Notas clase 7/8 MDS7203 modelos discriminativos vs generativos

Sean $C_1, C_2, \dots, C_K$ clases. Consideremos que nuestro conjunto de D datos está dado por

$$D = \{(x_i, t_i)\}_{i=1}^{N} \subset \mathbb{R}^M \times \{C_i\}_{i=1}^{K}$$

Consideremos el caso en el que queremos modelar la probabilidad condicional siguiente:

$$P(C_k \mid x)$$

Al plantearlo de este modo podemos computar el estimador máximo a posteriori.

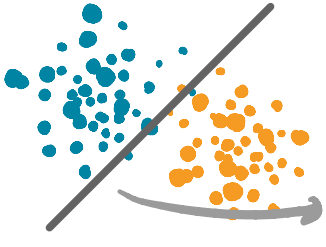
Acá queremos maximizar la probabilidad de que la etiqueta asignada a los datos sea aquella que corresponde. Esto se puede escribir, para el caso de clasificación binaria y un punto de dato (x_i, t_i):

$$L_i(\theta) = p(x_i, t_i \mid \theta) = \begin{cases} p(x_i, C_1 \mid \theta) & \text{si } t_i = C_1 \\ p(x_i, C_2 \mid \theta) & \text{si } t_i = C_2 \end{cases}$$

Asumimos que $C_1 = 1$ y $C_2 = 0$

$$= p(x_i, C_1 \mid \theta)^{t_i} \, p(x_i, C_2 \mid \theta)^{1 - t_i}$$

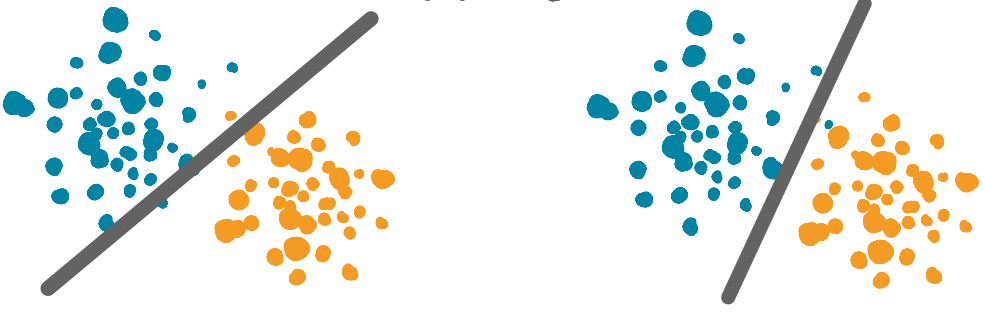
Consideremos por simplicidad el caso de clasificación binaria $K = 2$.



Impongamos que nuestros conjuntos de datos distribuyen de acuerdo a una distribución Gaussiana. (con la misma covarianza)

TENEMOS LA INTUICIÓN DE ENCONTRAR UN HIPERPLANO QUE SEPARE LOS DATOS DE LA MEJOR MANERA POSIBLE:

$$a^T x + b = 0$$



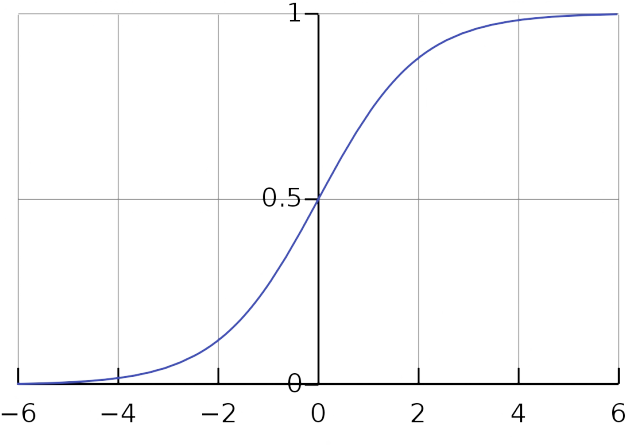
GEOMÉTRICAMENTE SABEMOS QUE DADOS a Y b, NUESTRA RECTA ESTÁ DETERMINADA POR $a^T x + b = 0$, LUEGO LOS CASOS $a^T x + b > 0$ Y $a^T x + b < 0$ DETERMINAN A QUE LADO DEL BORDE DE DECISIÓN QUEDA EL PUNTO.

LA LEJANÍA A LA RECTA NOS DA LA INTUICIÓN DE MÁS CERTIDUMBRE RESPECTO A LA CLASIFICACIÓN, PERO DEBEMOS PASAR ESTO A PROBABILIDADES...

RECORDEMOS LA FUNCIÓN LOGÍSTICA:

$$\sigma : \mathbb{R} \to (0,1)$$
$$r \mapsto \frac{1}{1+e^{-r}}$$

ES UN MAPEO CONTINUO DE LOS REALES AL DOMINIO $(0,1)$!!!



Esta función tiene además propiedades convenientes:

- Refleja: $\sigma(-r) = 1 - \sigma(r)$
- Diferenciable, más aún: $\frac{d}{dr}\sigma(r) = \sigma(r)(1-\sigma(r))$
- Con inversa: $r(\sigma) = \log\left(\frac{\sigma}{1-\sigma}\right)$

Esto nos sirve usar:

$$\mathbb{P}(C_k \mid x) = \frac{1}{1+e^{-a^T x - b}}$$

A esto le llamamos **regresión logística**.

Llegado este punto, podemos optimizar directamente los parámetros a y b. Consideremos el siguiente abuso de notación:

$$\omega = \begin{pmatrix} a \\ b \end{pmatrix}, \quad \tilde{x} = \begin{pmatrix} x \\ 1 \end{pmatrix}$$

Recordando que la verosimilitud está dada por

$$\hat{\theta}_{MAP} = \underset{\theta}{\operatorname{argmáx}} \prod_{i=1}^{N} L_i(\theta) = \underset{\theta}{\operatorname{argmáx}} \prod_{i=1}^{N} p(C_1 \mid x_i)^{t_i} p(C_2 \mid x_i)^{1-t_i}$$

Recordemos que $\mathbb{P}(C_k \mid x_i) = \frac{1}{1+e^{-\omega^T x_i}} = \sigma(\omega^T x)$

Denotando $\sigma_i = \sigma(\omega^T x_i)$ queda:

$$\hat{\theta}_{MAP} = \underset{\omega}{\operatorname{argmáx}} \prod_{i=1}^{N} \sigma_i^{t_i} (1-\sigma_i)^{1-t_i}$$

Entonces derivando la log verosimilitud nos da:

$$\frac{\partial \log L}{\partial \omega} = \frac{\partial}{\partial \omega}\left(\sum_{i=1}^{N} \left(t_i \log \sigma_i + (1-t_i)\log(1-\sigma_i)\right)\right)$$

$$= \sum_{i=1}^{N} t_i \frac{\partial \log \sigma_i}{\partial \omega} + (1-t_i)\frac{\partial}{\partial \omega}\log(1-\sigma_i)$$

Recordando que $\frac{d}{dr}\sigma(r) = \sigma(r)(1-\sigma(r))$

$$\frac{\partial}{\partial w}\log \sigma_i = \frac{1}{\sigma_i}\sigma_i(1-\sigma_i)x_i \quad y \quad \frac{\partial}{\partial w}\log(1-\sigma_i) = \frac{1}{1-\sigma_i}\cdot -\sigma_i(1-\sigma_i)\cdot x_i$$

$$= \sum_{i=1}^{N} t_i(1-\sigma_i)x_i - (1-t_i)\sigma_i x_i$$

$$t_i x_i - t_i\sigma_i x_i - \sigma_i x_i + t_i\sigma_i x_i = (t_i - \sigma_i)x_i$$

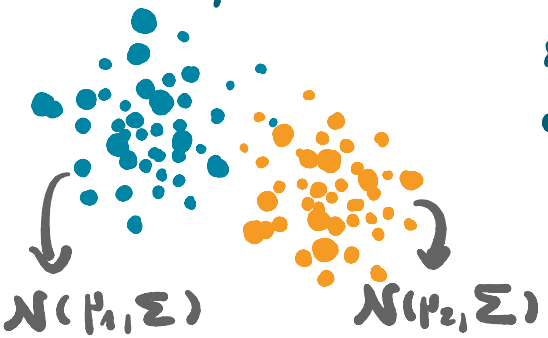
$$= \sum_{i=1}^{N}(t_i - \sigma_i)x_i$$

Lamentablemente no es llegar e igualar a cero en este caso, pero podemos usar el método de Gradiente Estocástico:

$$\theta \longmapsto \theta - \eta(t_i - \sigma_i)x_i$$

Se dice que regresión logística es un modelo **discriminativo** porque aprende las probabilidades condicionales $p(C_k|x)$ directamente, a diferencia del approach **GENERATIVO** donde modelaremos $p(x, C_k|\theta)$ y aplicaremos Bayes para obtener un estimador.

Volvamos al caso inicial. Asumimos que nuestros datos son una mixtura Gaussiana, una Gaussiana para cada clase, con (priors): $p(C_1) = \alpha$, $p(C_2) = 1-\alpha$ y $P(x|C_k) = \mathcal{N}(x; \mu_k, \Sigma)$



de este modo:

$$p(x,t) = p(t=C_1)p(x|C_1) + p(t=C_2)p(x|C_2)$$

Con esto nuestros parámetros a aprender son: $\mu_{1,2}, \Sigma$ y α

Recordando la densidad de una Gaussiana tenemos

$$p(x|C_k) = \frac{1}{(2\pi)^{\frac{N}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_k)\Sigma^{-1}(x-\mu_k)^T\right)$$

y además,

$$\log(p(x|C_k)) = \frac{1}{2}\left[\log\left(\frac{1}{(2\pi)^{\frac{N}{2}}|\Sigma|^{\frac{1}{2}}}\right) - (x-\mu_k)\Sigma^{-1}(x-\mu_k)^T\right]$$

juntando lo anterior tenemos:

$$\theta_{MAP} = \arg\max_\theta \prod_{i=1}^N L_i(\theta) = \arg\max_\theta \prod_{i=1}^N \underbrace{p(x_i, C_1|\theta)}_{p(x_i|C_1)p(C_1)}^{t_i} \underbrace{p(x_i, C_2|\theta)}_{p(x_i|C_2)p(C_2)}^{1-t_i}$$

$$\begin{aligned}\log() &= \arg\max_\theta \sum_{i=1}^N t_i\left(\log(p(x_i|C_1)) + \log(p(C_1))\right) \\ &\quad + (1-t_i)\left(\log(p(x_i|C_2) + \log(p(C_2))\right) \\ &= \arg\max_\theta \sum_{i=1}^N t_i \log\alpha + (1-t_i)\log(1-\alpha) \\ &\quad + t_i\frac{1}{2}\left[\log\left(\frac{1}{(2\pi)^{\frac{N}{2}}|\Sigma|^{\frac{1}{2}}}\right) - (x-\mu_1)\Sigma^{-1}(x-\mu_1)^T\right] \\ &\quad + (1-t_i)\frac{1}{2}\left[\log\left(\frac{1}{(2\pi)^{\frac{N}{2}}|\Sigma|^{\frac{1}{2}}}\right) - (x-\mu_2)\Sigma^{-1}(x-\mu_2)^T\right]\end{aligned}$$

TOMAMOS EL GRADIENTE E IGUALAMOS A CERO.

OJO: ESTO DEBEMOS HACERLO PARA CADA COMPONENTE (TOMAMOS EN CADA CASO SÓLO LOS TÉRMINOS RELEVANTES)

- $$\frac{\partial \log L}{\partial \alpha} = \frac{\partial}{\partial \alpha}\sum_{i=1}^N t_i \log\alpha + (1-t_i)\log(1-\alpha)$$
$$= \sum_{i=1}^N t_i \cdot \frac{1}{\alpha} - (1-t_i)\frac{1}{1-\alpha}$$

IMPONIENDO QUE $\frac{\partial L}{\partial \theta} = 0$ QUEDA $\frac{1}{\alpha}\sum_{i=1}^N t_i = \frac{1}{1-\alpha}\sum_{i=1}^N (1-t_i)$

$$\Leftrightarrow (1-\alpha)\sum_{i=1}^N t_i = \alpha\sum_{i=1}^N (1-t_i)$$

$$\Leftrightarrow \underbrace{\sum_{i=1}^N t_i}_{N_1} - \alpha\cancel{\sum_{i=1}^N t_i} = \alpha N - \alpha\cancel{\sum_{i=1}^N t_i}$$

N_1 = cantidad de datos con etiqueta 1

$$\Leftrightarrow \alpha = \frac{N_1}{N}$$ aka, proporción de datos C_1

ESTO JUSTIFICA LA INTUICIÓN DE QUE NUESTRO PRIOR ESTÁ DETERMINADO POR LA FRECUENCIA RELATIVA POR CLASE.

- $\frac{\partial \text{Log}}{\partial \mu_1} = \frac{\partial}{\partial \mu_1}\left(-\frac{1}{2}\sum_{i=1}^{N} t_i (x_i-\mu_1)\Sigma^{-1}(x_i-\mu_1)^T\right)$

$$= \sum_{i=1}^{N} t_i(\Sigma^{-1}(x_i-\mu_1)) = \Sigma^{-1}\sum_{i=1}^{N}(t_i x_i - t_i\mu_1)$$

Imponiendo que valga cero queda:

$$\underbrace{\sum_{i=1}^{N} t_i x_i}_{\sum_{x_i\in C_1} x_i} = \mu_1 \underbrace{\sum_{i=1}^{N} t_i}_{N_1} \iff \mu_1 = \frac{1}{N_1}\sum_{x_i\in C_1} x_i$$

} La media empírica!

Análogamente obtenemos $\mu_2 = \frac{1}{N_2}\sum_{x_i\in C_2}^{N} x_i$.

Cuando modelamos la distribución de los datos, i.e., damos una expresión para la probabilidad de un punto de dato dado, llamamos a esto un **Modelo generativo.**

La ventaja de lo que hicimos es que ahora conocemos la estructura probabilística de los datos, gracias a la cual podemos generar nuevos puntos de datos con métodos como Box-Muller. Muchas bibliotecas Python como numpy, scipy y pytorch tienen incorporados métodos para samplear Gaussianas.

Para cerrar con las ventajas de modelar nuestros datos de manera generativa, veamos que podemos deducir una forma cerrada de un clasificador con la misma forma que lo que teníamos en regresión logística:

$$\mathbb{P}(C_1|x) = \frac{1}{1+e^{-a^T x - b}}$$

En vez de usar el método de gradiente Estocástico, encontraremos una expresión buena para a y b recordando el teorema de Bayes:

$$\mathbb{P}(C_k|x) = \frac{\mathbb{P}(x|C_k)\,\mathbb{P}(C_k)}{\mathbb{P}(x)} = \frac{\mathbb{P}(x|C_k)\,\mathbb{P}(C_k)}{\sum_{j=1}^{k}\mathbb{P}(x|C_j)\,\mathbb{P}(C_j)}$$

para el caso de clasificación binaria queda

$$\mathbb{P}(C_1|x) = \frac{\mathbb{P}(x|C_1)\,\mathbb{P}(C_1)}{\mathbb{P}(x|C_1)\,\mathbb{P}(C_1)+\mathbb{P}(x|C_2)\,\mathbb{P}(C_2)} = \frac{1}{1+\frac{\mathbb{P}(x|C_2)\,\mathbb{P}(C_2)}{\mathbb{P}(x|C_1)\,\mathbb{P}(C_1)}}$$

Esta último expresión nos sugiere plantear lo siguiente:

$$\frac{\mathbb{P}(x|C_2)\,\mathbb{P}(C_2)}{\mathbb{P}(x|C_1)\,\mathbb{P}(C_1)} = \exp(-a^T x - b)$$

pues

$$\mathbb{P}(C_k|x) = \frac{1}{1+e^{-a^T x-b}} = \frac{1}{1+\frac{\mathbb{P}(x|C_2)\,\mathbb{P}(C_2)}{\mathbb{P}(x|C_1)\,\mathbb{P}(C_1)}}$$

Resolvamos $\frac{\mathbb{P}(x|C_2)\,\mathbb{P}(C_2)}{\mathbb{P}(x|C_1)\,\mathbb{P}(C_1)} = \exp(-a^T x + b)$

Primero asumimos distribuciones Gaussianas para $\mathbb{P}(x|C_k)$, con matriz de covarianza común:

$$p(x|C_k) = \frac{1}{(2\pi)^{\frac{N}{2}}|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_k)\Sigma^{-1}(x-\mu_k)^T\right)$$

La gracia de asumir covarianza común es que se simplifica la expresión:

$$\frac{\mathbb{P}(x|C_2)\mathbb{P}(C_2)}{\mathbb{P}(x|C_1)\mathbb{P}(C_1)} = \frac{\cancel{(2\pi)^{-\frac{N}{2}}}\cancel{|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_2)\Sigma^{-1}(x-\mu_2)^T\right)\mathbb{P}(C_2)}{\cancel{(2\pi)^{-\frac{N}{2}}}\cancel{|\Sigma|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(x-\mu_1)\Sigma^{-1}(x-\mu_1)^T\right)\mathbb{P}(C_1)}$$

Luego tomando logaritmo tenemos:

$$\log\left(\frac{\exp\left(-\frac{1}{2}(x-\mu_2)\Sigma^{-1}(x-\mu_2)^T\right)\mathbb{P}(C_2)}{\exp\left(-\frac{1}{2}(x-\mu_1)\Sigma^{-1}(x-\mu_1)^T\right)\mathbb{P}(C_1)}\right) = -a^Tx - b$$

$$\Leftrightarrow \underbrace{-\frac{1}{2}(x-\mu_2)\Sigma^{-1}(x-\mu_2)^T + \log\left(\frac{\mathbb{P}(C_2)}{\mathbb{P}(C_1)}\right) + \frac{1}{2}(x-\mu_1)\Sigma^{-1}(x-\mu_1)^T}_{} = -a^Tx - b$$

$$\frac{1}{2}\left(x^T\Sigma^{-1}(\mu_1-\mu_2) + (\mu_1-\mu_2)^T\Sigma^{-1}x - \mu_1\Sigma^{-1}\mu_1 + \mu_2\Sigma^{-1}\mu_2 + \log\left(\frac{\mathbb{P}(C_2)}{\mathbb{P}(C_1)}\right)\right)$$

$$\Leftrightarrow (\mu_1-\mu_2)^T\Sigma^{-1}x + \frac{1}{2}\left(\mu_2\Sigma^{-1}\mu_2 - \mu_1\Sigma^{-1}\mu_1\right) + \log\left(\frac{\mathbb{P}(C_1)}{\mathbb{P}(C_2)}\right) = -a^Tx - b$$

Con esto, $-a = (\mu_1-\mu_2)^T\Sigma^{-1}$

$$-b = \frac{1}{2}\left(\mu_2\Sigma^{-1}\mu_2 - \mu_1\Sigma^{-1}\mu_1\right) + \log\left(\frac{\mathbb{P}(C_1)}{\mathbb{P}(C_2)}\right)$$

Concluimos con esto la forma cerrada de nuestro clasificador, dada la estructura generativa de los datos.

Observación: La generalización multiclase de regresión logística es la función **Softmax** (o exponencial generalizada).

$$\mathbb{P}(C_k|x) = \frac{\exp(s_i)}{\sum_{j=1}^{k}\exp(s_j)}, \text{ con } s_i = \log(\mathbb{P}(x|C_i)\mathbb{P}(C_i))$$