4. alkalom

Adatelemzési platformok, BME, 2018. Február 20., III. gyakorlati óra.

Felügyelt tanítás fajtái

- Osztályozás: Logisztikus regresszió ide tartozik, mivel amit megpróbál megmagyarázni az
  - bináris vagy
  - kategória változó
- Regresszió

Adathalmaz

Brazil kereskedelmi bank hitelbírálati adatai. Valamennyire elő van már készítve.

Változó kódok

- 'f_': Flagek, binomiális
- 'cat_': Kategória, nominális
- 'num_': Numerikus

Együttható

- 0 és 1 még nem mondja meg, hogy mit fejez ki, ezt nekünk kell kideríteni
  - 0: rossz adós
  - 1-es: nem fizeti vissza
  - Például az 'age' együttható: negatív, vagyis minél idősebb annál kisebb valószínüséggel lesz rosz adós
- Az együttható önmagában még nem súly. Normalizálni kell a változókat, vagyis ugyanabba az értékkészletbe teszem őket, a [0, 1] tartományba

Normalizálás

Két főbb fajtája van

1. Range transzformáció: $X_i = \frac{x_i - x_{min}}{x_{max} - x_{min}} = [0, 1]$
2. Z-transzformáció (vagy 'standardizáció'): $\bar{x_i} = \frac{x - \bar{x}}{\sigma}$
   - átlaga: 0

    - szórása: 1
    - Az értéktartománya: $-\infty / +\infty$
    - Normálelszolás esetén majdnem mindig a +/-3 szórás tartományba esik, de itt nem tudunk semmit az eloszlásról

Normalizálás után már mások az együttható súlyok

- konfidencia(0,0): nullának mekkora a valószínüsége
- konfidencia(1,0): egynek mekkora a valószínüsége

Rapidminer annak az értéknek próbálja megbecsülni a valószínüségét, amelyikkel először találkozik. Mivel a kategória változókat számként tárolja, ezért nem tudhatjuk biztosra, hogy melyik valószínüséget becsli.

Teszt adathalmaz

- A teszt adathalmazt is normalizálni kellett volna!!
- Z-transzformáció esetén lehet ugyanazokat a súlyokat használni,
- Range-nél már lehet, hogy diszjunkt az adathalmaz (nagyon eltérő a min és max érték a két adathalmaznál, lehet, hogy nem is fedik egymást).
- Nem feltétlenül baj ha nem a 0-1 tartományon van normalizálva, mivel a model képes az extrapolációra,
- Rapidminer-ben: az Adatelőkészítő modelleknél a `preprocessing` bemenet segít ebben

Kiugró érték

- Például, num_age: 0.07,
- Logisztikus regresszió az adott sorra határozza meg az együtthatót, ezért rossz hatással van az egész modelre.

Best practice

- Training és tesztelési adathalmazt együtt normalizáljuk a közös min és max értékkel.
- De ezzel információt viszünk a jövőbe, ami vagy probléma vagy nem, a súlyokat mindenesetre befolyásolja.

Split validation operator

- Tanuló és tesztadathalmazra bontja az adathalmazt és méri a model hatékonyságát a teszten.
- A 'közel hasonló' eset már egy jó magyarázó model lehet.