

Clasificadores → Modulo Bayesiano

Por teoría de detección:

$$L(x) = \frac{P(x|A)}{P(x|B)} \rightarrow \frac{P(B)}{P(A)}$$

donde $x \in \mathbb{R}^q \rightarrow$ muestra en P atributos.

A, B classes.

$$P(x|A)p(A) - P(x|B)p(B) > 0$$

$$p(x, A) - p(x, B) > 0$$

Desde posterior:

$$\frac{p(A|x)}{p(B|x)} = \frac{\underbrace{p(x|A)p(A)}_{\text{evidencia}}}{\underbrace{p(x|B)p(B)}_{\text{prior}}} \stackrel{\text{verosimilitud}}{\longrightarrow} p(A|x) > p(B|x) ?$$

$$\frac{p(x|A)p(A)}{p(x)} - \frac{p(x|B)p(B)}{p(x)} > 0$$

$$p(x|A)p(A) - p(x|B)p(B) > 0$$

Cómo modelar $P(A|x) = ?$
 $P(B|x) = ?$

Bayes ingenuo \rightarrow Naïve Bayes

Asume características independientes.

$$P(x, A) = \prod_{j=1}^P p(x_j|A) p(A); \quad P(x, B) = \prod_{j=1}^P p(x_j|B) p(B)$$

$p(A), p(B)$ se estiman mediante frecuencia relativa \rightarrow conteo.

Multiclaro \rightarrow

$$y_{\text{new}} = \arg \max_{A_c} \prod_{j=1}^P p(x_j|A_c) p(A_c)$$

Clasificador Bayesiano → Normal Gaussiano

$$L(x) = P(x|A)p(A) - P(x|B)p(B) = 0 \rightarrow \text{Frontera}$$

$$\frac{P(x|A)}{P(x|B)} = \frac{P(B)}{P(A)} \Rightarrow \log\left(\frac{P(x|A)}{P(x|B)}\right) = \log\left(\frac{P(B)}{P(A)}\right)$$

$$\log(P(x|A)) - \log(P(x|B)) = \log(P(B)/P(A))$$

$$R(x) = \log(P(x|A)) - \log(P(x|B)) - \log(P(B)/P(A))$$

$$P(x|A) = \frac{1}{2\pi^{p/2} |\Sigma_A|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_A)^T \Sigma_A^{-1} (x-\mu_A)\right)$$

$$P(x|B) = \frac{1}{2\pi^{p/2} |\Sigma_B|^{1/2}} \exp\left(-\frac{1}{2}(x-\mu_B)^T \Sigma_B^{-1} (x-\mu_B)\right)$$

$$\log(P(x|A)) = -\frac{p}{2} \log(2\pi) - \frac{1}{2} \log(|\Sigma_A|) \\ - \frac{1}{2} (x-\mu_A)^T \Sigma_A^{-1} (x-\mu_A)$$

cuadrático

$$R(x) = -\frac{1}{2} (x-\mu_A)^T \Sigma_A^{-1} (x-\mu_A) + \frac{1}{2} (x-\mu_B)^T \Sigma_B^{-1} (x-\mu_B)$$

$$-\frac{1}{2} \log(|\Sigma_A|) + \frac{1}{2} \log(|\Sigma_B|) - \log(P(B)/P(A))$$

Clasificador cuadrático asume distribución Gaussiana.

Clasificador lineal Bayesiano

Si $\Sigma_A = \Sigma_B = \Sigma$ en cuadrático:

$$R(x) = -\frac{1}{2} (x - \mu_A)^T \Sigma^{-1} (x - \mu_A) + \frac{1}{2} (x - \mu_B)^T \Sigma^{-1} (x - \mu_B) + \text{cte}$$

$$R(x) = \frac{1}{2} \left[-\underbrace{x^T \Sigma^{-1} x}_{+} + x^T \Sigma^{-1} \mu_A + \mu_A^T \Sigma^{-1} x - \mu_A^T \Sigma^{-1} \mu_A + \right. \\ \left. + \underbrace{x^T \Sigma^{-1} x}_{+} - x^T \Sigma^{-1} \mu_B - \mu_B^T \Sigma^{-1} x + \mu_B^T \Sigma^{-1} \mu_B \right] + \text{cte}$$

$$R(x) = \frac{1}{2} \left[2 \mu_A^T \Sigma^{-1} x - 2 \mu_B^T \Sigma^{-1} x - \mu_A^T \Sigma^{-1} \mu_A + \mu_B^T \Sigma^{-1} \mu_B \right] + \text{cte}$$

$$R(x) = \frac{1}{2} \left[2 (\mu_A - \mu_B)^T \Sigma^{-1} x - \mu_A^T \Sigma^{-1} \mu_A + \mu_B^T \Sigma^{-1} \mu_B \right] + \text{cte}$$

$$R(x) = [(\mu_A - \mu_B)^T \Sigma^{-1} x - \frac{1}{2} \mu_A^T \Sigma^{-1} \mu_A + \mu_B^T \Sigma^{-1} \mu_B] + \text{cte}$$

Final $\approx w^T x + b$

$$\Sigma = \underbrace{p(A)\Sigma_A + p(B)\Sigma_B}_{\Sigma} \rightarrow \text{Aproximación para codificar} \\ \text{aparición de A y B}$$

Clasificador Bayesiano lineal por diferencia entre medias.

Si: $\Sigma_A = \Sigma_B = \sigma^2 I$

$$R(x) = \frac{1}{2} [\mu_A - \mu_B]^T x + \text{cte.}$$

Modelo logístico: $P(A|x) = \frac{P(x|A)p(A)}{P(x)}$

$$L(x) = \frac{P(A|x)}{P(B|x)} = \frac{P(x|A)}{P(x|B)} \rightarrow \frac{P(B)}{P(A)}$$

$$\Rightarrow L(x) = \frac{P(A|x)}{P(B|x)} = \frac{P(x|A)p(A)}{P(x|B)p(B)}$$

$$\log(L(x)) = \log\left(\frac{P(A|x)}{P(B|x)}\right) = \log\left(\frac{P(x|A)p(A)}{P(x|B)p(B)}\right) = w^T x + w_0$$

Se asume modelo lineal en $\log(L(x))$.

En biclassificación: $P(A|x) + P(B|x) = 1$.
 $x \in A \cup x \in B$.

$$Y(x) = [P(A|x) \quad P(B|x)] = [0.3 \quad 0.2]$$

$$L(x) = \frac{P(A|x)}{P(B|x)} = \frac{P(x|A) P(A)}{P(x|B) P(B)} = e^{w^T x + w_0}$$

$$P(B|x) = 1 - P(A|x); \quad P(A|x) + P(B|x) = 1$$

$$P(A|x) = \frac{(1 - P(A|x)) P(x|A) P(A)}{P(x|B) P(B)}$$

$$P(A|x) = \frac{P(x|A) P(A)}{P(x|B) P(B)} - \frac{P(A|x) P(x|A) P(A)}{P(x|B) P(B)}$$

$$P(A|x) = e^{w^T x + w_0} - P(A|x) e^{w^T x + w_0} =$$

$$P(A|x) [1 + e^{w^T x + w_0}] = e^{w^T x + w_0}$$

$$P(A|x) = \frac{e^{w^T x + w_0}}{1 + e^{w^T x + w_0}} \left(\frac{e^{-(w^T x + w_0)}}{e^{-(w^T x + w_0)}} \right)$$

$$P(A|x) = \frac{1}{1 + e^{-(w^T x + w_0)}}$$

función sigmoid

$$P(B|x) = 1 - P(A|x) = 1 - \frac{1}{1 + e^{-(w^T x + w_0)}}$$

$$P(B|x) = \frac{1 + e^{-(w^T x + w_0)} - 1}{1 + e^{-(w^T x + w_0)}}$$

$$P(B|x) = \frac{e^{-(w^T x + w_0)}}{1 + e^{-(w^T x + w_0)}}$$

Funcióñ de costo:

$$L(x) = \log\left(\frac{P(A|x)}{P(B|x)}\right) = \log(P(A|x)) - \log(P(B|x))$$

Si se asume X_n i.i.d:

$$P(A|X_n) = \frac{1}{1 + e^{-(w^T x + w_0)}}$$

$$P(A|X, w) = \prod_{n=1}^N P(A|X_n, w)$$

$$L(x) = \log\left(\prod_{n=1}^N P(A|X_n, w)\right) - \log\left(\prod_{n=1}^N P(B|X_n, w)\right)$$

Derivada de la Sigmoidal: $\sigma(t) = 1/(1+e^{-t})$

$$\frac{\partial}{\partial t} \left\{ \frac{1}{1+e^{-t}} \right\} = \frac{\partial}{\partial t} \left\{ (1+e^{-t})^{-1} \right\}$$

$$= - (1+e^{-t})^{-2} e^{-t} (-1)$$

$$= \frac{e^{-t}}{(1+e^{-t})^2} = \frac{1}{1+e^{-t}} \cdot \frac{e^{-t}}{1+e^{-t}} = \sigma(t)(1-\sigma(t))$$

Versión Simplificada

$$L(w) = \begin{cases} -\log(p(A|x_n)) & \text{Si } y_n = A \\ -\log(p(B|x_n)) & \text{Si } y_n = B \end{cases}$$

$$L(w) = \begin{cases} -\log(p(A|x_n)) & \text{Si } y_n = A \\ -\log(1-p(B|x_n)) & \text{Si } y_n = B \end{cases}$$

$$p(A|x_n) + p(B|x_n) = 1.$$

$$I(x) = \log(1/p(x)) \rightarrow \text{Medida de información}$$

$I(x) = -\log(p(x)) \rightarrow$ Medida de información sobre x



Si $A=1$ $\Sigma B=0$

$$L(x) = \begin{cases} -\log(p(A|x)) & \text{Si } y=1 \\ -\log(1-p(A|x)) & \text{Si } y=0 \end{cases}$$

Función de costo propuesta: log loss

$$J(w) = \frac{1}{N} \sum_{n=1}^N y_n \log(\hat{p}_n) + (1-y_n) \log(1-\hat{p}_n)$$

$$\hat{p}_n = P(A|X_n) = \frac{1}{1 + e^{-(w^T x_n + w_0)}}$$

Si $y_n=1$ $\hat{p}_n \rightarrow 1$; $\log(\hat{p}_n) \rightarrow 0$ (Si X_n bien clasif.)

Log-loss desde cross-entropia

Entropia $H(x) = \mathbb{E}\{I(x)\}$

$$I(x) = \log\left(\frac{1}{p(x)}\right) = -\log(p(x))$$

Shannon : $\mathbb{E}\{x\} = \int x p(x) dx$

$$H_s(x) = \int I(x) p(x) dx$$

$$H_s(x) = - \int p(x) \log p(x) dx$$

Para $p(x) > q(x)$: Entropia relativa
(Divergencia)

$$D_{KL}(p||q) = \int p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

$$D_{KL}(p||q) = \int p(x) \log(p(x)) dx - \int p(x) \log(q(x)) dx$$

$$D_{KL}(p||q) = H_s(p, q) - H(p)$$

$$H_s(p, q) = - \int p(x) \log(q(x)) dx$$

→ Cross entropia.

$$H(p, q) = \mathbb{E}_p [I(q)] = -\mathbb{E}_p [\log(q)]$$

Si $P = y \rightarrow$ etiqueta verdadera

$$q = P(A|X) = \frac{1}{1 + e^{-(w^T x + w_0)}}$$

$$J(w) = \frac{1}{N} \sum_{n=1}^N H(P(y|x_n), q(\hat{y}|x_n))$$

$$P(y=1|x_n) = y_n; \quad P(y=0|x_n) = 1 - y_n$$

$$q(y=1|x_n) = \frac{1}{1 + e^{-(w^T x_n + w_0)}} = \hat{y}_n$$

$$q(y=0|x_n) = \frac{e^{-(w^T x_n + w_0)}}{1 + e^{-(w^T x_n + w_0)}} = 1 - \hat{y}_n$$

$$J(w) = \frac{1}{N} \sum_{n=1}^N y_n \log(\hat{y}_n) + (1-y_n) \log(1-\hat{y}_n)$$

Derivada de $\log(\sigma(t))$

$$\frac{\partial}{\partial t} \left[\log(\sigma(t)) \right] = \frac{1}{\sigma(t)} \frac{\partial}{\partial t} \sigma(t)$$
$$= \frac{1}{\sigma(t)} \sigma(t)(1-\sigma(t))$$

$$\frac{\partial}{\partial t} \log(\sigma(t)) = 1 - \sigma(t)$$

Por ende:

$$\frac{\partial J(w)}{\partial w_j} = \frac{1}{N} \sum_{n=1}^N (\sigma(\tilde{w}^T x_n) - b_n) x_{nj}$$

DEMOSTRAR

Softmax regresión

Regresión logística multinomial

Softmax: $s_k(x) = x^T w_k$

$$\hat{P}_k = \sigma(s(x))_k = \frac{\exp(s_k(x))}{\sum_{k=1}^K \exp(s_k(x))}$$

Encontrar w_k con Cross-entropía

$$J(w) = -\frac{1}{n} \sum_{n=1}^N \sum_{k=1}^K y_n^k \log(\hat{P}_n^k)$$

$$y_n = [y_n^1 \ y_n^2 \ \dots \ y_n^K]$$

Demostrar