

MATEMATIKA ÉS STATISZTIKA DATA SCIENCE-HEZ



VALÓSZÍNŰÉGSZÁMÍTÁS I. VALÓSZÍNŰSÉG FOGALMA, FELTÉTELES VALÓSZÍNŰSÉG

VALÓSZÍNŰSÉGSZÁMÍTÁS ALAPFOGALMAI



JELENSÉGEK CSOPORTOSÍTÁSA

véletlen jellegük alapján

determinisztikus jelenségek

adott körülmények mellett minden esetben ugyanaz az esemény következik be — pl.: 100 celsius fokon a víz forr

sztochasztikus (véletlen) jelenségek

adott körülmények nem mindig azonos esemény következik be — pl.: ha egy pénzérmét feldobunk, akkor az hol a fej, hol az írás oldalával felfelé esik le

véletlen tömegjelenségek

amiknek közös jellemzője, hogy változatlan körülmények között akár végtelen sokszor megismételhetők — pl.: kockadobás, érmedobás, lottóhúzás, statisztikai mintavétel, stb.

egyszeri véletlen jelenségek

amiknek közös jellemzője, hogy az őket befolyásoló körülmények nem ismételhethők meg egy az egyben — pl.: sportesemények

VALÓSZÍNŰSÉGSZÁMÍTÁS ALAPFOGALMAI



VÉLETLEN (tömeg)JELENSÉGEK

vagy másnéven **kísérletek**hez tartozó fogalmak

elemi események – a kísérlet lehetséges kimenetelei (jele: ω_i)

eseménytér – az összes elemi eseményt tartalmazó halmaz (jele: Ω)

esemény – az eseménytér egy részhalmaza (jele: A, B, C, stb...)

lehetetlen esemény – 0 elemi eseményt tartalmazó halmaz (jele: \emptyset vagy $\{ \}$)

biztos esemény – az összes elemi eseményt tartalmazó halmaz, tehát maga az eseménytér (jele: Ω)

események halmaza – az összes lehetséges esemény tartalmazó halmaz (jele: \mathcal{A})

KOCKADOBÁS

szabályos hatoldalú dobókockával

$$\omega_1 = \{1\}, \omega_2 = \{2\}, \omega_3 = \{3\}, \omega_4 = \{4\}, \omega_5 = \{5\}, \omega_6 = \{6\}$$

$$\Omega = \{\omega_1; \omega_2; \omega_3; \omega_4; \omega_5; \omega_6\}, \quad \Omega = \{1; 2; 3; 4; 5; 6\}$$

$$A: \text{páros dobás}, \quad A = \{\omega_2; \omega_4; \omega_6\} = \{2; 4; 6\}$$

$$B: \text{prímszám dobás}, \quad B = \{\omega_2; \omega_3; \omega_5\} = \{2; 3; 5\}$$

$$C: 7\text{-es dobás}, \quad C = \emptyset \quad \text{vagy} \quad C = \{ \}$$

$$D: \text{egyjegyű számot dobunk}, \quad D = \Omega = \{1; 2; 3; 4; 5; 6\}$$

$$\mathcal{A} = \{ \emptyset, \{1\}, \{2\}, \{3\}, \dots, \{1; 2\}, \{1; 3\}, \dots, \{1; 2; 3\}, \dots, \{1; 2; 3; 4\}, \dots, \{1; 2; 3; 4; 5\}, \dots, \Omega \}$$

VALÓSZÍNŰSÉG FOGALMA



A valószínűség egy olyan függvény,

amely minden eseményhez egy számot rendel a 0 és 1 közötti intervallumból;
bármely A esemény bekövetkezésének valószínűségét **$P(A)$** -val jelöljük.

A valószínűségszámítás Kolmogorov három axiómájára épül:

- ▶ $0 \leq P(A) \leq 1$,
- ▶ $P(\Omega) = 1$, és
- ▶ ha $A \cap B = \emptyset$, akkor $P(A \cup B) = P(A) + P(B)$



Eseményekre valószínűségére vonatkozó tulajdonságok

- ▶ $P(\emptyset) = 0$
- ▶ $P(\overline{A}) = 1 - P(A)$
- ▶ $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- ▶ ha $A \subseteq B$, akkor $P(A) \leq P(B)$

KLASSZIKUS VALÓSZÍNŰSÉGI MEZŐ



Klasszikus valószínűségi mezőről beszélünk, ha

- (1) az eseménytérben található elemi események száma véges (jelölje a számukat $|\Omega| = n \in \mathbb{N}$),
- (2) az elemi események valószínűsége pozitív és egyenlő.

Amennyiben ez a két feltétel teljesül, akkor **az elemi események valószínűsége:**

$$\mathbf{P}(\omega_1) = \dots = \mathbf{P}(\omega_n) = 1/n \quad , \text{ hiszen } 1 = \mathbf{P}(\Omega) = \mathbf{P}\left(\bigcup_{i=1}^n \omega_i\right) = \sum_{i=1}^n \mathbf{P}(\omega_i) = n\mathbf{P}(\omega_1)$$

Egy **A esemény valószínűségének meghatározásához** tegyük fel, hogy az $A \in \mathcal{A}$ esemény k darab elemi — azaz az esemény bekövetkezése szempontjából kedvező — eseményből áll:

$$A = \{\omega_1, \dots, \omega_k\} \quad , \text{ ekkor } \mathbf{P}(A) = \mathbf{P}(\omega_1 \cup \dots \cup \omega_k) = \mathbf{P}(\omega_1) + \dots + \mathbf{P}(\omega_k) = k \cdot \frac{1}{n} = \frac{k}{n}$$

azaz az A esemény valószínűségét egy klasszikus valószínűségi mezőben
a kedvező események és az összes esemény számának hányadosaként számítjuk

PÉLDA



	<div>KOCKADOBÁS</div> <div>szabályos hatoldalú dobókockával</div>	<div>ÉRMEDO BÁS</div> <div>szabályos kétoldalú pénzérmével</div>
<div>Klasszikus</div> <div>valószínűségi mező</div>	<div>✓ elemi események (lehetséges kimenetek) száma véges</div> <div>✓ elemi események valószínűsége pozitív és egyenlő</div>	<div>✓ elemi események (lehetséges kimenetek) száma véges</div> <div>✓ elemi események valószínűsége pozitív és egyenlő</div>
<div>Elemi események</div> <div>valószínűsége</div>	<div>$P(\omega_1) = P(\omega_2) = P(\omega_3) = P(\omega_4) = P(\omega_5) = P(\omega_6) = \frac{1}{6}$</div>	<div>$P(\omega_1) = P(\omega_2) = \frac{1}{2}$</div>
<div>Összetettebb</div> <div>esemény</div> <div>valószínűsége</div>	<div>B: prímszám dobás, $B = \{\omega_2; \omega_3; \omega_5\} = \{2; 3; 5\}$</div> <div>$P(B) = \frac{k}{n} = \frac{3}{6} = \frac{1}{2}$</div>	<div>$A = \{\text{egy érme kétszeri feldobása után pontosan egy fejet kapunk}\}$</div> <div>$\Omega = \{FF; FI; IF; II\} \qquad A = \{FI; IF\}$</div> <div>$P(A) = \frac{k}{n} = \frac{2}{4} = \frac{1}{2}$</div>

NÉHA KIFEJEZETTEN ÖSSZETETT FELADAT...



KOMBINATORIKA FELADATOK

Hányféleképpen?

I. SORBARENDEZÉSI PROBLÉMÁK

n elemből az összeset elemmel dolgozunk

A) EGYENES mentén

B) KÖR mentén

$$(n - 1)!$$

ISMÉTLÉS NÉLKÜL

n **különböző** elemet kell
sorbarendezni

$$n!$$

ISMÉTLÉSEL

n olyan elemet kell sorba-
rendezni, amelyek között
ismétlődő elemek is vannak

$$\frac{n!}{k_1! k_2! \dots k_r!}$$

II. KIVÁLASZTÁSI PROBLÉMÁK

n elemből csak k elemmel dolgozunk

A) KOMBINÁCIÓ

NEM SZÁMÍT A SORREND — vagyis ha ugyanazokat
az elemeket más sorrendben választjuk ki, az
ugyanannak a kiválasztásnak számít

ISMÉTLÉS NÉLKÜL

Egy elemet csak egyszer
választhatunk ki

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

ISMÉTLÉSEL

Egy elemet többször is
kiválaszthatunk

$$\binom{n+k-1}{k}$$

B) VARIÁCIÓ

SZÁMÍT A SORREND — vagyis ha ugyanazokat az
elemeket más sorrendben választjuk ki,
az más kiválasztásnak számít

ISMÉTLÉS NÉLKÜL

Egy elemet csak egyszer
választhatunk ki

$$\frac{n!}{(n-k)!}$$

ISMÉTLÉSEL

Egy elemet többször is
kiválaszthatunk

$$n^k$$

FELTÉTELES VALÓSZÍNŰSÉG

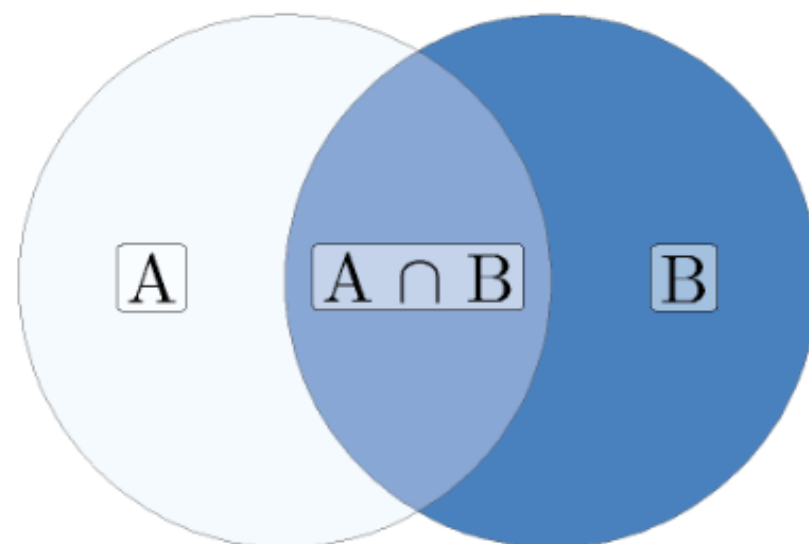


Gyakran arra vagyunk kíváncsiak, hogy egy esemény bekövetkezése befolyásolja-e, és ha igen, milyen mértékben egy másik esemény valószínűségét.

→ Tegyük fel, hogy egy A esemény $P(A)$ valószínűségét vizsgáljuk, majd tudomásunkra jut, hogy egy B esemény bekövetkezett ($P(B) \neq 0$). Ennek tükrében az A esemény valószínűsége megváltozhat.

→ Ilyenkor van szükségünk **a feltételes valószínűség** fogalmára.
Jelölése: $P(A|B)$

Például: ha egy társasjáték vége felé közeledve akkor tudok nyerni, ha egy hatoldalú kockával kétszer dobva több mint 10-et dobok (A), akkor az 1. dobás értéke (B) befolyásolja a nyerési esélyeimet.



$$P(A | B) = \frac{P(A \cap B)}{P(B)}$$

FELTÉTELES VALÓSZÍNŰSÉG



Például: ha egy társasjáték vége felé közeledve akkor tudok nyerni, ha egy hatoldalú kockával kétszer dobva több mint 10-et dobok (A), akkor az 1. dobás értéke (B) befolyásolja a nyerési esélyeimet.

$$A = \{ \text{a dobott számok összege} > 10 \}, \quad B = \{ \text{az első dobás 6-os} \} \quad P(A | B) = \frac{P(A \cap B)}{P(B)}$$

$$A \cap B = \{ \text{a dobott számok összege} > 10 \text{ ÉS az első dobás 6-os} \}$$

$$P(A) = \frac{3}{36} \approx 8 \%$$

$$P(A \cap B) = \frac{2}{36}$$

$$P(B) = \frac{1}{6}$$



$$P(A | B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{2}{36}}{\frac{1}{6}} = \frac{2}{6} = \frac{1}{3} \approx 33 \%$$

FELTÉTELES VALÓSZÍNŰSÉG A DS-BEN



Tulajdonképpen ez minden **ML modell predikció**jának az alapja a háttérben:

$$P (Y | X)$$

Mi a valószínűsége, hogy bekövetkezik valami (Y), **feltéve hogy** ismerjük a bemeneti jellemzőket (X)

DS project	Feltételes valószínűségi gondolkodás
Churn prediction	P(lemondás használati mintázat)
Fraud detection	P(csalás tranzakció jellemzői)
Recommendation system	P(érdeklődés profil, előzmények)
A/B testing	P(siker B verzió) VS P(siker A verzió)
Spam filter	P(spam email szavai, metaadatok)
Predictive maintenance	P(hiba a közel jövőben szenzoradatok, terhelés, környezet)

FELTÉTELES VALÓSZÍNŰSÉG A DS-BEN



Hogyan jelenik meg az ML modellek gondolkodásában?

1) Lineáris regresszió – mit jelent a gyakorlatban?

Cél: folytonos mennyiséget becsülünk (ár, idő, bevétel stb.) jellemzők alapján.

Gondolat:

„ha ismerem a bemeneti jellemzőket X , akkor *milyen értéket várok* a kimenetre Y ?”

Ezt így írjuk le:

$$E(Y \mid X) = \beta_0 + \beta_1 X_1 + \dots + \beta_k X_k$$

vagyis: a kimenet várható (átlagos) értéke a jellemzők *lineáris* kombinációja.

Miért „feltételes valószínűség”?

Mert nem azt mondjuk, hogy „pontosan ennyi lesz Y ”, hanem:

„**átlagosan ennyi lesz Y , feltéve hogy X ilyen.**”

Ha szeretnéd, ezt úgy is lehet nézni, hogy a modell a teljes eloszlást is feltételezi:

$$Y \mid X \sim \mathcal{N}(\beta^\top X, \sigma^2)$$

tehát $P(Y \mid X)$ egy normális eloszlás, amelynek **közepe** a fenti lineáris kifejezés, **szórása** pedig σ .

Mit csinál a tanítás (OLS)?

Olyan β -kat keres, amelyek mellett a becsült átlagtól való eltérések (a négyzetes hibák) a legkisebbek – ez empirikusan pont azt jelenti, hogy a **feltételes átlagot** jól közelítjük.

2) Logisztikus regresszió — mit csinál valójában?

A logisztikus regresszió akkor kell, amikor a kimenet (Y) **kategóriás**, tipikusan két értéket vehet fel, pl.:

- vásárolt / nem vásárolt
- leiratkozott / nem iratkozott le
- spam / nem spam

Mit próbál megtanulni?

Azt, hogy **mekkora a valószínűsége** annak, hogy valami megtörténik, **feltéve hogy** ismerjük az X bemeneti jellemzőket.

Vagyis:

$$P(Y = 1 \mid X)$$

— ez pontosan egy **feltételes valószínűség**.

Hogyan gondolkodik a modell?

Azt mondja:

ha az X jellemzők nőnek vagy csökkennek, a „pozitív esemény” (pl. vásárlás) esélye is változik — de **nem lineárisan**, hanem egy S-alakú görbe mentén.

Ezt az S-alakú függvényt hívjuk **logisztikus (szigmoid) függvénynek**:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_k X_k)}}$$

Mit tanul meg a modell?

A β **értékeket** úgy választja meg, hogy az adatok alapján a

$P(Y \mid X)$ feltételes valószínűség minél jobban illeszkedjen a valós megfigyelésekhez.

Ezt nevezzük **maximum likelihood becslésnek** —

olyan β -kat keresünk, amik a megfigyelt adatok bekövetkezését a legvalószínűbbé teszik.

FELTÉTELES VALÓSZÍNŰSÉG A DS-BEN



3) Döntési fa – hogyan használ feltételes valószínűséget?

A döntési fa sokkal „intuitívabb” modell, mint a regressziók.

Úgy működik, mint egy **sorozatos „ha... akkor...” döntésfa**, de a logikája mögött ugyanaz a gondolat: megpróbálja becsülni, **hogyan oszlik el a célváltozó Y a bemeneti jellemzők X ismeretében.**

Hogyan épül a fa?

A modell mindig azt kérdezi:

„Ha ezen a jellemzőnél kettévágom az adatokat (pl. életkor < 35?), akkor a csoportokban tisztábban elkülönülnek az Y értékek?”

Ha igen, akkor ez a vágás **javítja a $P(Y|X)$ becslését.**

Így épül a fa, lépésről lépésre, míg minden levélhez (végpont) tartozik egy becslés.



Klasszifikációs döntési fa

Ha az Y pl. „vásárolt / nem vásárolt”, akkor minden levélben megnézi, hogy az adott jellemzők kombinációja mellett **milyen gyakran** történt a vásárlás.

Ez a gyakoriság adja a becsült feltételes valószínűséget:

$$\hat{P}(Y = 1 \mid X \in \text{leaf})$$

Magyarul:

„Abban a csoportban, ahol X ilyen és ilyen, a vásárlás aránya 0.72 → tehát $P(Y = 1|X) = 0.72$.”



Regressziós döntési fa

Ha az Y folytonos (pl. bevétel), akkor minden levélben a modell kis mintaátlagokat számol:

$$\hat{E}(Y \mid X \in \text{leaf})$$

Tehát nem valószínűséget, hanem **feltételes várható értéket** becsül — ugyanazt a gondolatot, mint a lineáris regresszió, csak **nem lineárisan.**

GYAKORLÓ FELADATOK



Önállóan megoldandó feladatok az érintett témakörökből:

[3_alkalom_gyakorlo_feladatok.ipynb](#)