

# A sztochasztika alapjai

## MBNXK262

8. előadás: Huffman-kód, Véletlen vektorváltozók,  
függetlenség, kovariancia, korreláció

Kevei Péter

2024/25 tavasz

# Huffman-kód

$\mathcal{X} = \{x_1, x_2, \dots, x_n\}$  véges halmaz, a *forrásábécé*.

Kód  $f : \mathcal{X} \rightarrow \{\text{véges 0-1 sorozatok}\}$

Az  $f$ -hez tartozó lehetséges kódszavak  $f(x_1), f(x_2), \dots, f(x_n)$ .

# Huffman-kód

$\mathcal{X} = \{x_1, x_2, \dots, x_n\}$  véges halmaz, a *forrásábécé*.

Kód  $f : \mathcal{X} \rightarrow \{\text{véges 0-1 sorozatok}\}$

Az  $f$ -hez tartozó lehetséges kódszavak  $f(x_1), f(x_2), \dots, f(x_n)$ .

Az  $f$  kód *prefix*, ha a lehetséges kódszavak közül egyik sem folytatása a másiknak. Jelölje  $x \in \mathcal{X}$  esetén  $|f(x)|$  a kódszó hosszát.

## Példa

Legyen  $\mathcal{X} = \{a, b, c\}$ , és legyen  $f_1(a) = 0$ ,  $f_1(b) = 01$ ,  $f_1(c) = 011$ . Ekkor  $f_1$  nem prefix kód, de könnyen látható, hogy egyértelműen dekódolható. Az  $f_2(a) = 01$ ,  $f_2(b) = 00$ ,  $f_2(c) = 1$ , kód prefix.

Legyen  $X$  egy véletlen betű, és eloszlása  $\mathbf{P}(X = x_k) = p_k$ ,  $k = 1, 2, \dots, n$ . Tehát  $p_k$  a  $x_k$  betű gyakorisága az adott nyelvben.

Adott  $f$  kód esetén egy hosszú szövegben az egy karakterre eső átlagos kódszóhossz:

Feltehető, hogy  $p_1 \geq p_2 \geq \dots \geq p_n$ . Ha az  $f$  prefix kód optimális, akkor feltehető, hogy teljesülnek a következők:

(i) Hosszabb kódhoz ritkább betűk tartoznak, azaz

$$|f(x_1)| \leq |f(x_2)| \leq \dots \leq |f(x_n)|.$$

Feltehető, hogy  $p_1 \geq p_2 \geq \dots \geq p_n$ . Ha az  $f$  prefix kód optimális, akkor feltehető, hogy teljesülnek a következők:

(i) Hosszabb kódhoz ritkább betűk tartoznak, azaz

$$|f(x_1)| \leq |f(x_2)| \leq \dots \leq |f(x_n)|.$$

(ii) A két legkisebb valószínűséghez tartozó kód hossza egyenlő.

Feltehető, hogy  $p_1 \geq p_2 \geq \dots \geq p_n$ . Ha az  $f$  prefix kód optimális, akkor feltehető, hogy teljesülnek a következők:

(i) Hosszabb kódhoz ritkább betűk tartoznak, azaz

$$|f(x_1)| \leq |f(x_2)| \leq \dots \leq |f(x_n)|.$$

(ii) A két legkisebb valószínűséghez tartozó kód hossza egyenlő.

(iii)  $f(x_{n-1})$  és  $f(x_n)$  csak az utolsó bitben térnek el.

## Tétel

*Tegyük fel, hogy az*

$$\mathcal{X}' = \{x_1, \dots, x_{n-2}, y_{n-1}\}$$

*( $n - 1$ ) elemű forrásábécé és  $p_1, \dots, p_{n-2}, p_{n-1} + p_n$  eloszlás esetén  $g$  egy optimális prefix kód. Ekkor az eredeti problémához tartozó optimális prefix kódot kapunk, ha az  $x_{n-1}$ , ill.  $x_n$  kódját úgy választjuk, hogy a  $g(y_{n-1})$  kódszót kiegészítjük 0-val, ill. 1-gyel, a többi kódszót változatlanul hagyjuk.*

## Példa

Legyen  $n = 6$ ,  $\mathcal{X} = \{x_1, \dots, x_6\}$ , és  $p_1 = 0.132$ ,  $p_2 = 0.329$ ,  $p_3 = 0.329$ ,  $p_4 = 0.165$ ,  $p_5 = 0.041$ ,  $p_6 = 0.004$ .

(i)  $x_5, x_6 \rightarrow x_{56}$ ,  $p_{56} = 0.045$ ;

## Példa

Legyen  $n = 6$ ,  $\mathcal{X} = \{x_1, \dots, x_6\}$ , és  $p_1 = 0.132$ ,  $p_2 = 0.329$ ,  $p_3 = 0.329$ ,  $p_4 = 0.165$ ,  $p_5 = 0.041$ ,  $p_6 = 0.004$ .

(i)  $x_5, x_6 \rightarrow x_{56}$ ,  $p_{56} = 0.045$ ;

(ii)  $x_1, x_{56}$ ,  $p_{156} = 0.177$ ;

## Példa

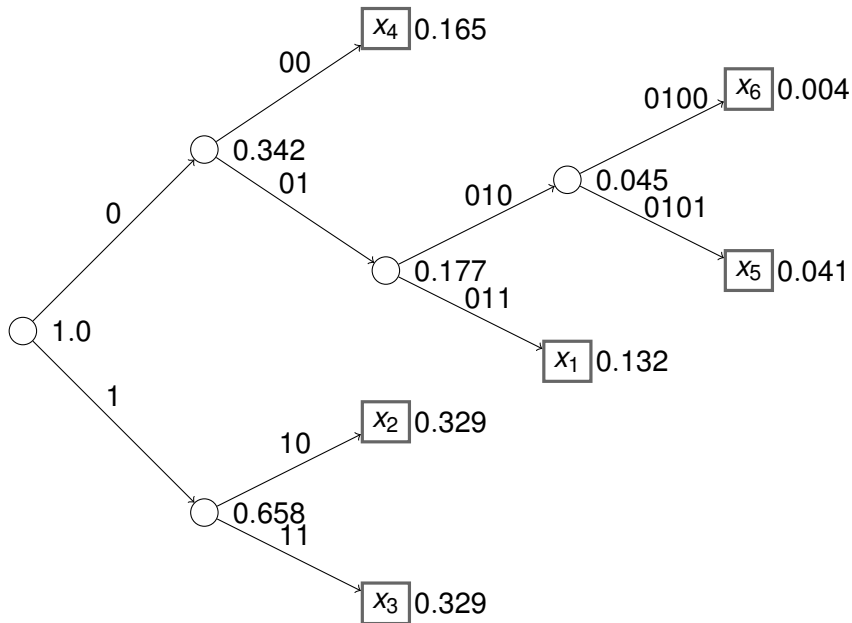
Legyen  $n = 6$ ,  $\mathcal{X} = \{x_1, \dots, x_6\}$ , és  $p_1 = 0.132$ ,  $p_2 = 0.329$ ,  $p_3 = 0.329$ ,  $p_4 = 0.165$ ,  $p_5 = 0.041$ ,  $p_6 = 0.004$ .

- (i)  $x_5, x_6 \rightarrow x_{56}$ ,  $p_{56} = 0.045$ ;
- (ii)  $x_1, x_{56}$ ,  $p_{156} = 0.177$ ;
- (iii)  $x_{156}, x_4$ ,  $p_{1564} = 0.342$ ;

## Példa

Legyen  $n = 6$ ,  $\mathcal{X} = \{x_1, \dots, x_6\}$ , és  $p_1 = 0.132$ ,  $p_2 = 0.329$ ,  $p_3 = 0.329$ ,  $p_4 = 0.165$ ,  $p_5 = 0.041$ ,  $p_6 = 0.004$ .

- (i)  $x_5, x_6 \rightarrow x_{56}$ ,  $p_{56} = 0.045$ ;
- (ii)  $x_1, x_{56}$ ,  $p_{156} = 0.177$ ;
- (iii)  $x_{156}, x_4$ ,  $p_{1564} = 0.342$ ;
- (iv)  $x_2, x_3$ ,  $p_{23} = 0.658$ .



Így az optimális kód:

| $x_1$ | $x_2$ | $x_3$ | $x_4$ | $x_5$ | $x_6$ |
|-------|-------|-------|-------|-------|-------|
| 011   | 10    | 11    | 00    | 0101  | 0100  |
| 0.132 | 0.329 | 0.329 | 0.165 | 0.041 | 0.004 |

A várható érték:

$$\begin{aligned} \mathbf{E}(f(X)) &= 0.132 \cdot 3 + 0.329 \cdot 2 + 0.329 \cdot 2 \\ &\quad + 0.165 \cdot 2 + 0.041 \cdot 4 + 0.004 \cdot 4 = 2.22 \end{aligned}$$

# Entrópia

Az optimális várható kódhosszra teljesül, hogy

$$\sum_{k=1}^n p_k \log_2 \frac{1}{p_k} \leq \mathbf{E}(|f(\xi)|) < \sum_{k=1}^n p_k \log_2 \frac{1}{p_k} + 1.$$

Pl.: JPEG, MP3.

# Véletlen vektorváltozók

## Definíció

$\xi = (\xi_1, \dots, \xi_n) : \Omega \rightarrow \mathbb{R}^n$  véletlen vektorváltozó, ha minden komponense véletlen változó. Eloszlásfüggvénye

$$F(x_1, \dots, x_n) = \mathbf{P}(\xi_1 < x_1, \dots, \xi_n < x_n).$$

$(\xi_1, \dots, \xi_n)$  véletlen vektorváltozó diszkrét, ha értékkészlete megszámlálható.

$\xi_i, i = 1, 2, \dots, n$ , eloszlása a peremeloszlás, vagy marginális eloszlás

## Trinomiális eloszlás

Szabályos dobókockával  $n$ -szer dobunk.  $\xi$  a hatosok,  $\eta$  egyesek száma

*peremeloszlások*: mind a hatosok, mind az egyesek száma *binomiális eloszlású* véletlen változó  $(n, 1/6)$  paraméterrel:

$$\mathbf{P}(\xi = k) = \mathbf{P}(\eta = k) = \binom{n}{k} \left(\frac{1}{6}\right)^k \left(\frac{5}{6}\right)^{n-k}, \quad k = 0, 1, \dots, n.$$

Így,

$$\mathbf{E}(\xi) = \frac{n}{6}, \quad \mathbf{D}^2(\xi) = n \frac{5}{36}.$$

# Trinomiális eloszlás

Szabályos dobókockával  $n$ -szer dobunk.  $\xi$  a hatosok,  $\eta$  egyesek száma

$(\xi, \eta)$  lehetséges értékei: olyan  $(k, \ell)$  párok, melyre  $0 \leq k, \ell \leq n$ , és  $k + \ell \leq n$ . A megfelelő valószínűségek

$$\mathbf{P}(\xi = k, \eta = \ell) =$$

# Függetlenség

## Definíció

$\xi_1, \dots, \xi_n$  függetlenek, ha minden  $x_1, \dots, x_n \in \mathbb{R}$  esetén

$$\mathbf{P}(\xi_1 < x_1, \dots, \xi_n < x_n) = \mathbf{P}(\xi_1 < x_1) \dots \mathbf{P}(\xi_n < x_n).$$

## Állítás

$\xi_1, \dots, \xi_n$  *diszkrét véletlen változók úgy, hogy  $\xi_i$  lehetséges értékei  $x_1^{(i)}, x_2^{(i)}, \dots, i = 1, 2, \dots, n$ . Ekkor  $\xi_1, \dots, \xi_n$  pontosan akkor függetlenek, ha*

$$\mathbf{P}(\xi_1 = x_{i_1}^{(1)}, \dots, \xi_n = x_{i_n}^{(n)}) = \mathbf{P}(\xi_1 = x_{i_1}^{(1)}) \dots \mathbf{P}(\xi_n = x_{i_n}^{(n)})$$

*tetszőleges  $i_1, \dots, i_n$  indexekre.*

# Kockadobás

## Példa

Egy szabályos dobókockával kétszer dobunk. Jelölje  $\xi$  az első,  $\eta$  a második dobás eredményét.

## Állítás

$\xi, \eta$  független diszkrét véletlen változók.

$$\mathbf{E}(g_1(\xi)g_2(\eta)) = \mathbf{E}(g_1(\xi)) \mathbf{E}(g_2(\eta)).$$

*Speciálisan, ha  $\xi$  és  $\eta$  függetlenek, akkor  $\mathbf{E}(\xi\eta) = \mathbf{E}(\xi)\mathbf{E}(\eta)$ .*

# Kovariancia, korreláció

## Definíció

Az  $\xi$  és  $\eta$  véletlen változók *kovarianciája*

$$\mathbf{Cov}(\xi, \eta) = \mathbf{E}[(\xi - \mathbf{E}(\xi))(\eta - \mathbf{E}(\eta))],$$

*korrelációja*

$$\rho(\xi, \eta) = \frac{\mathbf{Cov}(\xi, \eta)}{\mathbf{D}(\xi)\mathbf{D}(\eta)}.$$

# Tulajdonságok

(i)  $\mathbf{Cov}(\xi, \xi) = \mathbf{D}^2(\xi);$

(ii)  $\mathbf{Cov}(\xi, \eta) = \mathbf{Cov}(\eta, \xi);$

# Tulajdonságok

(iii)  $\mathbf{Cov}(\xi, \eta) = \mathbf{E}(\xi\eta) - \mathbf{E}(\xi)\mathbf{E}(\eta);$

# Tulajdonságok

(iv)  $\mathbf{Cov}(a(\xi + c), b(\eta + d)) = ab\mathbf{Cov}(\xi, \eta);$

# Tulajdonságok

$$(v) \mathbf{Cov} \left( \sum_{i=1}^n \xi_i, \sum_{j=1}^m \eta_j \right) = \sum_{i=1}^n \sum_{j=1}^m \mathbf{Cov}(\xi_i, \eta_j);$$

# Tulajdonságok

(vi) ha  $\xi$  és  $\eta$  függetlenek, akkor **Cov** $(\xi, \eta) = 0$ .

# Tulajdonságok

## Állítás

*Tetszőleges  $\xi, \xi_1, \dots, \xi_n, \eta, \eta_1, \dots, \eta_m$  véletlen változók és  $a, b$  valós számok esetén igazak az alábbiak.*

- (i)  $\mathbf{Cov}(\xi, \xi) = \mathbf{D}^2(\xi);$
- (ii)  $\mathbf{Cov}(\xi, \eta) = \mathbf{Cov}(\eta, \xi);$
- (iii)  $\mathbf{Cov}(\xi, \eta) = \mathbf{E}(\xi\eta) - \mathbf{E}(\xi)\mathbf{E}(\eta);$
- (iv)  $\mathbf{Cov}(a(\xi + c), b(\eta + d)) = ab\mathbf{Cov}(\xi, \eta);$
- (v)  $\mathbf{Cov}\left(\sum_{i=1}^n \xi_i, \sum_{j=1}^m \eta_j\right) = \sum_{i=1}^n \sum_{j=1}^m \mathbf{Cov}(\xi_i, \eta_j);$
- (vi) *ha  $\xi$  és  $\eta$  függetlenek, akkor  $\mathbf{Cov}(\xi, \eta) = 0$ .*

## Állítás

(i) *Bunyakovszkij–Cauchy–Schwarz-egyenlőtlenség:*

$$|\mathbf{Cov}(\xi, \eta)| \leq \mathbf{D}(\xi)\mathbf{D}(\eta).$$

*Innen adódik, hogy  $\rho(\xi, \eta) \in [-1, 1]$ ;*

(ii) *ha  $\rho(\xi, \eta) = 1$ , akkor*

$$\xi = \mathbf{E}(\xi) + \frac{\mathbf{D}(\xi)}{\mathbf{D}(\eta)}(\eta - \mathbf{E}(\eta));$$

(iii) *ha  $\rho(\xi, \eta) = -1$ , akkor*

$$\xi = \mathbf{E}(\xi) - \frac{\mathbf{D}(\xi)}{\mathbf{D}(\eta)}(\eta - \mathbf{E}(\eta)).$$

# Bizonyítás

## Összeg szórásnégyzete

Legyenek  $\xi, \eta$  véletlen változók,  $\rho$  a korrelációjuk. Ekkor

$$\mathbf{D}^2(\xi + \eta) =$$

# Összeg szórásnégyzete

Ha  $\xi$  és  $\eta$  függetlenek, akkor  $\rho = 0$ , így

$$\mathbf{D}^2(\xi + \eta) = \mathbf{D}^2(\xi) + \mathbf{D}^2(\eta).$$

Indukcióval, ha  $\xi_1, \xi_2, \dots, \xi_n$  páronként független (korrelálatlan) véletlen változók, akkor

$$\mathbf{D}^2\left(\sum_{i=1}^n \xi_i\right) = \sum_{i=1}^n \mathbf{D}^2(\xi_i).$$

# Kovariancia

## Példa

Egy szabályos dobókockával  $n$ -szer dobunk. Jelölje  $\xi$  a hatosok,  $\eta$  egyesek számát! Adjuk meg a várható értéket, szórást, kovarianciát, korrelációt!

# Lineáris regresszió felé

## Példa

3, külsőre egyforma érmével a fejdobás valószínűsége  $1/4, 2/4, 3/4$ . Véletlenszerűen választunk egy érmét, és azzal kétszer dobunk. Legyen  $\eta$  a választott érmével dobva a fej valószínűsége,  $\xi$  a dobott fejek száma.  $\xi$ -ből szeretnénk  $\eta$  értékére következtetni.

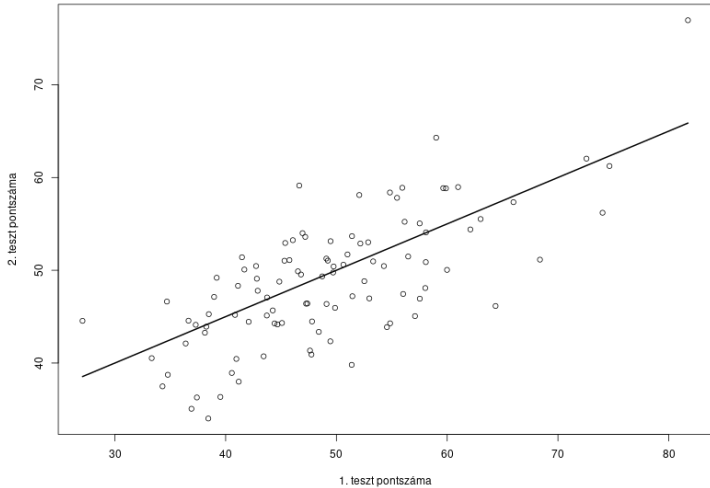
# Lineáris regresszió felé

$(\xi, \eta)$  együttes eloszlása:

| $\eta \backslash \xi$ | 0               | 1               | 2               | $\Sigma$      |
|-----------------------|-----------------|-----------------|-----------------|---------------|
| $\frac{1}{4}$         | $\frac{9}{48}$  | $\frac{6}{48}$  | $\frac{1}{48}$  | $\frac{1}{3}$ |
| $\frac{1}{2}$         | $\frac{1}{12}$  | $\frac{1}{6}$   | $\frac{1}{12}$  | $\frac{1}{3}$ |
| $\frac{3}{4}$         | $\frac{1}{48}$  | $\frac{6}{48}$  | $\frac{9}{48}$  | $\frac{1}{3}$ |
| $\Sigma$              | $\frac{14}{48}$ | $\frac{20}{48}$ | $\frac{14}{48}$ | 1             |

# Lineáris regresszió felé

A sztochasztika alapjai kurzus elején a hallgatók kitöltenek egy tesztet, mely az eddigi matematika tudásukat méri. A kurzus teljesítéséhez a félév végén is kitöltenek egy tesztet a kurzus anyagából. A korábbi évek tapasztalatai alapján feltehető, hogy az  $i$ -edik hallgató első teszten elért pontszáma  $\xi_i$ , a másodikon  $\eta_i = \frac{\xi_i + \xi'_i}{2}$ , ahol  $\xi_1, \xi_2, \dots$  és  $\xi'_1, \xi'_2, \dots$  független normális eloszlású változók  $\mu = 50$  várható értékkel és  $\sigma = 10$  szórással,  $i = 1, 2, \dots, N$ , ahol  $N$  a hallgatók száma.



ábra: 100 hallgató pontszámának alakulása

Számítsuk ki a két változó,  $\xi$  és  $\eta$  korrelációs együtthatóját!  
( $\eta = \frac{\xi + \xi'}{2}$ ,  $\xi, \xi'$  független  $N(50, 100)$ )

Ha ismerjük  $\xi$  értékét, akkor

$$\mathbf{E}[\eta|\xi] = \mathbf{E}\left[\frac{\xi + \xi'}{2}|\xi\right] = 25 + \frac{\xi}{2},$$

hiszen  $\xi$  értékét pontosan tudom,  $\xi'$  várható értéke pedig 50, és  $\xi$  független  $\xi'$ -től. Ez éppen a regressziós egyenes.

# Lineáris regresszió

$(\xi, \eta)$  véletlen vektorváltozó. Az  $\eta$  változót tekintem *függő* változónak, ennek az értékére szeretnék következtetni a  $\xi$  *független* változó értékéből. Vagyis ismert  $\xi$  esetén szeretném megmondani  $\eta$ -t. Keressük azokat az  $a, b$  valós számokat, melyre a  $\eta - (a\xi + b)$  változó kicsi. A kicsiséget négyzetes hibában mérve, keressük az

$$h(a, b) = \mathbf{E} \left[ (\eta - (a\xi + b))^2 \right]$$

függvény minimumhelyét, azaz a legjobb  $a, b$  választást.

# Lineáris regresszió

# Lineáris regresszió

# Lineáris regresszió

Ezek szerint a legjobb lineáris közelítést a

$$g(x) = \frac{\mathbf{Cov}(\eta, \xi)}{\mathbf{D}^2(\xi)}(x - \mathbf{E}(\xi)) + \mathbf{E}(\eta)$$

függvény adja. Ő a *regressziós egyenes*.

Ha  $\xi$  és  $\eta$  korrelálatlanok, azaz  $\mathbf{Cov}(\xi, \eta) = 0$ , akkor a legjobb közelítés  $\mathbf{E}(\eta)$ , vagyis  $\xi$  semmi információt nem ad  $\eta$  értékéről.

# Teszt

A Huffman-kódban a két legkisebb valószínűségű betű kódja

- ▶ pontosan az első bitben tér el
- ▶ pontosan az utolsó bitben tér el
- ▶ nem egyforma hosszú
- ▶ megegyezik

# Teszt

A  $\xi$  véletlen változó várható értéke 1, szórásnégyzete 5. Ekkor  $D^2(2\xi + 1) =$

- ▶ 11
- ▶ 21
- ▶ 20
- ▶ más

# Teszt

A 3 egyforma cinkelt érmés feladatban, mennyi a valószínűsége, hogy pontosan 1 fejet dobok a szabályos érmével?

- ▶  $1/6$
- ▶  $1/2$
- ▶  $1/12$
- ▶ más

# Teszt

Legyen  $\text{Cov}(\xi, \eta) = 3$ . Ekkor  $\text{Cov}(\xi + 2, \eta + 1) =$

- ▶ 3
- ▶ 4
- ▶ 5
- ▶ 6