Entropía condicional

Salvador López Mendoza

Mayo de 2018

Agregar contexto

¿Qué pasa si se sabe más acerca de un segmento de texto?

La palabra *carne*, ¿está presente o ausente en un segmento que contiene la palabra *come*?

¿La presencia de *come* ayuda a predecir la presencia de *carne*?

La presencia de *come*, ¿reduce la incertidumbre acerca de *carne* ($H(X_{carne})$)?

La ausencia de la palabra *come*, ¿también ayuda?

PEUVI

Entropía condicional

En vez de considerar $p(X_{carne} = 1)$ se utiliza $p(X_{carne} = 1|X_{come} = 1)$

En vez de considerar $p(X_{carne} = 0)$ se utiliza $p(X_{carne} = 0|X_{come} = 1)$

La entropía condicional es:

$$H(X_{carne}|X_{come} = 1) = -p(X_{carne} = 0|X_{come} = 1)log_2 p(X_{carne} = 0|X_{come} = 1) - p(X_{carne} = 1|X_{come} = 1)log_2 p(X_{carne} = 1|X_{come} = 1)$$

$H(X_{carne}|X_{come} = 0)$ se define de forma similar.

PEUVI

Definición completa

$$H(X_{carne}|X_{come}) = \sum_{u\in\{0,1\}} [p(X_{come} = u)H(X_{carne}|X_{come} = u)]$$

$$= \sum_{u\in\{0,1\}} [p(X_{come} = u) \sum_{v\in\{0,1\}} [-p(X_{carne} = v|X_{come} = u)log_2 p(X_{carne} = v|X_{come} = u$$

Para cualquier par de variables aleatorias discretas, X e Y, se tiene que $H(X) \geq H(X|Y)$.

¿Cuál es el posible valor mínimo de $H(X|Y)$?

Capturar relaciones sintagmáticas

$$H(X_{carne}|X_{come}) = \sum_{u \in \{0,1\}} [p(X_{come} = u) H(X_{carne}|X_{come} = u)]$$

¿Cuál es más pequeña? ¿$H(X_{carne}|X_{el})$ o $H(X_{carne}|X_{come})$?

¿Para qué palabra w, $H(X_{carne}|x_w)$ alcanza su mínimo?

¿Para qué palabra w, $H(X_{carne}|x_w)$ alcanza su máximo?

PEUVI

Minando relaciones sintagmáticas

- Para cada palabra w_1,
  - Para cada otra palabra w_2, calcular la entropía condicional $H(X_{w_1}|X_{w_2})$.
  - Ordenar todas las palabras candidatas en orden ascendente de los valores de $H(X_{w_1}|X_{w_2})$.
  - Tomar las palabras candidatas con los valores más altos.
    Esas palabras tienen el potencial de tener relaciones sintagmáticas con w_1.

¡Cuidado!

$H(X_{w_1}|X_{w_2})$ y $H(X_{w_1}|X_{w_3})$ son comparables.

Pero $H(X_{w_1}|X_{w_2})$ y $H(X_{w_3}|X_{w_2})$ no son comparables.

¿Cómo podemos minar las k relaciones sintagmáticas más fuertes en una colección de documentos?

PEUVI