

11. Óra

Adatelemzési platformok, BME, 2018. Március 27., V. Gyakorlati óra.

A legutolsó Rapidminer gyarkolat.

Adat

telco.txt

- churn: label, binomial
- customer id: id
- Data format: MM/dd/yyyy
- Gender: binomial

Döntési fa

Validation

- Training
 - Decision tree
- Testing
 - Apply model
 - Performance

Elsőre csak egy tönk az egész.

Nem egyértelmű, hogy a rapidminer melyik célváltozót fogja tekinteni primary outcome-nak (valamilyen sorrend alapján csinálja).

Akkor jó, ha a döntési fa pontosságok közel vannak egymáshoz is és az átlagos pontossághoz.

AUC

Az AUC: 0.5, ami gyakorlatilag a véletlenszerű eseteket jelöli ('egyenes vonal' a ROC

görbén: véletlenszerű kapcsolat). Ez ROC görbénél itt azt jelenti, hogy nincs különbség a konfidenciák között (mivel itt egyelőre csak egy levél van).

ROC (threshold): Ez maga a konfidencia. Itt nagyon gyorsan alacsony értéket vesz fel.

- A ROC görbét nem a konfidencia érdekli, hanem konfidencia alapján képzett sorrendet.

Pesszimista/optimista ROC görbe (hogyan kötjük össze a pontokat).

ROC görbe alapok

- Konfidencia (döntési fánál): az 1-esek és 0-asok aránya[?].
- Konfidencia (logisztikus regresszió): maga a logisztikus regresszió egyenes.
- Pontosság: a helyes és helytelen értékek aránya.

Decision tree

- Pruning: Miután felépítette a fát, visszanyes egyes ágakat valamilyen konfidencia kritérium alapján.
- Prepruning: A fa kiépítése előtt már lenyes.

Futtatások

- Prepruning nélkül: 20 mélységű, szálamizós fa.
- Maximal Depth = 5: Ez csak egy vágást csinál (az egyik egy elemű).
- Gini index: Ez már egy jobb model.

Lift görbe

Compare ROCs paraméter

- KNN
- Döntési fa
- ROC