Podstawowe Metody Klasyfikacji

Wojciech Kosztyła, Wojciech Węgrzynek

Aby uzyskać zbiór danych do klasyfikacji połączyliśmy zbiór danych o hrabstwach z amerykańskiego censusu 2015 z danymi o wynikach wyborów prezydenckich 2016. Klasami były wówczas *Trump* i *Clinton* w zależności od tego który kandydat uzyskał większość głosów w danym hrabstwie.
Z naszego datasetu musieliśmy więc usunąć hrabstwa z Alaski (gdzie wybory prezydenckie są prowadzone w granicach różnych od hrabstw) oraz Puerto Rico (gdzie wybory prezydenckie nie odbywają się).
Cechy zostały znormalizowane.

Regresja logistyczna

Przeprowadziliśmy wpierw klasyfikację logistyczną, najpierw na wszystkich dostępnych cechach (później będziemy nazywać ten model modelem 1), a następnie na tych, które można było uznać za istotne (co najmniej “.”) na podstawie tej pierwszej (ten model nazwiemy modelem 2).
Na wszystkich cechach osiągnęliśmy dewiację 1055.3 (AIC: 1119.3) gdzie dla zerowego modelu mamy dewiację 2699.7. Z kolei w przypadku ograniczenia się do cech uznanych za istotne otrzymaliśmy dewiację na poziomie 1415.7 (AIC: 1433.7). W obu przypadkach dostaliśmy więc wyraźne polepszenie. Ciekawa jest dosyć spora różnica pomiędzy modelami, być może wskazuje ona na to, że odrzucone dane są mimo wszystko w jakiś sposób ważne.
W obu przypadkach klasa 1 odpowiada hrabstwom gdzie wygrał Donald Trump, a więc dodatnie wartości współczynników oznaczają większe prawdopodobieństwo takiego hrabstwa.
Przez ~0 rozumiem wartości poniżej 10^-6.

Istotne Dane:

- White:
    - Opis: Procent populacji identyfikujący się jako biały.
    - Model 1, P(>|z|)= ~0.002
    - Model 2, P(>|z|)=~0.0
    - Współczynnik: w obu modelach dodatni, ~19 dla modelu 1, ~12.5 dla modelu 2. Jedna z najbardziej wpływowych cech.
- Pacific:
    - Opis: Procent ludności o etniczności pacyficznej.
    - Model 1, P(>|z|)=~0.06

    - Model 2, P(>|z|)=~0.0003
    - Współczynnik: W obu przypadkach ujemny, ~-8.5 dla modelu 1, ~-13 dla modelu 2, najbardziej wpływowa cecha w modelu 2.
- Citizen:
    - Opis: Procent populacji posiadający obywatelstwo amerykańskie.
    - Model 1, P(>|z|)=~0.00002
    - Model 2, P(>|z|)=~0
    - Współczynnik: w obu modelach ujemny, ~-5 dla modelu 1, -7 dla modelu 2.
- Income:
    - Opis: Mediana dochodu gospodarstwa domowego.
    - Model 1, P(>|z|)=~0.05
    - Model 2, P(>|z|)=~0.015
    - Współczynnik: W obu przypadkach dodatni, ~4 dla modelu 1, ~3.5 dla modelu 2.
- IncomePerCap:
    - Opis: Zarobki per capita.
    - Model 1, P(>|z|)=~0.007
    - Model 2, P(>|z|)=~0
    - Współczynnik: W obu przypadkach ujemny, ~-6.5 dla modelu 1, ~-1129 dla modelu 2.
- MeanCommute:
    - Opis: Średni czas dojazdu do pracy.
    - Model 1, P(>|z|)=~0.08
    - Model 2, P(>|z|)=~0.33 (nieistotny)
    - Współczynnik: ~-1.3 dla modelu 1, ~0.5 dla modelu 2.
- Employed:
    - Opis: Procent ludności, który stanowią osoby po 16 roku życia i są zatrudnione.
    - Model 1, P(>|z|)=~0
    - Model 2, P(>|z|)=~0
    - Współczynnik: W obu przypadkach ujemny, ~-6.5 dla modelu 1, ~-11 dla modelu 2.
- Unemployment:
    - Opis: Współczynnik bezrobocia.
    - Model 1, P(>|z|)=~0
    - Model 2, P(>|z|)=~0
    - Współczynnik: W obu przypadkach ujemny, ~-6.5 dla modelu 1, ~-11 dla modelu 2.

Obserwacje:
MeanCommute wyglądało w modelu 1 jakby miało szansę być istotne (“.”), ale w modelu 2 zupełnie straciło na wiarygodności. Z kolei Pacific wyglądało podobnie w modelu 1, ale w modelu 2 znacząco zyskało na wiarygodności.
Potencjalnie silnie skorelowane cechy (Unemployment i Employed oraz Income i IncomePerCap) nie wpływają na model zgodnie z taką korelacją, ale raczej wspólnie modelują bardziej złożone zależności.

Porównanie modeli regresji

Aby porównać modele regresji podzieliliśmy nasz dataset na zestaw “train” i “test” w proporcjach 80% do 20%.
Modele: “regresja logistyczna”, “LDA”, “QDA” i “kNN” były uczone na zbiorze “train”, a zbiór “test” był wykorzystywany do sprawdzania wyników.

W przypadku uczenia modelu “kNN” wymagane było podanie wartości parametru “k”. Dobór tego parametru jest nietrywialny, więc napisaliśmy skrypt sprawdzający wartość “accuracy” wyniku dla różnych wartości “k” i wybraliśmy w ten sposób najlepiej sprawdzającą się wartość (`Best accuracy: 0.929373996789727 Best k: 9`).

Zebrane wyniki wszystkich czterech modeli:

```
Regresja logistyczna    accuracy = 0.905296950240771
LDA                     accuracy = 0.906902086677368
QDA                     accuracy = 0.874799357945425
kNN   k=9               accuracy = 0.929373996789727
```

Własność “accuracy” tych czterech modeli nie odbiega znacząco od siebie, jednak najlepiej poradził sobie model “kNN”, a najgorzej “QDA”.

```
Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.6696   0.0387   0.1022   0.2489   3.8914

Coefficients: (1 not defined because of singularities)
              Estimate Std. Error z value Pr(>|z|)
(Intercept)   249.9080   169.8553   1.471  0.14121
TotalPop       -0.5835     3.5789  -0.163  0.87049
Men            -1.5579     1.3568  -1.148  0.25086
Women              NA         NA      NA       NA
Hispanic        6.3327     5.9402   1.066  0.28638
White          18.9854     6.1831   3.071  0.00214 **
Black           5.0260     5.1924   0.968  0.33307
Native          8.1684     5.9116   1.382  0.16705
Asian          -0.1018     3.9146  -0.026  0.97925
Pacific        -8.4599     4.5016  -1.879  0.06021 .
Citizen        -5.7104     1.3551  -4.214 2.51e-05 ***
Income          4.3566     2.2037   1.977  0.04805 *
IncomePerCap   -6.5571     2.4283  -2.700  0.00693 **
Poverty        -1.9285     1.8481  -1.044  0.29671
ChildPoverty   -0.3715     1.5454  -0.240  0.81002
Professional   -0.3504    77.4107  -0.005  0.99639
Service        -2.6792    40.4424  -0.066  0.94718
Office          5.8746    40.0178   0.147  0.88329
Construction    9.3563    49.3605   0.190  0.84966
Production      4.1628    52.3368   0.080  0.93660
Drive        -115.7742   102.9358  -1.125  0.26071
Carpool       -38.8318    34.8005  -1.116  0.26449
Transit       -97.2830    71.7736  -1.355  0.17529
Walk          -48.1789    36.8955  -1.306  0.19161
OtherTransp   -27.4421    23.1231  -1.187  0.23531
WorkAtHome    -51.1187    43.2702  -1.181  0.23745
MeanCommute    -1.3333     0.7667  -1.739  0.08204 .
Employed      -10.1350     1.3934  -7.273 3.50e-13 ***
PrivateWork  -121.0363    84.4753  -1.433  0.15191
PublicWork   -119.5489    86.7235  -1.379  0.16805
SelfEmployed  -72.6070    52.5523  -1.382  0.16709
FamilyWork    -11.3885    14.2118  -0.801  0.42293
Unemployment   -5.6234     1.0478  -5.367 8.02e-08 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2699.7  on 3110  degrees of freedom
Residual deviance: 1055.3  on 3079  degrees of freedom
```

```
Call:
glm(formula = Victor ~ White + Citizen + Income + IncomePerCap +
    Employed + Unemployment + Pacific + MeanCommute, family = binomial,
    data = data_for_classification)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-3.7146   0.0736   0.1749   0.3509   2.6331

Coefficients:
             Estimate Std. Error z value Pr(>|z|)
(Intercept)    8.6193     0.7745  11.129  < 2e-16 ***
White         12.4859     0.6349  19.666  < 2e-16 ***
Citizen       -7.1397     0.8649  -8.255  < 2e-16 ***
Income         3.4163     1.4267   2.395 0.016638 *
IncomePerCap -11.2904     1.6938  -6.666 2.64e-11 ***
Employed     -11.4631     1.0572 -10.843  < 2e-16 ***
Unemployment  -8.0903     0.8595  -9.413  < 2e-16 ***
Pacific      -13.0576     3.6476  -3.580 0.000344 ***
MeanCommute    0.5043     0.5224   0.965 0.334317
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 2699.7  on 3110  degrees of freedom
Residual deviance: 1415.7  on 3102  degrees of freedom
  (1 observation deleted due to missingness)
AIC: 1433.7

Number of Fisher Scoring iterations: 6
```