

Sztuczna inteligencja i inżynieria wiedzy – laboratorium

Sprawozdanie: Uczenie maszynowe

Kajetan Pynka, 254495

Spis treści

Wstęp.....	3
Klasyfikator bayesowski	4
SVM.....	5
Wnioski.....	6

Wstęp

Problemem rozważanym w ramach tego zadania jest klasyfikacja książek tzn. określenie ich przynależności gatunkowej na podstawie dostarczonego streszczenia. Dane pochodzą ze zbioru CMU Book Summary Dataset. W ramach zadania wykorzystałem dwa podejścia: naiwny klasyfikator bayesowski oraz maszynę wektorów nośnych (SVM). W ramach implementacji wykorzystałem gotową bibliotekę scikit-learn dla języka Python.

W ramach wstępnego oczyszczania danych wykonałem następujące kroki:

- Pozbyłem się wszystkich kolumn poza kolumną gatunków oraz kolumną streszczenia
- Pomiąłem rekordy, które nie posiadały określonego gatunku (NULL w kolumnie)
- Pomiąłem rekordy, których streszczenie było krótsze niż 100 znaków (w tym NULL'e)
- Wybrałem tylko rekordy, które jako gatunek mają przypisany jeden z 6 najczęściej występujących gatunków.
- Dla rekordów n-gatunkowych (tzn. dla danego streszczenia było przypisanych n gatunków) próbowałem je najpierw rozbić na n rozłącznych rekordów (to samo streszczenie pojawiałoby się n-krotnie). Zauważyłem jednak, że dokładność klasyfikacji jest dość niska, więc przyjąłem, że biorę pierwszy lepszy gatunek.
- Na koniec z każdego streszczenia usunąłem zbędne znaki interpunkcyjne i pozostawiłem jedynie litery a-z (wszystko zrzutowane do małych liter).

W ramach treningu zastosowałem 10-krotną walidację krzyżową. Dane podzielone zostały w 90% na dane treningowo-walidacyjne oraz w 10% na dane testowe. W ramach SVM zastosowałem model SVC (C-support Vector Classification). W ramach klasyfikatora bayesowskiego zastosowałem model MultinomialNB.

Do wyznaczania „najlepszych” hiperparametrów skorzystałem z biblioteki scikit-optimize.

Do ekstrakcji cech wykorzystałem klasę TfidfVectorizer.

Klasyfikator bayesowski

Wyniki po przepuszczeniu przez optymalizator:

```
siema 2022-06-14 12:53:41 ~/Desktop/uczelnia/AI/zad4 @main $ python .\nb.py
{'Fiction': 4747, 'Speculative fiction': 4314, 'Science Fiction': 2870, 'Novel': 2463, 'Fantasy': 2413, "Children's literature": 2122}
C:\Python310\lib\site-packages\skopt\optimizer\optimizer.py:449: UserWarning: The objective has been evaluated at this point before.
  warnings.warn("The objective has been evaluated "
0.5204795204795205
0.5283628067579234
OrderedDict([('alpha', 0.21128133153466064)])
MultinomialNB(alpha=0.21128133153466064)
siema 2022-06-14 12:55:18 ~/Desktop/uczelnia/AI/zad4 @main $
```

Najlepsze wyniki udało mi się osiągnąć dla wartości parametru $\alpha \approx 0.21128$:

- Średnia z wyników precyzji 10-krotnej walidacji krzyżowej $\approx 52.04\%$
- Uzyskana dokładność predykcji zbioru testowego $\approx 52.84\%$

SVM

```
siema 2022-06-14 11:30:13 ~/Desktop/uczelnia/AI/zad4 @main $ python .\svm.py
{'Fiction': 4747, 'Speculative fiction': 4314, 'Science Fiction': 2870, 'Novel': 2463, 'Fantasy': 2413, "Children's literature": 2122}
She was really happy to meet this lovely teddy bear! He had many friends and tiger was the best of them.
["Children's literature"]
OrderedDict([('C', 4.286496948879817), ('degree', 6), ('gamma', 1.6994661213725195), ('kernel', 'rbf')])
SVC(C=4.286496948879817, degree=6, gamma=1.6994661213725195)
0.5289160192378838
siema 2022-06-14 12:19:30 ~/Desktop/uczelnia/AI/zad4 @main $
```

Najlepsze wyniki udało mi się osiągnąć dla wartości parametrów $C \approx 4.2865$, $\text{degree} = 6$, $\text{kernel} = \text{'rbf'}$ oraz $\gamma \approx 1.6995$:

- Uzyskana dokładność predykcji zbioru testowego $\approx 52.89\%$

Wnioski

```
siema 2022-06-14 12:57:00 ~/Desktop/uczelnia/AI/zad4 main $ python .\nb.py  
{'Fiction': 4747, 'Speculative fiction': 4314, 'Science Fiction': 2870, 'Novel': 2463, 'Fantasy': 2413, 'Children's literature': 2122}  
0.5228888888888889  
0.5394605394605395  
siema 2022-06-14 12:57:23 ~/Desktop/uczelnia/AI/zad4 main $ python .\svm.py  
{'Fiction': 4747, 'Speculative fiction': 4314, 'Science Fiction': 2870, 'Novel': 2463, 'Fantasy': 2413, 'Children's literature': 2122}  
0.5296666666666666  
0.4975024975024975  
siema 2022-06-14 13:07:54 ~/Desktop/uczelnia/AI/zad4 main $
```

Ostatecznie wyniki w obu podejściach wyszły dość podobne (w granicach 50-54% skuteczności klasyfikacji). Co ciekawe to klasyfikator bayesowski przetrenował cały model w ciągu kilkunastu sekund natomiast na wyniki od SVC musiałem czekać około 10min (a wyniki wyszły lekko gorsze).

Przyjmując podejście w pełni losowe, dla sześciu gatunków można by się spodziewać dokładności w okolicach 16-17% także wynik 50-54% jest w pewnym stopniu satysfakcjonujący.