σ_0 and Σ to be isotropic

and

$$\Sigma_N = \Sigma_0^{-1} + \beta \cdot X^T \cdot X$$

$$= \left(\alpha^{-1} I \right)^{-1} + \beta \cdot X^T X$$

PSEUDO-CODE for online update of Gaussian-Gaussian conjugate prior relationship:

Iteration $t = 0$

(1) Initialize prior parameters $\mu_0^{(t)}$ and $\Sigma_0^{(t)}$

(2) As we collect more data:

(2.1) Compute the data likelihood

(2.2) Estimate the parameters of posterior $\mu_N^{(t)}$ and $\Sigma_N^{(t)}$

(2.3) Update the prior parameters

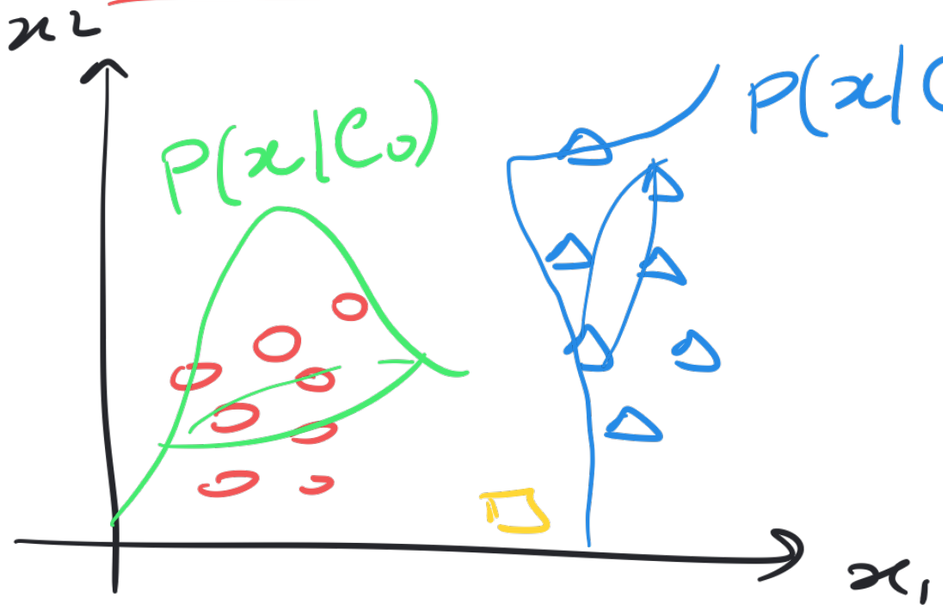
$$\mu_0^{(t+1)} \leftarrow \mu_N^{(t)}$$

$$\Sigma_0^{(t+1)} \leftarrow \Sigma_N^{(t)}$$

(3) $t \leftarrow t + 1$

classifiers

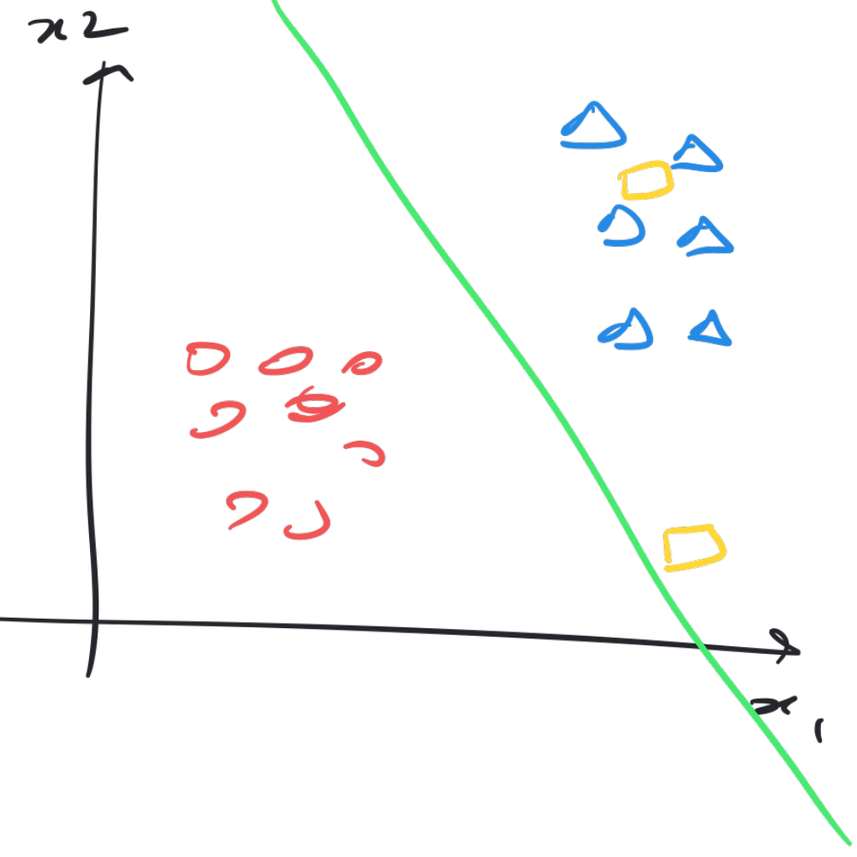
Generative



$P(C_0|\square) > P(C_1|\square)$

$\Rightarrow \square \in C_0$

Discriminative



NAÏVE BAYES CLASSIFIER

2 class-problem

$$P(C_0|x) \underset{C_1}{\overset{C_0}{\gtrless}} P(C_1|x)$$

$$P(C_i|x) = \frac{\overbrace{P(x|C_i)}^{\text{DATA likelihood}} \times \overbrace{P(C_i)}^{\text{Prior}}}{P(x)}$$

↙ Bayes' Theorem

$$P(C_i) = \frac{\#\text{samples in } C_i}{\#\text{total samples}} = \frac{N_i}{N}$$

↳ NOTE that if data is <u>imbalanced</u>, the prior for the under-represented class will be small which will make it unlikely to assign a sample to that class.

Law of Total Prob.

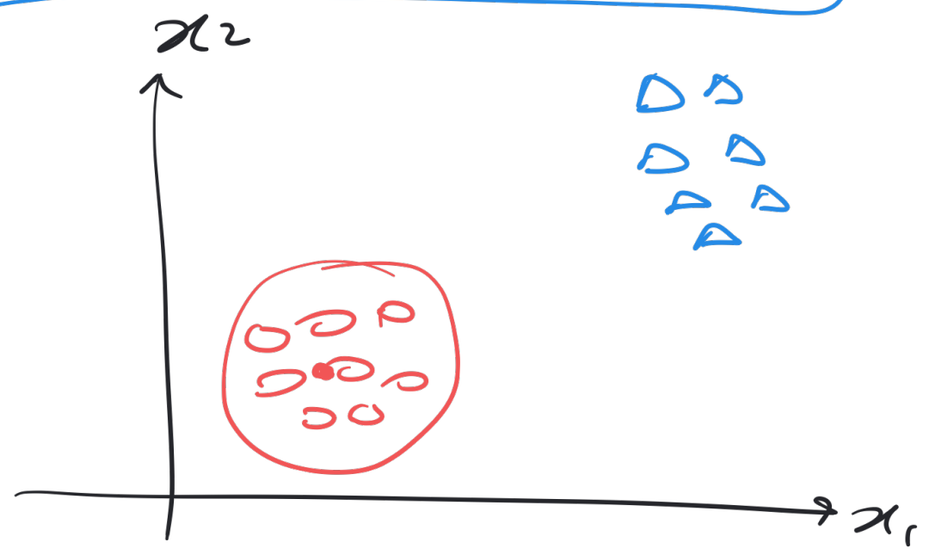
$$P(x) = P(x|C_0).P(C_0) + P(x|C_1).P(C_1)$$

Evidence probability

$$= \sum_j P(x|C_j).P(C_j)$$

Let's consider the data likelihood for each class to be a multivariate Gaussian distribution:

$$P(x \mid C_k) = \frac{1}{(2\pi)^{d/2} \cdot |\Sigma_k|^{1/2}} \cdot \exp\left(-\frac{1}{2}(x-\mu_k)^T \cdot \Sigma_k^{-1} \cdot (x-\mu_k)\right)$$



Key:

x is $\mathbb{R}^d$

$d \equiv$ # features

$\mu_k \equiv$ mean, $\mu_k \in \mathbb{R}^d$

$\Sigma_k \equiv$ covariance, Σ_k is $d \times d$, symmetric

$|\Sigma_k| \equiv$ determinant of Σ_k

The covariance for each class k is <u>isotropic</u>, i.e., $\Sigma_k = \sigma_k^2 \cdot I$

$$\Sigma_k = \begin{bmatrix} \sigma_k^2 & 0 & \cdots & 0 \\ 0 & \sigma_k^2 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_k^2 \end{bmatrix}$$

with this assumption:

$$\Sigma_k^{-1} = (\sigma_k^2 \cdot I)^{-1} = (\sigma_k^2)^{-1} \cdot I$$

$$|\Sigma_k| = (\sigma_k^2)^d$$

with this, we can simplify the data likelihood:

$$P(x|C_k) = \frac{1}{(2\pi)^{d/2}\left((\sigma_k^2)^d\right)^{1/2}} \cdot \exp\left(-\frac{1}{2\sigma_k^2}(x-\mu_k)^T\cdot(x-\mu_k)\right)$$

Given some training data $\{(x_i, t_i)\}_{i=1}^{N}$

$t_i \in \mathbb{N}$, $x_i \in \mathbb{R}^d$

MLE APPROACH for estimating the unknowns, $\theta = \{\mu_k, \Sigma_k\}_{k=1}^{2}$

(1) OBSERVED DATA LIKELIHOOD:

$$\mathcal{L}^{o} = \prod_{i=1}^{N} P(x_i | C_k)$$

$$= \prod_{i=1}^{N} \frac{1}{(2\pi)^{d/2} (\sigma_k^2)^{d/2}} \cdot \exp\left(- \frac{1}{2 \cdot \sigma_k^2} \cdot (x_i - \mu_k)^T (x_i - \mu_k) \right)$$

② log - data likelihood:

$$\mathcal{L} = \ln(\mathcal{L}^{\circ})$$

$$= \sum_{i=1}^{N} \left[-\frac{d}{2} \cdot \ln(2\pi) - \frac{d}{2} \cdot \ln(\sigma_k^2) - \frac{1}{2\sigma_k^2} \cdot (x_i - \mu_k)^T (x_i - \mu_k) \right]$$

③ Find the parameters μ_k and σ_k^2, $\forall k$:

③.1 START with μ_k:

$$\frac{\partial \mathcal{L}}{\partial \mu_k} = 0 \Leftrightarrow \sum_{n \in C_k} \left[\frac{1}{\sigma_k^2} \cdot (x_n - \mu_k) \right] = 0 \Leftrightarrow (*)$$

$$\circledast \Leftrightarrow \sum_{n \in C_k} (x_n - \mu_k) = 0$$

σ_k^2 does NOT depend on n.

$$\Leftrightarrow \sum_{n \in C_k} x_n - \sum_{n \in C_k} \mu_k = 0$$

$$\Leftrightarrow \sum_{n \in C_k} x_n - N \cdot \mu_k = 0$$

$$\Leftrightarrow \mu_k = \frac{\sum_{n \in C_k} x_n}{N} \equiv \text{sample mean for class } C_k$$

$$\mathcal{L} = \sum_{i=1}^{N} \left[-\frac{d}{2}\ln(2\pi) - \frac{d}{2}\ln(\sigma_k^2) - \frac{1}{2\sigma_k^2}(x_i - \mu_k)^T (x_i - \mu_k) \right]$$

(3.2) solve for σ_k^2:

$$\frac{\partial \mathcal{L}}{\partial \sigma_k^2} = 0 \iff \sum_{n \in C_k} \left[-\frac{d}{2} \cdot \frac{1}{\sigma_k^2} + \frac{2(x_n - \mu_k)^T (x_n - \mu_k)}{(2\sigma_k^2)^2} \right] = 0$$

quotient rule in 2nd term

$$\iff \sum_{n \in C_k} \left[-d + \frac{2(x_n - \mu_k)^T (x_n - \mu_k)}{2\sigma_k^2} \right] = 0$$

multiplied on numerator both sides by $2\sigma_k^2$

$$\Leftrightarrow \sum_{n \in C_k} d = \sum_{n \in C_k} \frac{1}{\sigma_k^2} \cdot (x_n - \mu_k)^T (x_n - \mu_k)$$

$$\Leftrightarrow N_k \cdot d = \frac{1}{\sigma_k^2} \cdot \sum_{n \in C_k} (x_n - \mu_k)^T (x_n - \mu_k)$$

$$\Leftrightarrow \sigma_k^2 = \frac{\sum_{n \in C_k} (x_n - \mu_k)^T (x_n - \mu_k)}{d \cdot N_k}$$

$\equiv$ Sample variance

$$\Sigma_k = \sigma_k^2 \cdot I$$