Ciencia de Datos y Lenguaje Natural

Teoría de la Información - 2

Grupo PLN - INCO

Universidad de la República

Variantes de la entropía

Veremos algunas cantidades asociadas a la entropía y aplicaciones donde se han utilizado.

- Perplejidad
- Entropía conjunta y condicional
- Entropía cruzada ($cross - entropy$)
- Divergencia de Kullback Leibler
- Información mutua

Perplejidad y modelos de lenguaje

- La perplejidad, PP(X) con X variable aleatoria, es una cantidad asociada a la entropía cruzada, y que varía en el mismo sentido:

$$PP(M) = 2^{H(L,M)}$$

  siendo L el lenguaje y M el modelo
- La perplejidad es creciente según la entropía. Tiene básicamente ´la misma información´
- Se ha utilizado intensivamente en modelos de lenguaje.

Modelos de Lenguaje

- Un modelo probabilístico de lenguaje asigna probabilidades a secuencias de una lengua.
- Suelen ser secuencias de palabras, podrían ser también de caracteres.
- Los modelos de lenguaje tienen variadas aplicaciones en PLN.

Modelos de Lenguaje

- Un modelo de lenguaje es una distribución de probabilidad sobre secuencias de símbolos del lenguaje.
- Una cantidad de interés es $P(w_n|w_{n-1}...w_1)$, o sea, la probabilidad de la próxima palabra dada una secuencia de palabras.
- Se calcula a partir de los datos de un corpus
- Nos restringiremos en lo que sigue a modelos de n-gramas, que son más tradicionales.

Modelos de Lenguaje

- Las probabilidades se estiman a partir de las frecuencias relativas de las palabras en el corpus (estimador de máxima verosimilitud) con factores de descuento (*smoothing*)
- Se hacen simplificaciones de modo de tener 'masa crítica' como para hacer los cálculos.
- Hipótesis markoviana de bigramas $P(w_n|w_{n-1}...w_1) \approx P(w_n|w_{n-1})$
- Hipótesis markoviana de trigramas $P(w_n|w_{n-1}...w_1) \approx P(w_n|w_{n-1}w_{n-2})$

Modelos de Lenguaje y perplejidad

- Los modelos de lenguaje se evalúan, como ocurre en general con los sistemas de aprendizaje, mediante un conjunto de testeo $W = w_1, w_2, ..., w_n$
- En vez de maximizar directamente la probabilidad del conjunto de testeo, se minimiza la perplejidad.
- Se define la perplejidad PP(W), como $PP(W) = \sqrt[N]{\frac{1}{P(w_1 w_2 ... w_n)}}$ (ver J&M 4.7 para la relación entre ambas definiciones)
- Para calcular la probabilidad de la secuencia de palabras $w_1 w_2 ... w_n$ se utilizan simplificaciones a n-grama

Entropía conjunta y condicional

La entropía conjunta del par de variables aleatorias (X, Y) y las entropías condicionales de una variable respecto a la otra están vinculadas por las siguientes fórmulas

$$H(X, Y) = H(X|Y) + H(Y) = H(Y|X) + H(X)$$

Se cumple: $H(X|Y) \leq H(X)$, con igualdad cuando son independientes

Entropía cruzada

La entropía cruzada de una distribución p respecto a una distribución q, sobre iguales conjuntos de base, se define como
$H(p, q) = -E_p \log q$

$$H(p, q) = -\sum_{i=1}^{n} p(x_i) \log_2 q(x_i).$$

Divergencia de Kullback Leibler, o entropía relativa

Es una medida de discrepancia entre 2 variables aleatorias, que no satisface los axiomas de una distancia (no es simétrica).

Si p(x) y q(x) son dos distribuciones de probabilidad definidas sobre el mismo conjunto de valores x, su entropía relativa es

$$D_{KL}(p||q) = -\sum_{x} p(x) \log_2(p(x)/q(x_i)).$$

Notar que $D_{KL}(p||q) \geq 0$,

y si $p(x) = q(x)$ su distancia $D_{KL}(p||q) = 0$

Información mutua

En algún sentido la información mutua $I(X;Y)$ es la intersección entre H(X) y H(Y), dado que representa su dependencia estadística.

$$I(X;Y) = H(X) - H(X|Y)$$
$$I(X;Y) = H(Y) - H(Y|X) = I(Y;X)$$
$$I(X;Y) = H(X) + H(Y) - H(X,Y)$$

Se usa para ver correlaciones entre variables.

Información mutua puntual

La información mutua puntual $PMI(x; y)$ de dos valores x y y de las variables aleatorias X y Y intenta discriminar las ocurrencias conjuntas de ambos valores simplemente debidas al azar.

$$PMI(x; y) = \log_2 \frac{p(x, y)}{p(x)p(y)}$$

Información mutua puntual, propiedades

- La información mutua $I(X;Y)$ es la esperanza en la distribución conjunta p(X,Y) de PMI(x;y)
- PMI(x;y) $= 0$ cuando X e Y son independientes
- PMI(x;y) puede tomar valores negativos
- PMI(x;y) toma su valor máximo cuando x e y están perfectamente asociados

Información mutua puntual, aplicaciones

- Es ampliamente utilizado
- Un ejemplo de uso es en el descubrimiento de colocaciones lingüísticas
- Las colocaciones son palabras con asociación de algún modo convencionalizada
- por ejemplo, en el caso de nombre y adjetivo

Algunas colocaciones sustantivo/adjetivo

pérdida	irreparable
atentos	saludos
testigo	ocular
flagrante	delito
mercado	negro
bajo	consumo

Información mutua puntual, problemas

- PMI puede ser negativo
- No tiene una interpretación clara un valor negativo

  Se usa entonces PPMI, variante positiva en la se llevan a 0 los valores negativos,
- PMI prioriza eventos de muy baja frecuencia

Información mutua puntual, normalización

$$PMI(x;y) = \log_2 \frac{p(x,y)}{p(x)p(y)}$$

Se normaliza de modo de atenuar el problema de la baja frecuencia

$$PMI(x;y) = \log_2 \frac{p(x,y)}{p(x)p(y)} / -\log_2 p(x,y)$$

https://svn.spraakdata.gu.se/repos/gerlof/pub/www/Docs/npmi-pfd.pdf