Információs rendszerek elméleti alapjai

Információelmélet

Az entrópia tulajdonságai

Jensen egyenlőtlenség

$h(x)$ konvex függvény,

$h(\mathrm{E}(\xi)) \le \mathrm{E}(h(\xi))$

$h(x)$ szigorúan konvex, akkor az egyenlő feltétele: $P(\xi = \mathrm{E}(\xi)) = 1$ $-\xi$ 1 val.gel konstans

konvexitás

$x, y \in [a,b], 0 < \lambda < 1$

$h(\lambda x + (1-\lambda)y) \le \lambda h(x) + (1-\lambda)h(y)$

szigorú konvexitás

$\mathrm{E}(\xi)$ pontban, $\lambda x + (1-\lambda)y = \mathrm{E}(\xi), \forall x, y$

h logaritmus, E az entrópia

Forrás jellemzése

A forrás közelítései:

Megfigyeljük a forrást M hosszú üzenetet véve.

Ebből közelítjük a forrás valószínűségi eloszlását, statisztikákból. (Stacionáris forrásokat vizsgálunk. Azaz időtől ne függjön az eloszlás

Az üzenet (diszkrét) lehetséges szimbólumai, x_1, x_2, …, x_n.

Forrás jellemzése

0-d rendű közelítés: csak n értékét használjuk fel.

A legbizonytalanabb az egyenletes eloszlás;

Egy szimbólum választás: ξ_i valószínűségi változó

$$P\left(\xi_i = x_j\right) = \frac{1}{n}, \quad j = 1,\ldots,n$$

N szimbólum egymás után (n^N lehetőség)

$$P\left(\xi_1 = x_{i_1}, \xi_2 = x_{i_2} \ldots, \xi_n = x_{i_N}\right) = n^{-N}, \quad j = 1,\ldots,n$$

Független , egyenletes, azonos eloszlású változó.

Forrás jellemzése

Entrópia : $H(\xi_i) = log_2 n$

$$H(\xi_1, \xi_2, \ldots, \xi_N) = NH(\xi) = N \, log_2 n, \quad j = 1$$

$$H(\xi) = log_2 n = \lim_{N \to +\infty} \left(\frac{1}{N} H(\xi_1, \xi_2, \ldots, \xi_N) \right)$$

1 szimbólumra jutó entrópia/bizonytalanság

Forrás jellemzése

Első-rendű közelítés

M hosszú üzenetet megfigyelünk

x_i előfordulásainak száma: $m_i \; \forall \; i=1,\ldots, n$

Rögzítjük ξ_i eloszlásának közelítését,

$P(\xi_i=x_j)=m_j/M=p_j \; \forall \; j=1,\ldots, n$

A vetületi eloszlások rögzítése mellett a legbizonytalanabb együttes eloszlás a független eloszlás

Forrás jellemzése

Első-rendű közelítés

Független , azonos eloszlású változók.

$$P\left(\xi_1 = x_{i_1}, \xi_2 = x_{i_2} \ldots, \xi_n = x_{i_N}\right) = \prod_{j=1}^{N} P\left(\xi_j = x_{i_j}\right) =$$

$$\prod_{i=1}^{n} p_i^{\mu_i}, \quad \mu_i = \text{az} \quad x_i \ \text{száma} \ \{x_{i_1}, x_{i_2} \ldots, x_{i_N}\} \text{-ben}$$

$$H\left(\xi_i\right) = -\sum_{j=1}^{m} p_j \log_2 p_j \cdots \text{Entrópia}$$

Forrás jellemzése

Együttes entrópia (függetlenség miatt additív)

Egy szimbólumra jutó entrópia:

$$H(\xi_1,\ldots,\xi_N) = -N\sum_{j=1}^{n} p_j \log_2 p_j$$

$$H(\xi) = \lim_{N\to\infty} \frac{1}{N} H(\xi_1,\ldots,\xi_N) = -\sum_{j=1}^{n} p_j \log_2 p_j$$

Forrás jellemzése

Másod-rendű közelítés:

Megfigyeljük az egymás utáni párok gyakoriságát,

m_{ij}: az x_i x_j előfordulásainak száma

m_i : az x_i előfordulásainak száma

A $p_{i|j} = m_{ij} / m_i$ hányados megfelel feltételes valószínűségnek:

$$P\left(\xi_t = x_j \middle| \xi_{t-1} = x_i\right) = \frac{P\left(\xi_t = x_j, \xi_{t-1} = x_i\right)}{P\left(\xi_{t-1} = x_i\right)}$$

Ezzel rögzítjük az átmeneti valószínűségeket. (átmenet valószínűség mátrix, sorok összege 1.

$$H\left(\xi_3 \middle| \xi_1, \xi_2\right) \leq H\left(\xi_3 \middle| \xi_2\right)$$

A 2’ Tétel szerint a legbizonytalanabb eloszlás, ami ennek eleget tesz az egylépéses homogén Markov lánc

Forrás jellemzése

$$\forall t - re \quad P\left(\xi_t = x_i \middle| \xi_1 = 1,\ldots,\xi_{t-1} = x_j\right) = P\left(\xi_t = x_i \middle| \xi_{t-1} = x_j\right) = p_{i|j}$$

Az együttes eloszlás :

$$P\left(\xi_1 = x_{i_1},\ldots,\xi_N = x_{i_N}\right) =$$

$$P\left(\xi_1 = x_{i_1},\ldots,\xi_{N-1} = x_{i_{N-1}}\right) \cdot P\left(\xi_N = x_{i_N} \middle| \xi_1 = x_{i_1},\ldots,\xi_{N-1} = x_{i_{N-1}}\right) =$$

$$P\left(\xi_1 = x_{i_1},\ldots,\xi_{N-1} = x_{i_{N-1}}\right) P\left(\xi_N = x_{i_N} \middle| \xi_{N-1} = x_{i_{N-1}}\right) =$$

$$P\left(\xi_1 = x_{i_1}\right) P\left(\xi_2 = x_{i_2} \middle| \xi_1 = x_{i_1}\right) \ldots P\left(\xi_N = x_{i_N} \middle| \xi_{N-1} = x_{i_{N-1}}\right) = P\left(\xi_1 = x_{i_1}\right) \prod_{i,j} p_{i|j}^{\mu_{i,j}}$$

ahol $\mu_{i,j}$ az x_i, x_j átmenetek száma, $P\left(\xi_t = x_i \middle| \xi_{t-1} = x_j\right)$

Forrás jellemzése

Markov lánc entrópiája határérték lesz

Szükséges fogalom az *ergodikus (határ) eloszlás*.

(Ergodikus Markov-lánc: minden állapotból minden állapot elérhető (véges lépéssel , pozitív valószínűséggel) és nem periodikus,

Minden érték visszatérő legyen.

Mi a viszony?

$$\overbrace{H(\xi_1,\ldots,\xi_N)}=\underbrace{H(\xi_N|\xi_1,\ldots,\xi_{N-1})+\overbrace{H(\xi_1,\ldots,\xi_{N-1})}}_{H(\xi,\eta)=H(\xi)+H(\eta)\wedge H(\xi)=H(\xi|\eta)}=$$

$$H(\xi_N|\xi_{N-1})+H(\xi_1,\ldots,\xi_{N-1})=$$

$$H(\xi_1)+H(\xi_2|\xi_1)+\ldots+H(\xi_N|\xi_{N-1})$$

$$H(\xi_t|\xi_{t-1})=\sum_{i=1}^{n}P(\xi_{t-1}=x_i)H(\xi_t|\xi_{t-1}=x_i)$$

$$H(\xi_t|\xi_{t-1}=x_i)=\sum_{j=1}^{n}P(\xi_t=x_j|\xi_{t-1}=x_i)\cdot log_2\,P(\xi_t=x_j|\xi_{t-1}=x_i)=$$

$$-\sum_{j=1}^{n}p_{j|i}\,log_2\,p_{j|i}=H(\xi|x_i)$$

$$H(\xi_t|\xi_{t-1})=\sum_{i=1}^{n}P(\xi_{t-1}=x_i)\cdot H(\xi|x_i)$$

Forrás jellemzése

A Markov lánc tulajdonság miatt :

$$H\left(\xi_N \middle| \xi_1,\ldots,\xi_{N-1}\right) = \sum_{i_1,i_2,\ldots,i_{N-1}} P\left(\xi_1 = x_{i_1},\ldots,\xi_{N-1} = x_{i_{N-1}}\right) \cdot$$

$$H\left(\xi_N \middle| \xi_1 = x_{i_1},\ldots,\xi_{N-1} = x_{i_{N-1}}\right) = \sum_{i_1,i_2,\ldots,i_{N-1}} P\left(\xi_1 = x_{i_1},\ldots,\xi_{N-1} = x_{i_{N-1}}\right) \cdot$$

$$\left(-\sum_{i=1}^{n} P\left(\xi_N = x_i \middle| \xi_1 = x_{i_1},\ldots,\xi_{N-1} = x_{i_{N-1}}\right) \cdot log_2\, P\left(\left(\xi_N = x_i \middle| \xi_1 = x_{i_1},\ldots,\xi_{N-1} = x_{i_{N-1}}\right)\right)\right)$$

Markov lánc és ξ_{N-1} eloszlása az együttes eloszlásból:

$$\sum_{j=1}^{n} P\left(\xi_{N-1} = x_j\right)\left(-\sum_{i=1}^{n} p_{i|j}\, log_2\, p_{i|j} \right) =$$

$$\sum_{j=1}^{n} P\left(\xi_{N-1} = x_j\right) H\left(\xi \middle| x_j\right)$$

Forrás jellemzése

Visszahelyettesítve $H(\xi_1,\ldots,\xi_N)$-be,

és alkalmasan visszafelé, $N=2$-ig, kapjuk :

$$H(\xi_1,\ldots,\xi_{N-1},\xi_N)=\sum_{t=2}^{N}\sum_{j=1}^{n}P(\xi_{t-1}=x_i)H(\xi|x_i)=$$

$$\sum_{j=1}^{n}H(\xi|x_i)\left(\sum_{t=1}^{N}P(\xi_{t-1}=x_j)\right)$$

Kihasználva, hogy ergodikus Markov lánc

$$\exists \quad \lim_{t\to\infty}P(\xi_t=x_i)=\pi_i \text{ határeloszlás, } \forall\, i=1,\ldots,n$$

Forrás jellemzése

π_i az átmenet valsz. mátrix sajátvektora $(\lambda = 1$ sajátértékkel$)$:

$$P(\xi_t = x_i) = \sum_{j=1}^{n} P(\xi_{t-1} = x_j) p_{i|j}$$

Mindkét oldal határértékét véve:

$$\pi_i = \sum_{j=1}^{n} \pi_j p_{i|j} \quad \forall i$$

Ez egy lineáris egyenletrendszer

$\pi_1, \ldots, \pi_n$, az átmenet valsz. mátrix sajátvektorai,

$P(\xi_t = x_i) = \pi_i \cdots$ettőt kezdve π_i nem függ az időtől

azaz stacionáris.

Forrás jellemzése

visszatérve az entrópiára

$$H(\xi_1,\ldots,\xi_N) = H(\xi_1) + \sum_{j=1}^{n} H(\xi|x_j)\left(\sum_{t=1}^{N} P(\xi_{t-1} = x_j)\right)$$

$$\lim_{N\to\infty} \frac{1}{N} \sum_{j=1}^{n} P(\xi_{t-1} = x_j) = \pi_j$$

ezért

$$H(\xi) = \lim_{N\to\infty} \frac{1}{N} H(\xi_1,\ldots,\xi_N) = \sum_{j=1}^{n} \pi_j H(\xi|x_j)$$

Forrás jellemzése

Lehet 3-ad, 4-ed stb rendű közelítés, ami két, három stb. lépéses Markov lánc.

Nyelvek esetén szokás/lehetőség, szavakat tekinteni a forrás választásának. Lehet a szavak felett 0-ad, 1-ső, 2-od, stb. közelítés.

(2-od rendű alapján szimulálva értelmes mondattöredékek nagy gyakorisággal keletkeznek.)

Matematikai kitérő – A valószínűségszámítás alapfogalmairól