**Döntéstámogatás - 5. - október 5.**

# Tartalom

* Sampling
  + Mintavétel
  + Tanuló – teszt
* Exploring
* Modificating
* Modelling
* Assessment

## Minták

Populáció minta: P1 = és P2 = , Normális: Ɲ(γ, ϭ)

Normális eloszlás mintavételi tétele: ha a populáció normális, akkor a minta is az:

γS = γ, ϭS = , nS: minta számossága

Centrális határeloszlás tétel: n → ∞, a minta normális

## Adatsor szétválasztása

Tanuló (2/3) és teszt (1/3) részre általában. Teszt állomány lehetőleg legyen véletlenszerű.

Szétválasztási metódusok:

### Keresztvalidáció

Osszuk n (≈ 5, 10) részre, legyen a tanuló az 1...n-1., a teszt az n. rész, majd válasszuk a következőt a tesztállománynak 1…n-2,n, teszt: n-1. A hiba a hibák átlaga lesz.

### Leave-One-Out

1 adat lesz a teszt a többi a tanuló. Csak ha kicsi az adatsor!

### Bootstrap aggregating (Bagging)

a1...aN húzunk véletlenül visszatevéssel (vagyis egy adatot többször is kihúzhatunk).

N-szer húzunk, amiket 1-szer se húztunk ki, azokból fog állni a teszt állomány.

Méret becslés: az, hogy egy körben nem húzzuk ki az i.-et: pi = 1 - vagyis (1- )N = ≈ 27,8 % lesz a teszt. (ezért is hívják még „0,632 Bootstrap módszer”-nek).

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
|  | **V1** | **…** | **Vj** | **…** | **VN** |
| **a1** | 3 |  | 19 |  |  |
| **…** | 4 |  | 18 |  |  |
| **ai** | 452 |  | 21 |  |  |
| **…** | 7 |  | 185 |  |  |
| **aN** | 2 |  | 13 |  |  |

## Outlier adatok

Azok az adatok, amelyek „kilógnak”, valószínűsíthetőleg hiba miatt.

### Trimmerelés

Végigmegyünk a táblázaton az 1. oszloptól, oszloponként és soronként, először az 1. oszlopból töröljük azokat a sorokat, amelyeknek az értéke kilóg. Veszély: lehet, hogy az N. oszlopnál már alig marad sorunk!

### Windzorizálás

Kihúzás helyett csonkolunk: a kilógó érték helyére a megengedett max (vagy min) értéket írjuk.

## Hiányzó értékek

Az érték nem hibás, hanem egyáltalán nincs megadva. A hiányzó helyre az adott oszlopban szereplő értékek átlagát írhatjuk. Amennyiben az értékek ordinálisak, pl: nyelvvizsga (A,B,C), vagy színek:

* Gyakoriság alapján választunk. (Módusz)
* Medián: az egyforma értékeket csoportosítva, valamilyen sorrendbe felírva, a középső.
* Korrelált változóval való becslés: egy másik oszlop, amelyikkel korrelál, segíthet, pl. amikor ott alacsony az érték itt piros, mikor ott magas itt kék. Ez alapján ki tudunk találni egy átlagos értéket.

## Hisztogram

Az adatokat megjeleníthetjük hisztogramon is. (Grafikonon, ahol az x tengelyen az értékek, az y tengelyen az adott értékű adatok száma található). Fontos, hogy jó felbontást találjunk, túl kicsi/túl nagy nem jó, pl: életkornál 2-5 éves felbontás.