Etiquetación con modelos de máxima entropía

Miguel A. Alonso

Departamento de Computación, Facultad de Informática, Universidade da Coruña

Índice

Modelos de Máxima Entropía: MaxEnt

- Es un marco probabilístico de aprendizaje automático
- Se basa en **regresión logística multinomial**
- Cuando se utiliza para clasificar secuencias de datos suele tomar la forma de un modelo de Markov de máxima entropía o MEMM
- En el caso de PoS tagging:
  - la sencuencia a clasificar son las palabras de un texto
  - Se trata de asignar a cada palabra su PoS

Modelos de Máxima Entropía: MaxEnt

- MaxEnt pertenece a la familia de los clasificadores **exponenciales** o **log-lineales**
  - extrae un conjunto de características (**features**) de la entrada
  - las combina de modo **lineal**
  - utiliza el resultado como un **exponente**
- Dada una entrada x con features f_i ponderadas mediante w_i, la probabilidad de asignar a x la clase c es

$$P(c \mid x) = \frac{1}{Z} \exp\left(\sum_i w_i f_i\right)$$

  donde Z es un factor normalizador para hacer que todo sume 1
- Si el resultado pertenece a un conjunto discreto se habla de **clasificación**, si es un conjunto real se habla de **regresión**

Lectura recomendada

- Daniel Jurafsky and James H. Martin
  capítulo 6 de *Speech and Language Processing. Second Edition*
  Pearson Education, Upper Saddle River, New Jersey, 2009

Ejemplo de regresión lineal

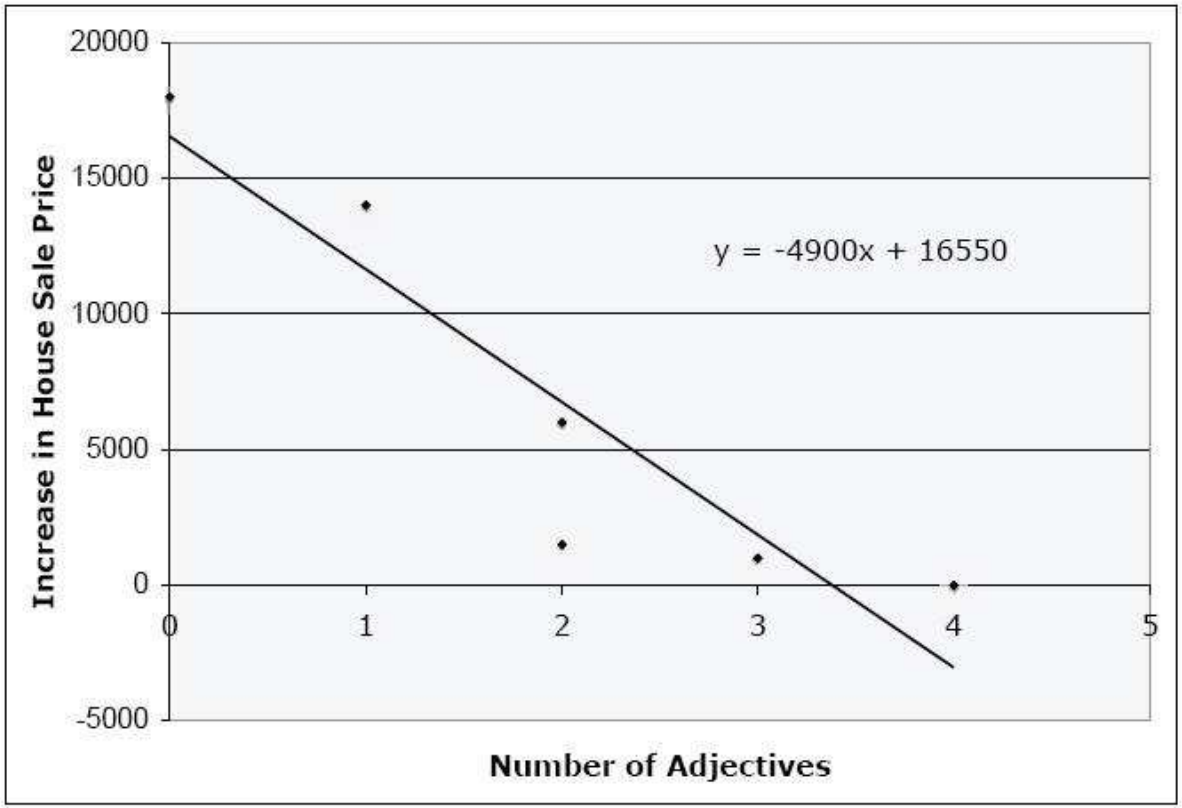
En *Freakonomics* se sugiere que las palabras de los anuncios de compra-venta pueden ayudar a predecir si se venderán por más o menos del precio anunciado. La hipótesis es que las palabras de significado vago (bonita, acogedora, luminosa, ...) se utilizan para enmascarar la ausencia de propiedades positivas concretas.

Supongamos los datos siguientes datos (ficticios):

# of Vague Adjectives	Amount House Sold Over Asking Price
4	0
3	$1000
2	$1500
2	$6000
1	$14000
0	$18000

Ejemplo de regresión lineal

de los que podemos obtener la **línea de regresión** siguiente:



$precio = w_0 + w_1 * f_1$ donde $w_0 = 16.550$ y $w_1 = -4.900$ y $F_1 = \#\text{adjetivos}$

Regresión lineal múltiple

- Si hay más de una feature se habla de **regresión lineal múltiple**

$$y = \sum_{i=0}^{N} w_i \ f_i$$

donde en general $w_0 = 1$, que se puede escribir como el producto escalar de los vectores $\vec{w}$ y $\vec{f}$

- El aprendizaje de los pesos $\vec{w}$ se suele hacer por minimización del error cuadrático medio entre los valores previstos y los observados

Regresión logística

- No estamos interesados en y sino en $P(y)$, la probabilidad de y, donde y es un conjunto de valores discretos (clasificación probabilística)
- En consecuencia, nos gustaría calcular $P(y = \text{true} \mid x) = \sum_{i=0}^{N} w_i f_i$ ... pero resulta que $\sum_{i=0}^{N} w_i f_i$ produce valores entre $-\infty$ e ∞
- Solución: hacer que $\sum_{i=0}^{N} w_i f_i$ sea una ratio entre dos probabilidades (y no una probabilidad)

Regresión logística

- La ratio que utilizamos son los odds de una probabilidad:

$$\frac{P(y = \text{true} \mid x)}{1 - P(y = \text{true} \mid x)} = \sum_{i=0}^{N} w_i f_i$$

- La parte izquierda puede variar entre 0 e ∞, para hacer que varíe entre $-\infty$ e ∞ aplicamos logaritmos:

$$\ln\left(\frac{P(y = \text{true} \mid x)}{1 - P(y = \text{true} \mid x)}\right) = \sum_{i=0}^{N} w_i f_i$$

- **La regresión logística es un modelo de regresión que usa una función lineal para estimar el logaritmo de los odds de una probabilidad**

Obtención de $P(y = \text{true} \mid x)$

$$\ln\left(\frac{P(y = \text{true} \mid x)}{1 - P(y = \text{true} \mid x)}\right) = \sum_{i=0}^{N} w_i f_i$$

$$\frac{P(y = \text{true} \mid x)}{1 - P(y = \text{true} \mid x)} = \exp\left(\sum_{i=0}^{N} w_i f_i\right)$$

$$P(y = \text{true} \mid x) = (1 - P(y = \text{true} \mid x)) \exp\left(\sum_{i=0}^{N} w_i f_i\right)$$

$$P(y = \text{true} \mid x) = \exp\left(\sum_{i=0}^{N} w_i f_i\right) - P(y = \text{true} \mid x) \exp\left(\sum_{i=0}^{N} w_i f_i\right)$$

$$P(y = \text{true} \mid x) + P(y = \text{true} \mid x) \exp\left(\sum_{i=0}^{N} w_i f_i\right) = \exp\left(\sum_{i=0}^{N} w_i f_i\right)$$

Obtención de $P(y = \text{true} \mid x)$

$$P(y = \text{true} \mid x)(1 + \exp\left(\sum_{i=0}^{N} w_i f_i\right) = \exp\left(\sum_{i=0}^{N} w_i f_i\right)$$

$$P(y = \text{true} \mid x) = \frac{\exp\left(\sum_{i=0}^{N} w_i f_i\right)}{1 + \exp\left(\sum_{i=0}^{N} w_i f_i\right)}$$

y de aquí

$$P(y = \text{false} \mid x) = \frac{1}{1 + \exp\left(\sum_{i=0}^{N} w_i f_i\right)}$$

para que $P(y = \text{true} \mid x) + P(y = \text{false} \mid x) = 1$

Obtención de $P(y = \text{true} \mid x)$

Si tomamos la expresión $P(y = \text{true} \mid x) = \frac{\exp\left(\sum_{i=0}^{N} w_i f_i\right)}{1+\exp\left(\sum_{i=0}^{N} w_i f_i\right)}$ y dividimos numerador y denominador por $\exp\left(-\sum_{i=0}^{N} w_i f_i\right)$ obtenemos

$$P(y = \text{true} \mid x) = \frac{1}{1 + \exp\left(-\sum_{i=0}^{N} w_i f_i\right)}$$

que ya está en forma de **función logística** $\frac{1}{1+e^{-x}}$

Como consecuencia

$$P(y = \text{false} \mid x) = \frac{\exp\left(-\sum_{i=0}^{N} w_i f_i\right)}{1 + \exp\left(-\sum_{i=0}^{N} w_i f_i\right)}$$

Aprendizaje en regresión logística

- Se resuelve a través de técnicas matemáticas complejas, de programación no lineal, denominadas de **optimización convexa**
- Se suele utilizar algoritmos como L-BFGS, algoritmos de gradiente ascendente, algoritmos de gradiente conjugado, algoritmos de escalado iterativo, ...
- Se suelen suavizar los pesos w_i mediante **regularización** con el fin de penalizar los pesos grandes

Clasificación con regresión logística

- El resultado será la clase *true* si $P(y = \text{true} \mid x) > P(y = \text{false} \mid x)$
- o lo que es lo mismo si $\frac{P(y=\text{true}|x)}{P(y=\text{false}|x)} > 1$
- y como $\frac{P(y=\text{true}|x)}{1-P(y=\text{true}|x)} = \exp(\sum_{i=0}^{N} w_i f_i)$
- se trata entonces de ver si $\exp(\sum_{i=0}^{N} w_i f_i) > 1$
- lo cual implica ver si se cumple

$$\sum_{i=0}^{N} w_i f_i > 0$$

Regresión logística multinomial

Se aplica cuando queremos clasificar una observación en **más de dos clases**

$$P(c \mid x) = \frac{1}{Z} \exp\left(\sum_{i=0}^{N} w_{ci} f_i\right)$$

donde hacemos que los pesos de f_i dependan de la clase $c \in C$ y

$$Z = \sum_{c' \in C} \exp\left(\sum_{i=0}^{N} w_{c'i} f_i\right)$$

lo que resulta en

$$P(c \mid x) = \frac{\exp\left(\sum_{i=0}^{N} w_{ci} f_i\right)}{\sum_{c' \in C} \exp\left(\sum_{i=0}^{N} w_{c'i} f_i\right)}$$

Regresión logística multinomial en PoS tagging

- En PoS tagging las features f_i no son reales sino discretas
- Más en concreto, son booleanas, indican si una propiedad está presente o no
- Denotamos por $f_i(c, x)$ la **función indicador** que nos indica si una feature i está presente para al clase c en x.

$$P(c \mid x) = \frac{\exp\left(\sum_{i=0}^{N} w_{ci} f_i(c, x)\right)}{\sum_{c' \in C} \exp\left(\sum_{i=0}^{N} w_{c'i} f_i(c, x)\right)}$$

- La ventaja con respecto a otros modelos es que las funciones indicador pueden referirse a "prácticamente cualquier cosa"

Ejemplo de clasificación no secuencial

Secretariat/NNP is/BEZ expected/VBN to/TO race/?? tomorrow/RB

$$f_1(c,x) = \left\{ \begin{array}{l} 1 \text{ si } w_i = \text{"race"} \ \& \ c = \text{NN} \\ 0 \text{ otherwise} \end{array} \right\}$$

$$f_2(c,x) = \left\{ \begin{array}{l} 1 \text{ si } t_{i-1} = \text{TO} \ \& \ c = \text{VB} \\ 0 \text{ otherwise} \end{array} \right\}$$

$$f_3(c,x) = \left\{ \begin{array}{l} 1 \text{ si suffix}(w_i) = \text{"ing"} \ \& \ c = \text{VBG} \\ 0 \text{ otherwise} \end{array} \right\}$$

$$f_4(c,x) = \left\{ \begin{array}{l} 1 \text{ si is\_lower\_case}(w_i) \ \& \ c = \text{VB} \\ 0 \text{ otherwise} \end{array} \right\}$$

$$f_5(c,x) = \left\{ \begin{array}{l} 1 \text{ si } w_i = \text{"race"} \ \& \ c = \text{VB} \\ 0 \text{ otherwise} \end{array} \right\}$$

$$f_6(c,x) = \left\{ \begin{array}{l} 1 \text{ si } t_{i-1} = \text{TO} \ \& \ c = \text{NN} \\ 0 \text{ otherwise} \end{array} \right\}$$

Ejemplo de clasificación no secuencial

Dada la entrada actual $x =$"race" y suponiendo los siguientes pesos:

		f1	f2	f3	f4	f5	f6
VB	**f**	0	1	0	1	1	0
VB	**w**		.8		.01	.1	
NN	**f**	1	0	0	0	0	1
NN	**w**	.8					-1.3

$$P(NN \mid x) = \frac{e^{0.8}e^{-1.3}}{e^{0.8}e^{-1.3} + e^{0.8}e^{0.01}e^{0.1}} = 0.20$$

$$P(VB \mid x) = \frac{e^{0.8}e^{0.01}e^{0.1}}{e^{0.8}e^{-1.3} + e^{0.8}e^{0.01}e^{0.1}} = 0.80$$

Al utilizar MaxEnt para clasificación obtenemos una distribución de probabilidad sobre las clases

Clasificación secuencial: HMM vs. MEMM

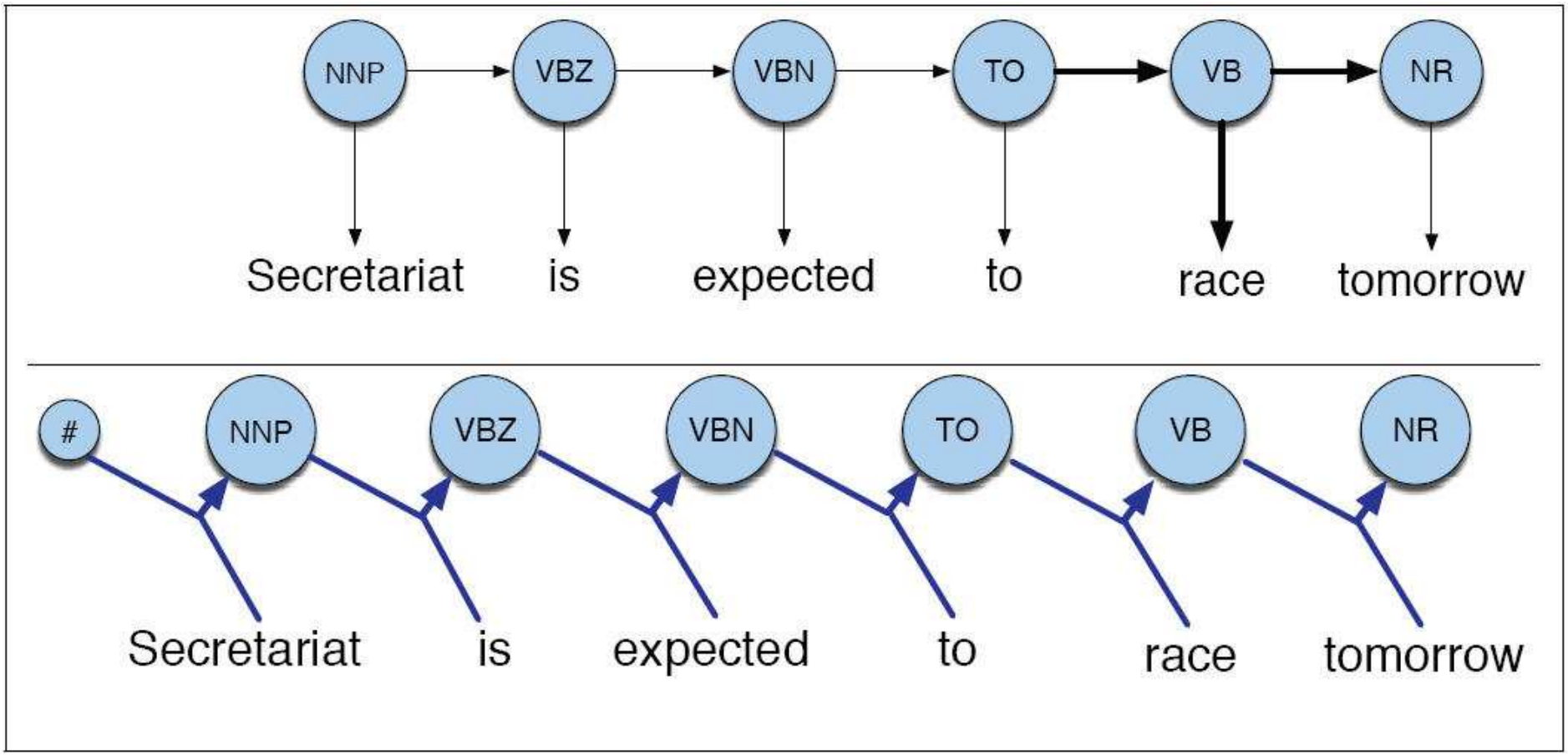
- Hidden Markov Models (HMM) son un modelo generativo:

$$\begin{aligned} T = & \arg\max_T P(T \mid W) \\ & \arg\max_T P(W \mid T) P(T) \\ & \arg\max_T \prod_i P(w_i \mid t_i) \prod_i (t_i \mid t_{i-1}) \end{aligned}$$

- MaxEnt Markov Models (MEMM) son un modelo discriminativo:

$$\begin{aligned} T = & \arg\max_T P(T \mid W) \\ & \arg\max_T \prod_i (t_i \mid w_i, t_{i-1}) \end{aligned}$$

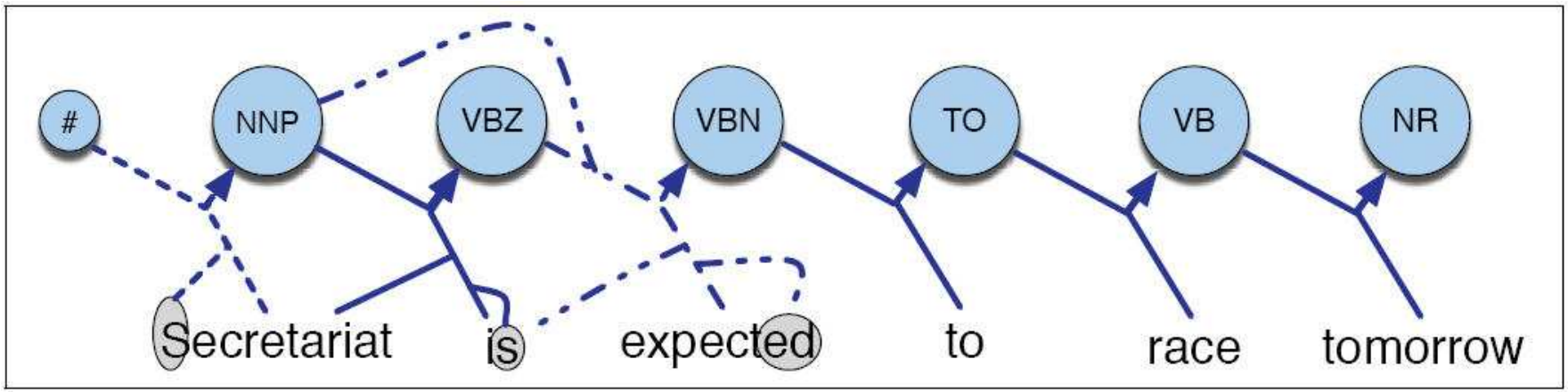
HMM


NNP
VBZ
VBN
TO
VB
NR
Secretariat
is
expected
to
race
tomorrow
#
NNP
VBZ
VBN
TO
VB
NR
Secretariat
is
expected
to
race
tomorrow

MEMM

Ventajas de los EMMM

- Los HMM manejan probabilidades de emisión y probabilidades de transición
- Los MEMM pueden incoporar propabilidades sobre las features que queramos:



por lo que la probabilidad de trasición de un estado q a un estado q' que produce la observación o se define como:

$$P(q \mid q', o) = \frac{1}{Z(q', 0)} \exp\left(\sum_i w_i f_i(q, o)\right)$$

Ejecución de un MEMM

- Viterbi para HMM:

$$\delta_{t+1}(j) = \max_{1 \le i \le N} \delta_t(i)\, a_{ij}\, b_j(o_{t+1}) \quad t = 1, 2, \ldots, T-1, \quad 1 \le j \le N$$

$$= \max_{1 \le i \le N} \delta_t(i)\, P(t_j \mid t_i) P(o_{t+1} \mid t_j) \quad t = 1, 2, \ldots, T-1, \quad 1 \le j \le N$$

- Viterbi para MEMM:

$$\delta_{t+1}(j) = \max_{1 \le i \le N} \delta_t(i)\, P(t_j \mid t_i, o_{t+1}) \quad t = 1, 2, \ldots, T-1, \quad 1 \le j \le N$$

¿Porqué le llaman amor cuando quieren decir sexo?

O dicho de otra manera **¿porqué le llaman máxima entropía cuando quieren decir regresión logística multinomial?**

¿Dónde está la entropía?

$$H(x) = -\sum_{x} P(x) \log_2 P(x)$$

Pues porque...

- La intuición de MaxEnt es que el modelo probabilístico debe seguir las restricciones que impongamos, pero no debe asumir nada especial sobre todo lo demás (i.e. debe dejar todo lo demás con la máxima entropía posible).
- Formalmente:

  Seleccionar un modelo de un conjunto de distribuciones de probabilidad permitidas, elegir el el modelo $p^* \in \mathcal{C}$ con máxima entropía $H(p)$:

$$p^* = \arg\max_{p \in \mathcal{C}} H(p)$$

La solución a este problema es la distribución de probabilidad de un modelo de regresión logística multinomial cuyos pesos maximizan la verosimilitud de los datos de entrenamiento

The end