Információs rendszerek elméleti alapjai

Információelmélet

Hosszú sorozatok jellemzése

Független, azonos eloszlású, változók esetére bizonyítás (Első rendű közelítés)

$\vec{\xi}, \xi_1, \ldots, \xi_N$ Független azonos eloszlású sorozat

$P(\xi = x_i) = p_i$ közös eloszlás

$\vec{x} = x_{i_1}, x_{i_2}, \ldots, x_{i_N}$, megfigyelés (kísérlet) olyan, hogy az x_i elf. száma $\mu_i \sim Np_i$ legyen

$$\text{Ekkor } P\left(\vec{\xi} = \vec{x}\right) \cong \prod_{i=1}^{n} p_i^{Np_i} = 2^{-N\left(-\sum_{i=1}^{n} p_i \log_2 p_i\right)} = 2^{-NH}$$

Hosszú sorozatok jellemzése

(A feltevés a nagy számok törvényének megfelelő tipikus gyakoriságnak tekinthető).

Igen hosszú (*N* elég nagy) sorozatokra, az x_i gyakorisága μ_i , tipikus esetben az alábbi két érték közé esik:

$$(1-\delta)Np_i < \mu_i < (1+\delta)Np_i$$

Ebből az ilyen feltételt kielégítő sorozatok valószínűsége alulról és felülről becsülhető

$$P\left(\vec{\xi} = \vec{x}\right) \cong 2^{\left(\sum_{i=1}^{n} \mu_i \log_2 p_i\right)}$$

Hosszú sorozatok jellemzése
Tétel

4. Tétel: $\forall\ \varepsilon > 0$, $\rho > 0$-hoz $\exists N_0$, $N > N_0$ a N hosszú sorozatok két halmazra oszthatók:

1. A lényegtelen halmaz. Ebbe legfeljebb ε valószínűséggel esik a választás.

2. A tipikus halmaz. Ebbe legalább 1- ε valószínűséggel esik a választás.

$\forall\ \vec{x}$ tipikus sorozatra

$$\left| \underbrace{\frac{log_2 P^{-1}\left(\vec{\xi} = \vec{x}\right)}{N}}_{\approx H} - H \right| < \rho$$

Hosszú sorozatok jellemzése
Becslés

Kifejezve:

$P\left(\vec{\xi} = \vec{x}\right)$ becslése

$$2^{-N(H+\rho)} < P\left(\vec{\xi} = \vec{x}\right) < 2^{-N(H-\rho)}$$

Bizonyítás: Csebisev egyenlőtlenségen alapul

A μ_i , $i=1,\ldots,n$ gyakoriságokat kell becsülni.

A μ_i , (N, p_i) paraméterű binomiális eloszlású változó.

Hosszú sorozatok jellemzése
Csebisev egyenlőtlenség

$$E(\mu_i) = Np_i$$

$$D^2(\mu_i) = N(1-p_i)p_i$$

$$D(\mu_i) = \sqrt{N(1-p_i)p_i}$$

(μ_i , $i=1,\ldots,n$ felírható N db független 0-1 értékű valószínűségi változó összegeként, - Bernoulli kisérlet)

Csebisev egyenlőtlenség

$$\forall \gamma > 0, P\left((\eta - E(\eta))^2 \geq \gamma^2\right) \leq \frac{D^2(\eta)}{\gamma^2}$$

Hosszú sorozatok jellemzése
Csebisev egyenlőtlenség alkalmazása

Csebisev egyenlőtlenséget alkalmazva

$$P\left(\left(\mu_i - Np_i\right)^2 \geq \gamma^2\right) \leq \frac{N\left(1-p_i\right)p_i}{\gamma^2}$$

Legyen $\gamma = \rho' N$,

$$P\left(\left(\mu_i - Np_i\right)^2 \geq \left(\rho' N\right)^2\right) \leq \frac{N\left(1-p_i\right)p_i}{N^2\ \rho'^2} = N^{-1}K < \varepsilon'$$

Ha N elég nagy ε' viszonyítva. Tehát

(*) $P\left(-\rho' N < \left(\mu_i - Np_i\right) < \rho' N\right) \geq 1-\varepsilon, \forall i - re,$ Ha N elég nagy

Hosszú sorozatok jellemzése

Ebből a lényegtelen halmaz: legalább egy i - re

$\left|\mu_i - Np_i\right| > \rho' N$ teljesül, és

$$P\left(\exists i : \left|\mu_i - Np_i\right| > \rho' N\right) < \sum_{i=1}^{n} P\left(\left|\mu_i - Np_i\right| > \rho' N\right) \leq n\varepsilon'$$

Tehát $\varepsilon' = \frac{\varepsilon}{n}$ választással kaptuk lényegtelen halmazt.

Hosszú sorozatok jellemzése

A lényeges halmaz ennek a komplementere

tehát minden $\vec{x}$ elemére $\vec{\xi} = \vec{x}$

a bekövetkezéséhez tartozó μ_i értékekre:

$N(p_i - \rho') < \mu_i < N(p_i + \rho'), i = 1, \ldots, n$ teljesül

Ebből $P(\vec{\xi} = \vec{x}) = 2^{\sum_{i=1}^{n} \mu_i \log_2 p_i}$ megbecsülhető mindkét oldalról

$$2^{N\left(\sum_{i=1}^{n} p_i \log_2 p_i - \rho' \sum_{i=1}^{n} \log_2 p_i\right)} < P(\vec{\xi} = \vec{x}) < 2^{N\left(\sum_{i=1}^{n} p_i \log_2 p_i + \rho' \sum_{i=1}^{n} \log_2 p_i\right)}$$

Hosszú sorozatok jellemzése

Reciprokot véve (egyenlőtlenség megfordul), majd a logaritmusát véve és N - el osztva

$$\left(2^{N\left(\sum_{i=1}^{n} p_i \log_2 p_i - \rho' \sum_{i=1}^{n} \log_2 p_i\right)}\right)^{-1} > \left(P\left(\vec{\xi} = \vec{x}\right)\right)^{-1} > \left(2^{N\left(\sum_{i=1}^{n} p_i \log_2 p_i + \rho' \sum_{i=1}^{n} \log_2 p_i\right)}\right)^{-1}$$

$$H + \rho' \sum_{i=1}^{n} \log_2 p_i > \frac{\log_2 P^{-1}\left(\vec{\xi} = \vec{x}\right)}{N} > H - \rho' \sum_{i=1}^{n} \log_2 p_i$$

$\rho = -\rho' \sum_{i=1}^{n} \log_2 p_i$ választással kapjuk a 4. Tétel bizonyitását

Hosszú sorozatok jellemzése

Következmény : $N > N_0$ esetén a

tipikus halmaz M_N elemszámára alsó és felső becslést kapunk a

(valószínűségek reciproka segítségével)

$$(1-\varepsilon)2^{N(H-\rho)} < M_N < (1-\varepsilon)2^{N(H+\rho)} < 2^{N(H+\rho)}$$

Nagyon hosszú jelsorozatok esetén a tipikus halmazon

majdnem egyenletessé teszi a választást

Hosszú sorozatok jellemzése

Legvalószí nűbb választáso k :

Különböző valószínűs égű N hosszú sorozatok

$p_1 \geq p_2 \geq \ldots \geq p_k$ legvalószí n. választás :

N db x_i, $\left(\left|\vec{x}_i\right| = N\right)$

De ez nem lesz tipikus (nincs benne a tipikus halmazban) .

$\vec{x}_1, \vec{x}_2, \ldots$ esetén : $P\left(\vec{\xi} = \vec{x}_i\right) \leq P\left(\vec{\xi} = \vec{x}_{i-1}\right)$

$\sum_{i=1}^{L} P\left(\vec{\xi} = \vec{x}_i\right) \geq 1 - \lambda$ - lényeges sorozat

$\sum_{i=1}^{L-1} P\left(\vec{\xi} = \vec{x}_i\right) < 1 - \lambda$

$L : \lambda -$ tól függő

$2^{N(H-\delta)} < \underbrace{L(\lambda)}_{\substack{\text{tipikus halmaz elemszámán ak becslése} \\ \text{igaz rá}}} < 2^{N(H+\delta)}$

$L(\lambda)$ és M_N nem ugyanaz,

de kevés elem esik a metszetükö n kivül.

Matematikai kitérő – A valószínűségszámítás alapfogalmairól

Binomiális (Bernoulli) eloszlás

A megfigyelésünk kétféle eredményt adhat: egy *A* esemény bekövetkezik vagy nem következik be.

n számú (egymástól független) megfigyelést végzünk.

Az *A* esemény valószínűsége minden egyes kísérletnél: *p*.

A valószínűségi változó: *n* megfigyelésből az *A* esemény bekövetkezésének száma (gyakorisága, *k*). Ez a visszatevéses mintavétel tipikus esete, ahol annak a valószínűsége, hogy az egymás után kiválasztott *n* elemből *k* db. a selejtes:

$$P(\xi = k) = p_k = \binom{n}{k} \cdot \frac{S^k (N-S)^{(n-k)}}{N^n}$$

A jelöléssel ez megfelel az alábbiaknak: $\frac{S}{N} = p, \quad 1 - p = q$

$$P(\xi = k) = p_k = \binom{n}{k} \cdot p^k (1-p)^{n-k}$$

Annak a valószínűsége, hogy az *A* esemény *n* számú megfigyelésből *k*-szor következik be:
A binomiális eloszlásban az ***n*** és a ***p*** un. paraméterek
A binomiális eloszlás várható értéke: $E(\xi) = np$.
A binomiális eloszlás szórásnégyzete:
$D^2(\xi) = npq$.
A binomiális eloszlás szórása $D(\xi) = \sqrt{npq}$

Matematikai kitérő – A valószínűségszámítás alapfogalmairól

Binomiális (Bernoulli) eloszlás

A Bernoulli eloszlás úgy is felfogható mint n független , 0,1 értékű, azonos eloszlású, valószínűségi változók összege.

A várható, érték, szórásnégyzet független valószínűségi változók esetén fennálló tulajdonsága miatt adódnak a tulajdonságai:

A binomiális eloszlásban az ***n*** és a ***p*** un. paraméterek

A binomiális eloszlás várható értéke: $E(\xi)= np$**.**

A binomiális eloszlás szórásnégyzete: $D^2(\xi)=npq$.

A binomiális eloszlás szórása $D(\xi)=\sqrt{npq}$