3. alkalom

Adatelemzési platformok, BME, 2018. Február 14.
I. elméleti óra

Esettanulmány: Hitelbírálati probléma

Három kategóriába tesszük a hitelt igénylőket.

1. Elutasítottak
2. Szürke sáv
3. Elfogadottak

Prediktív analitika

A célja, hogy az összes többi (bemeneti, vagy független) változó segítségével (egy adott példán belül) előre tudjam jelezni a célváltozót (függő változót).

Modelt tanítom a múltbeli adatokkal, ami alapján ő majd megbecsüli a későbbi célváltozókat.

- Lehetséges több célváltozót is megjósolni
- Egy speciális eset, amikor ezek hierarchikus viszonyban állnak egymással

Adatmodell

- 'Oszlopok' vagy 'Változók'
- 'Sorok' vagy 'Entitások' vagy 'Példák'

Adattípusok és hozzájuk kapcsolódó elemzési módszerek

- 'Kategória': Osztályozás
- 'Numerikus': Regresszió

De ezek nem annyira egyértelműek

Adatelőkészítés

Át kell kódolni az adatokat, hogy az RM helyesen tudja őket értelmezni:

- Adattípus: nominális, polinominális, numerikus, természetes, binominális stb.

- Adatszerep: célváltozó, azonosító, stb.

Sokféle adatot nem egyértelmű, hogy hogyan rögzítünk és használunk.

Credit scoring vizsgálati módok

- Application credit scoring: Az ügyfél által, a hitelkérelemmel kapcsolatban megadott adatokat tudjuk.
  - Rengeteg változó
  - Ezek egy sorban jelennek meg az adatmodellben
- Viselkedési: Az ügyfél korábbi viselkedéséről van adatunk
  - Itt egy ügyfélről több sorban is találunk adatokat
  - Múltra is vonatkozik
  - Nem (csak) saját bevalláson alapszik, hanem 'megfigyelésen'

Célváltozó

Nem egyértelmű, hogy ki a jó vagy rossz adós.

- Basel II alapján, Európában a rossz adós az aki akár csak egyszer is 3 * 30 napon meghaladóan tartozott.
  - Nyilvánosan csak erről van lista, tehát adat.
  - Menet közben a bankoknak folyamatosan becsülniük kell a bedőlt hiteleket és tartalékot kell kialakítaniuk ezzel kapcsolatban.

Tanulási adat minta

- Nem feltétlenül kell az egész adathalmazra nézve reprezentatívnek lennie, de pl sokkal inkább a jelenre és jövőre nézve.
- Hosszú futamidőnél nincs sok adat, ezért viszonylag friss adatokból kell dolgozni.
- Három hónapnál fiatalabb ügyletek viszont nem kerülnek be (mert ott még nem lehet tudni, hogy bedőlt-e vagy sem).

Konkrét gyakorlat

2 dimenzióban néz:

- Késedelmesek aránya
- Hitel élettartalama

Ez a függvény telítődik egy idő után és emiatt a telítődési utáni pont utáni élettartamú hiteleket érdemes behozni a tanuló adathalmazba.

Modellezés: Logisztikus regresszió

- Egyelőre osztályozási problémaként tekintünk erre.
- Logisztikus regresszió
  - Ez alapvetően egy osztályozási model
  - Feltételezzük, hogy az X és a célváltozó között van összefüggés
  - Mindegyik bemeneti változóhoz számol egy együtthatót (koefficienst)

$$X = \sum \omega_i * BV_i$$

- Az értéktartomány egyelőre $-\infty/\infty$, ezért logit transzformációt végzünk rajta (vagy normáljuk) a [0,1] tartományban.

$$P = \frac{e^x}{1 + e^x} \rightarrow [0, 1]$$

- Kifejezi, hogy mekkora valószínüséggel esik egyik vagy másik tartományba.
- Az attribútomok csak számszerű adatokat kezelnek
  - Ez nem foglalkozik azzal, hogy a változóknak mi az értelme
  - Numerikus változókat és hiányzó értékeket tartalmazhat, kategória változókkal nem tud mit kezteni
  - 'Dummyzás': kategória változókat át kell alakítani valamilyen numerikus változóvá.