Clasificador → Modelo Bayesiano

Por teoría de decisión:

$$l(x) = \frac{p(x|A)}{p(x|B)} \gtrless \frac{p(B)}{p(A)}$$

donde $x \in \mathbb{R}^p$ → muestra en p atributos.

A, B clases.

$$p(x|A)p(A) - p(x|B)p(B) \gtrless 0$$

$$p(x,A) - p(x,B) \gtrless 0$$

Desde posterior:

$$\underbrace{p(A|x)}_{\text{posterior}} = \frac{\overbrace{p(x|A)}^{\text{verosimilitud}}\overbrace{p(A)}^{\text{prior}}}{\underbrace{p(x)}_{\text{evidencia}}}$$

$$p(A|x) \gtrless p(B|x) \; ?$$

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

$$p(x|A)p(A) - p(x|B)p(B) \gtrless 0$$

Cómo modelar $p(A|x) = ?$

$p(B|x) = ?$

Bayes ingenuo → 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 → conteo.

Multiclase →

$$y_{new} = \underset{A_c}{\arg\max} \prod_{j=1}^{p} p(x_j|A_c)\,p(A_c)$$

Clasificador Bayesiano → Normal Gaussiano

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

$$\frac{p(X|A)}{p(X|B)} = \frac{p(B)}{p(A)} \Rightarrow \log\left(\frac{p(X|A)}{p(X|B)}\right) = \log\left(\frac{p(B)}{p(A)}\right)$$

$$\log\left[p(X|A)\right) - \log\left(p(X|B)\right| = \log\left(p(B)/p(A)\right)$$

$$R(X) = \log\left(p(X|A)\right) - \log\left(p(X|B)\right) - \log\left(p(B)/p(A)\right)$$

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

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

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

cuadrático

$$R(X) = -\frac{1}{2}(X-\mu_A)^T \Sigma_A^{-1}(X-\mu_A) + \frac{1}{2}(X-\mu_B)^T \Sigma_B^{-1}(X-\mu_B) - \frac{1}{2}\log\left(|\Sigma_A|\right) + \frac{1}{2}\log\left(|\Sigma_B|\right) - \log\left(p(B)/p(A)\right)$$

Clasificador cuadrático asume distribución Gaussiana.

Clasificador lineal Bayesiano

Si $\underline{\Sigma_A = \Sigma_B = \Sigma}$ en cuadrático:

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

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

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

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

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

lineal $\rightarrow w^T x + b$

$\underline{\underline{\Sigma}} = \underline{p(A)\Sigma_A + p(B)\Sigma_B}$ → Aproximación para codificar aparición de A y B

Clasificador Bayesiano lineal por diferencia entre medias.

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

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

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

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

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

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

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

En biclase $p(A|x) + p(B|x) = 1$.

$x \in A$ o $x \in B$.

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

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

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

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

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

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

$$p(A|x)\left[1 + e^{w^T x + w_0}\right] = e^{w^T x + w_0}$$

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

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

función sigmoide

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

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

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

Función de costo:

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

Si se asume x_n i.i.d:

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

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

$$\boxed{l(x) = \log\left(\prod_{n=1}^{N} p(A|x_n, w)\right) - \log\left(\prod_{n=1}^{N} p(B|x_n, w)\right)}$$

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

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

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

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

Versión Simplificada

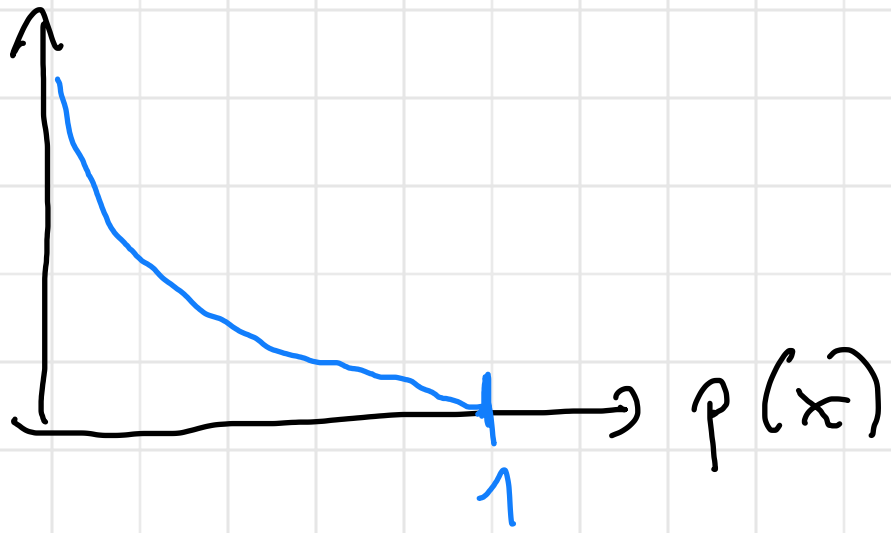
$$L(w) = \begin{cases} -\log(p(A|X_n)) & \text{Si: } y_n = A \\ -\log(p(B|X_n)) & \text{Si: } y_n = B \end{cases}$$

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

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

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

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



Si $A=1$ y $B=0$

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

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 $H(X) = \mathbb{E}\{I(X)\}$

$I(X) = \log(1/p(X)) = -\log(p(X))$

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

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

$$\boxed{H_s(X) = -\int p(x) \log p(x)\, dx}$$

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

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

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

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

$$\boxed{H_s(p,q) = -\int p(x) \log(q(x))\, dx} \rightarrow \text{Cross entropia.}$$

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

Si $p = y$ → etiqueta verdadera

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

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

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

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

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

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

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 $\ln \mathcal{L}$:

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

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-entropia

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

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

Demostrar