Informazio-Teoria

"Konputazio-Zientzien Metodo Matematikoak" irakasgairako oinarrizko kontzeptuak

Titulazioa: Informatikan ingeniaria
Konputazio Zientzia eta Adimen Artifiziala saila
Universidad del País Vasco - Euskal Herriko Unibertsitatea

Informazio-teoria

- Hainbat gaitan erabiliko ditugu informazio-teoriako kontzeptuak: K-NN sailkatzaileetan, sailkapen-zuhaitzetan, aldagai-aukeraketan, sailkatzaile Bayestarretan

Bibliografia

1. Informazio-kantitatea.
   http://en.wikipedia.org/wiki/Self−information
2. Entropia: $H(C)$, $H(C|X)$
   http://en.wikipedia.org/wiki/Entropy_(information_theory)
   http://en.wikipedia.org/wiki/Conditional_entropy
3. Elkarrekiko informazio-kantitatea: $I(X,C)$, $I(X,Y|C)$
   http://en.wikipedia.org/wiki/Mutual_information

1. Informazio-kantitatea

- Ziurgabetasuna murrizteko neurri moduan informazio-kantitatea erabiltzen da

Adibidea

Kutxa batean 9 bola beltz eta bola txuri bat daude. Bolak atera ondoren ez dira berriro kutxan jarriko

- Bola txuri bat ateratzen da. Gertakari honek informazio kantitate handia ematen du, aterako den hurrengo bolari buruzko ziurgabetasuna desagertu egiten delako
- Bola beltz bat ateratzen da. Gertakari honek informazio kantitate txikia ematen du, aterako den hurrengo bolari buruzko ziurgabetasuna antzeko mantentzen delako

1. Informazio-kantitatea

Formalizatuz

- C zorizko aldagaiak (z.a.) $c_1, \ldots, c_n$ balio posibleak hartzen baditu $p(c_1), \ldots, p(c_n)$ probabilitatez
- $\boxed{I(c_i) = -\log_2 p(c_i)}$
  - Baldin $p(c_i) \approxeq 1 \quad \Rightarrow \quad I(c_i) \approxeq 0$
  - Baldin $p(c_i) \approxeq 0 \quad \Rightarrow \quad I(c_i) \approxeq +\infty$
- Gertakari batek zenbat eta probabilitate handiagoa izan orduan eta informazio-kantitate txikiagoa emango du

2. Entropia: $H(C)$

- C zorizko aldagai (z.a.) diskretuak $c_1, \ldots, c_n$ balioak hartzen ditu $p(c_1), \ldots, p(c_n)$ prob.
- C aldagaiari dagokion $I(C)$ informazio-kantitatea zorizko aldagaiak $I(c_1), \ldots, I(c_n)$ balio posibleak hartzen ditu $p(c_1), \ldots, p(c_n)$ probabilitatez
- C zorizko aldagai diskretuaren $H(C)$ Shannon-en entropia (1948): dagokion $I(C)$-ren itxaropen matematikoa

$$H(C) = E(I(C)) = -\sum_{i=1}^{n} p(c_i) \log_2 p(c_i)$$

- Prob. banaketa bat emanik, ziurgabetasun maila neurtu
- $H(C) = 0 \iff \exists c_i$ non $p(c_i) = 1$, (denak klase berekoak)
- Baldin $p(c_i) = 0$, orduan $p(c_i) \log_2 p(c_i)$ indeterminazioari 0 balioa emango zaio

2. Baldintzapeko Entropia: $H(C|X)$

C, X eta (C,X) zorizko aldagaiak (z.a.)

- C: $c_1, \ldots, c_n$ balioak, $p(c_1), \ldots, p(c_n)$ probabilitatez
- X: $x_1, \ldots, x_m$ balioak, $p(x_1), \ldots, p(x_m)$ probabilitatez
- (C,X): $(c_1,x_1), \ldots, (c_1,x_m), \ldots, (c_n,x_1), \ldots, (c_n,x_m)$. $p(c_1,x_1), \ldots, p(c_1,x_m), \ldots, p(c_n,x_1), \ldots, p(c_n,x_m)$ prob.

$C|X=x_j$ baldintzapeko z.a. $p(c_1|x_j), \ldots, p(c_n|x_j)$ prob. C-ren entropia X-ren baldintzapean:

$$H(C|X) = \sum_{j=1}^{m} p(x_j) H(C|X=x_j) = -\sum_{i=1}^{n}\sum_{j=1}^{m} p(c_i,x_j) \log_2 \frac{p(c_i,x_j)}{p(x_j)}$$

3. Elkarrekiko informazio-kantitatea: $I(X,C)$

- X eta C bi zorizko aldagaien arteko elkarrekiko informazio-kantitateak aldagai baten balioa ezagutzeak beste aldagaiaren ziurgabetasuna zenbat murrizten duen neurtzen du
- Elkarrekiko informazioa simetrikoa da: $I(X,C) = I(C,X)$
- Elkarrekiko informazioa positiboa da: $I(X,C) \geq 0$

$$\boxed{I(X,C) = H(C) - H(C|X)}$$

Kalkuluak eginez zera lortzen da:

$$\boxed{I(X,C) = H(C) - H(C|X) = \sum_{i=1}^{n}\sum_{j=1}^{m} p(x_i,c_j)\log_2 \frac{p(x_i,c_j)}{p(x_i)\cdot p(c_j)}}$$

3. Elkarrekiko informazio-kantitatea: $I(X, Y|C)$

C klase-aldagaiaren baldintzapean

C klase-aldagaiaren baldintzapean X eta Y bi aldagai diskreturen artean dagoen elkarrekiko informazio-kantitatea honela definitzen da:

$$I(X, Y|C) = \sum_{i=1}^{n} \sum_{j=1}^{m} \sum_{k=1}^{r} p(x_i, y_j, c_k) \log_2 \frac{p(x_i, y_j|c_k)}{p(x_i|c_k) \cdot p(y_j|c_k)}$$