Sztuczna inteligencja i inżynieria wiedzy – laboratorium  
  
Sprawozdanie: Uczenie maszynowe  
  
  
Kajetan Pynka, 254495

Spis treści

[Wstęp 3](#_Toc106110266)

[Klasyfikator bayesowski 4](#_Toc106110267)

[SVM 5](#_Toc106110268)

[Wnioski 6](#_Toc106110269)

# 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
* Pominąłem rekordy, które nie posiadały określonego gatunku (NULL w kolumnie)
* Pominąłem 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 ja 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).

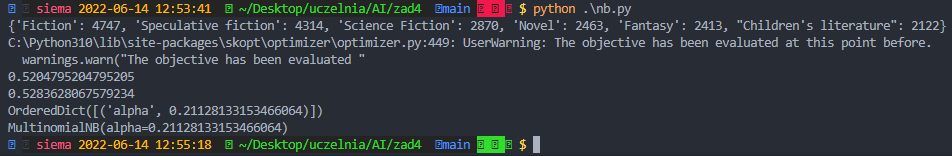
W ramach treningu zastosowałem 10-krotną walidację krzyżową. Dane podzielone zostały w 90% na dane treningowo-walidacyjne oraz w 10% na dane testowe. W ramach SVM zastosowałem model SVC (C-support Vector Classification). W ramach klasyfikatora bayesowskiego zastosowałem model MultinomialNB.

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

Do ekstrakcji cech wykorzystałem klasę TfidVectorizer.

# Klasyfikator bayesowski

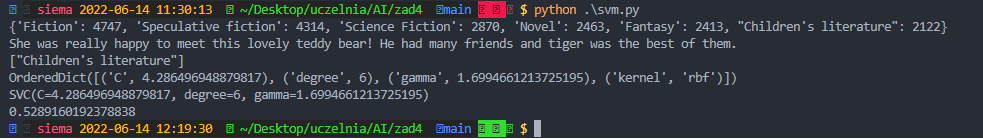
Wyniki po przepuszczeniu przez optymalizator:



Najlepsze wyniki udało mi się osiągnąć dla wartości parametru α ≈ 0.21128:

* Średnia z wyników precyzji 10-krotnej walidacji krzyżowej ≈ 52.04%
* Uzyskana dokładność predykcji zbioru testowego ≈ 52.84%

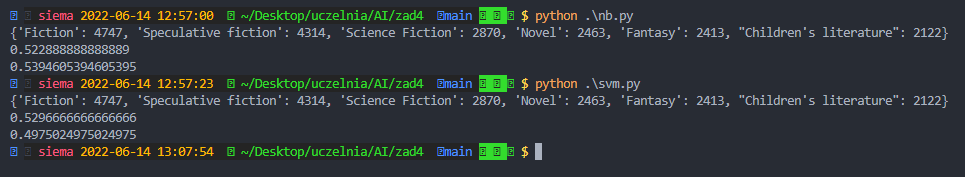
# SVM



Najlepsze wyniki udało mi się osiągnąć dla wartości parametrów C ≈ 4.2865, degree = 6, kernel=’rbf’ oraz γ ≈ 1.6995:

* Uzyskana dokładność predykcji zbioru testowego ≈ 52.89%

# Wnioski



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.