5. óra

Adatelemzési platformok, BME, 2018. Február 27., II. elméleti óra.

Felügyelt tanulás (ismétlés)

- Ugyanúgy mint a nem felügyeltnél, a táblázatos formából indulunk ki.
- Célváltozó: erre szeretnék következtetni

Egy olyan modelt szeretnék, ami a bemeneti változók ismeretében meg tudja becsülni a célváltozót.

Két nagy csoport

- Osztályozás: kategórikus célváltozóval (leginkább bináris)
- Regresszió: célváltozó numerikus

A regresszióhoz a lineáris regressziót fogjuk megnézni.

Üzleti probléma: ügyfélérték

Ha sok ügyfelem van, szeretném tudni, hogy mennyire profitábilisak számomra.

- Customer value: jelenlegi ügyfélérték, eddig mennyi értéket termelt számomra
- Customer lifetime value: nemcsak a jelenig, hanem a jövőbeli potenciált is tartalmazza, ezen belül
  - Value by product (adott termékre mennyi profitot termel)
    - e.g. pr1 * 5 * 0.05 * fee
  - Affinitás modell
    - Cross-sellinghez (keresztértékesítéshez) kapcsolódó termék (kóla a sült krumpli)
    - Upsell: magasabb értékű terméket próbálok eladni (nagyonbb kóla)
  - Időszak: mindenki más időtartamban van, nem lehet egyszerűen mindenkire ugyanúgy mérni
  - Lemorzsolódás

Az adott ügyfél mekkora arányban hajlandó az ajánlatot megfizetni.

- Egy numerikus változóval próbáljuk megbecsülni.
- De valószínűség jellegű változót nem regresszióval becsülünk, hanem osztályozással.

Lineáris regresszió

Célváltozó: Customer Lifetime Value (CLV)

1 változó: életkor

- Egy sor a táblázatban: 1 'pont' a függvényen
- A CLV és a kor közötti lineáris összefüggést próbáljuk becsülni

$$\bar{y} = \omega_o + \omega_{age} * age$$

- egy vektorhoz egy változót rendelünk
- ω_{age}: a meredeksége a függvénynek

Egy olyan egyenest próbálunk találni, ami esetében a ponthalmaztól vett négyzetes távolsága minimális legyen. A négyzetes távolság jobban bünteti a nagy hibákat. Az abszolút értékkel szemben a négyzetes érték folytonoan deriválható, ami megkönnyíti a minimalizálást.

Többváltozós lineáris regresszió

$$\bar{y} = \omega_o + \sum_i \omega_i * age$$

- Ez egy hipersíkot vázol fel [bár erről volt egy vita, hogy nem csak-e egy egyenest]
  - Hipersíkot leginkább lineáris szeparálás esetén használjuk, ahol egy osztályozási problémát oldunk meg.
  - Itt egy darab célváltozóhoz egy vektor tartozik.

Visszamérési függvények

A logisztikus regressziónál: logit transzformációt végeztünk, itt nincs ilyen.

Pontosság: a sorok mekkora arányában sikerült megtalálni a helyes választ

Itt a kimenet numerikus lesz

- Egy olyan hipersíkot/egyenest keresünk, amitől a pontok szórása minimális.
- CV-t és PCV-t(?) hasonlítjuk össze

Alapvetően a lineáris regresszió a négyzetes hibára optimalizál, viszont a visszamérési függvényeket elsősorban az üzleti célok tekintetében kell meghatározni, ezért használunk másokat is.

1. Négyzetes hiba

MSE (mean squared error)

$$\frac{\sum (y - \hat{y})^2}{n}$$

- A kiugró értékek túlságosan is befolyásolhatják
- Nem mutatja az eltérés irányát
- Nem a célváltozó nagyságrendjében (értékkészletében) adja meg a hibát hanem annak négyzetében.

2. Abszolút hiba (?)

$$\frac{|(y - \hat{y})^2|}{n}$$

[Ez nem biztos, hogy így néz ki]

3. Abszolút különbség

$$\frac{|(y - \hat{y})|}{n}$$

Ez akkor használható, ha nem az egyes sorok szerinti pontosság a fontos, hanem az összesített becsülhetőség. Ilyen példa az energia kereskedés, ahol elsősoraban a keresletet és a kínálatot kell kiegyenlíteni, és nem az egyiket kell növelni.

4. RMSE

Root mean square error

$$\sqrt{MSE}$$

Problémák a lineáris regresszióval

1. Hiányzó értékek

- Ahol hiányzik a CV: tesztadat lesz belőle.
- A hiányzó értékeket mindenképpen kezelnünk kell.

2. Model értelmezése

Csak akkor lehet a változókat súlyként használni, ha normalizáljuk, de itt ezt nem tesszük

3. Kategória változók

- Kihagyjuk
- Számmá kell valahogy konvertálni
    - Dummy változókat csinálhatunk belőlük:
        1. Túl sok plusz változót (tehát dimenziót) fog teremteni, csökkenti a model teljesítményét.
        2. Az adott változó sokszoros mértékben fog beszámítódni (pl négy db családi állapot négyszeres súlyt jelent).
        3. Az 'utolsó' változót akár ki is hagyhatjuk (lineáris regresszió pl nem szereti ha van összefüggés a változók között).
- Sorrendezhetőségnél nem tudjuk meghatározni a tényleges 'távolságot' a különböző értékek között, ezért problémás lehet.
- Weight of evidence: az adott értéknek a célváltozóhoz való kapcsolatát használja.
    - Hasonló lehet a változó gyakoriságát hozzárendelni.

4. Kiugró értékek

Maga felé rángatja az egyenest, a négyzet miatt ráadásul ezt még meg is erősíti.

5. Bemeneti változók

- A lineáris regressziónak nagyon nagy az extrapolációs képessége (szemben pl a döntési fával).
- A lineáris regresszió minden bemeneti változót (akár zajt is) beépít a modelben szolgai módon (szembe más modellekel).
- Megoldás: Step-wise, iteratívan csinálunk egy egyváltozós modelt és addig adogatunk hozzá továbbiakat, amig már nem tudjuk javítani a model pontosságát.
    - Ez viszont egy 'mohó' algoritmus: csak a lokális optimumot találja meg.