

Informazio-Teoria

“Datu Meatzaritza” irakasgairako oinarritzko kontzeptuak

Espezialitatea: **Konputazioa**, hirugarren ikasmaila
Titulazioa: **Informatika Ingeniaritzako Gradua**
Konputazio Zientzia eta Adimen Artifiziala saila
Universidad del País Vasco - Euskal Herriko Unibertsitatea

1/8

Informazio-teoria

- ▶ Irakasgaian zehar hainbat gaitan agertuko diren informazio-teoriako kontzeptuak: K-NN sailkatzaileetan, sailkapen-zuhaitzetan, aldagai-aukeraketan, sailkatzaile Bayestarretan
- ▶ Shannon-ek 1948. urtean sortua. Gaurregun asko erabiltzen diren kontzeptuak dira
- ▶ Abantaila handi bat dute: ez-parametrikoak dira

Oinarritzko kontzeptuak

1. Informazio-kantitatea
2. Entropia: $H(C)$, $H(C|X)$
3. Elkarrekiko informazio-kantitatea: $I(X, C)$, $I(X, Y|C)$

2/8

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

3/8

1. Informazio-kantitatea

Formalizatuz

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

4/8

2. Entropia: $H(C)$

- ▶ C zorizko aldagai (z.a.) diskretuak $c_1, \dots, c_n$ balio posibleak hartzen ditu $p(c_1), \dots, p(c_n)$ probabilitatez
- ▶ C aldagaiari dagokion $I(C)$ informazio-kantitatea zorizko aldagaiak $I(c_1), \dots, I(c_n)$ balio posibleak hartzen ditu $p(c_1), \dots, p(c_n)$ probabilitatez
- ▶ C zorizko aldagai diskretuaren $H(C)$ Shannon-en entropia (1948), honela definitzen da: dagokion $I(C)$ zorizko aldagaiaren itxaropen matematikoa

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

- ▶ $H(C) = 0 \iff \exists c_i \text{ non } p(c_i) = 1$
- ▶ Baldin $p(c_i) = 0$, orduan $p(c_i) \log_2 p(c_i)$ indeterminazioari 0 balioa emango zaio

5/8

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

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

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

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

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

6/8

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: $I(X, C) \geq 0$

$$I(X, C) = H(C) - H(C|X)$$

Kalkuluak eginez zera lortzen da:

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

7/8

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:

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

8/8