Úspěšnost podle formy studia

KIV/DBM2

Pavel Třeštík

Téma

- Úspěšnost podle formy studia
- Úkol je zjistit úspěšnost prezenční/ kombinované/ distanční formy jiným způsobem než je podíl: |úšpěšný absolventi| / |studenti|
- Primárně založeno na známkách
  - Statut, počet kreditů, pokus,…
- Výstup: model, který odhadne úspěšnost
  - Z modelu by mělo jít poznat, které příznaky jsou důležité pro úspěšnost

Realizace

- Myšlenka při vybírání tématu
  - Výstup: SQL nebo regresní model (podle počtu příznaků)
  - Vybrat co největší počet příznaků, které by se ohodnotily Wekou
  - Podle výstupu z Weky → SQL nebo model
- Reálně
  - Výběr příznaků byl proveden manuálně
  - SQL určitě nemožné – jak “vymyslet z ničeho” rovnici? Nejde
  - Model – regresní? Ten ale potřebuje reálné hodnoty k učení..

Realizace

- Data neposkytly vhodný způsob reálných hodnot
  - → použití klasifikačního modelu
- Klasifikátor také potřebuje “správné” hodnoty z dat, kvůli učení
  - Třídy z dat lze poměrně snadno vytvořit
  - Třídami se staly kombince: “absolvoval” + “forma studia”
    - př. 1P (absoloval prezenční), 0P (“neabsoloval” prezenční), 1D
- Úspěšnost založena na známkách – za každého studenta až destíky záznamů
  - Celkem ~2.7M záznamů z toho ~460k unikátních kombinací
  - Vybrané příznaky ve výsledkách

Realizace

- Použitým modelem byl Naivní Bayes
  - Porušení podmínky závislosti příznaků
  - Navzdory porušení podmínky, by měly být výsledky poměrně dobré
  - Rychlý
- Pokus použít SMO (implementace SVM) pro porovnání
  - Velmi pomalé trénování modelu
    - 50k záznamů – několik minut, 100k+ hodiny, destíky hodin
  - I při vytvoření modelu s 50k záznamy, velmi pomalá klasifikace (desítky hodin)

Výsledky

Sloupec	Alias
~~NOVE_PRIJATY~~	~~STD_NOVE_PRIJATY~~
~~STUPEN_PRED_VZDELANI~~	~~STD_STUPEN_PRED_VZDELANI~~
~~POCET_ZAPISOVYCH_PROPUSTEK~~	~~STD_POCET_PROPUSTEK~~
FORMA	SP_FORMA
FAKULTA_SP	SP_FAKULTA_PROGRAMU
~~ROK_PLATNOSTI~~	~~SVR_ROK~~
POC_KRED	ZN_KREDITY_ZA_PREDMET
STATUT	ZN_STATUT_PREDMETU
POKUS_CISLO	ZN_POKUS_CISLO
HODNIDNO_ZKZP	ZN_HODNOCENI
TYP_ZK	ZN_TYP_ZKOUSKY
PRAC_ZKR	ZN_PRACOVISTE_ZKRATKA
ZAPOCET_POKUS	ZN_ZAPOCET_POKUS

Tabulka 3: Použité vlastnosti

Výsledky

```
=== Summary ===

Correctly Classified Instances     2263619               83.1571 %
Incorrectly Classified Instances    458482               16.8429 %
Kappa statistic                          0.584
Mean absolute error                      0.0666
Root mean squared error                  0.1899
Relative absolute error                 50.4718 %
Root relative squared error             73.9342 %
Total Number of Instances          2722101

=== Detailed Accuracy By Class ===

               TP Rate  FP Rate  Precision  Recall   F-Measure  MCC      ROC Area  PRC Area  Class
               0.914    0.005    0.956      0.914    0.934      0.928    0.999     0.989     0K
               0.890    0.295    0.906      0.890    0.898      0.584    0.914     0.971     0P
               0.427    0.096    0.386      0.427    0.405      0.317    0.855     0.362     1P
               0.884    0.001    0.874      0.884    0.879      0.878    1.000     0.868     0D
               0.590    0.008    0.418      0.590    0.489      0.491    0.990     0.396     1K
               0.384    0.001    0.359      0.384    0.371      0.371    0.999     0.331     1D
Weighted Avg.  0.832    0.237    0.841      0.832    0.836      0.585    0.916     0.890

=== Confusion Matrix ===

      a       b       c       d       e       f   <-- classified as
 245444     132       0     107   22799       0 |       a = 0K
      0 1842874  227981     339       0       0 |       b = 0P
      0  192163  143168       4       0       0 |       c = 1P
     80       0       0   14722       0    1853 |       d = 0D
  11347       2       2       5   16371       0 |       e = 1K
      0       0       0    1668       0    1040 |       f = 1D
```

Obrázek 1: Výstup N. Bayes klasifikátoru

Výsledky

	Stat.	N. Bayes
prezenční	$\frac{335335}{2406529} = 13.93\%$	$\frac{143168}{2406529} = 5.95\%$
kombinované	$\frac{27727}{296209} = 9.36\%$	$\frac{16371}{296209} = 5.53\%$
distanční	$\frac{2708}{19363} = 13.99\%$	$\frac{1040}{19363} = 5.37\%$

Tabulka 2: Porovnání úspěšnosti modelu a statistiky

Problémy

- Nejasné jména a popisky některých sloupců a způsob uložení dat
- Způsob realizování modelu – klasifikátor nebyl ani možností při vymýšlení tématu
- Podezřele dlouhé trénování a použití jiných modelů než NB

Konec

Otázky?