Politechnika Łódzka
Wydział Fizyki Technicznej, Informatyki
i Matematyki Stosowanej

Praca końcowa

Klasyfikacja bólów głowy z wykorzystaniem algorytmów uczenia maszynowego*

Konrad Pławik

Promotor:
dr hab. inż. Agnieszka Wosiak

Czerwiec 2024

* SVN: `https://github.com/kplawik/HeadacheClassification`

Spis treści

Spis rysunków

1. Wstęp

Bóle głowy bywają trudne do sklasyfikowania. O ile z obserwacji własnych miałem niestety okazję się o tym przekonać to nawet i świat nauki od lat również boryka się z tym problemem. Brytyjski instytut znany jako Headache Classification Committee of the International Headache Society (IHS) rozróżnia 13 kategorii bólów głowy - a samej tylko migreny - 29 typów [1]. Co więcej instytut ten wyraźnie mówi o tym że pacjent może cierpieć na więcej niż jeden z rodzaj ([1] punkt 9 we wstępie). Badania przeprowadzone przez EHF (European Headache Federation) [2] również potwierdzają że dominujący ból głowy nie musi być jedynym [3].

W pomocą przychodzi nam zagadnienie Uczenia Maszynowego oraz powiązane z nim algorytmy klasyfikacyjne. Poniższa praca dokumentuje wyniki kilkudziesięciu eksperymentów mających na celu automatyczną klasyfikację przy użyciu zarówno algorytmów regresyjnych (np. kNN) jak i głębokich Sieci Neuronowych (Deep Learning).

2. Dane

2.1. Zbiór danych

Wykorzystany zbiór danych pozyskano z serwisu `codeocean.com` [4]. Zbiór ten udostępniona na licencji GNU General Public License (GPL) a jego autorami są:

1. Paola A. Sánchez-Sánchez
2. José Rafael García-González
3. Juan Manuel Rua Ascar.

Cała trójka z pochodzi Universidad Simón Bolívar, Barranquila w Kolumbii.

Zbiór zawierał anonimowe dane 400 rozpoznanych przypadków a każdy z przypadków 23 cechy. Cechy miały różny typ (np. wiek pacjenta (typ całkowity) czy wystąpienie danego objawu (typ binarny)) co przemawiało za użyciem normalizacji przy użyciu MinMaxScalera z biblioteki Scikit-learn.

W zbiorze znajdowały się dane dotyczące 7 rodzajów bólu głowy. Zbiór nie był zbiorem zbalansowanym (co należy mieć na uwadzę w dajszej analizie):

```
Type
Basilar-type aura              18
Familial hemiplegic migraine   24
Migraine without aura          60
Other                          17
Sporadic hemiplegic migraine   14
Typical aura with migraine    247
Typical aura without migraine  20
dtype: int64
```

Zbiór nie posiadał brakujących danych więc nie zaistniała konieczność imputacji.

2.2. Informacje prawne

Zbiór udostępniony został na licencji GNU General Public License (GPL) [4].

Wykorzystane oprogramowanie korzystało z licenji:
Język Python: Python Software Foundation License [5]
Biblioteka Pandas: BSD 3-Clause License [6]
Biblioteka NumPy: BSD 3-Clause License [7]
Biblioteka Seaborn: BSD 3-Clause License [8]
Biblioteka TensorFlow: Apache License 2.0 [9]

Wspomniane biblioteki zostały szczegołowo opisane w następujących publikacjach naukowych:
Język Python [10]
Pandas: `https://zenodo.org/records/10957263` [11]
NumPy: `https://www.nature.com/articles/s41586-020-2649-2` [12]

Seaborn: https://joss.theoj.org/papers/10.21105/joss.03021 [13]
TensorFlow: https://zenodo.org/records/10798587 [14]

3. Klasyfikacja

3.1. Wstęp teoretyczny

3.1.1. Klasyfikator kNN (k-najbliszych sąsiadów)

Klasyfikator kNN, ze względu na swoją intuicyjność, jest jednym z najpopularniejszych klasyfikatorów. Działa on zgodnie z regułą: obserwacja x zostaje sklasyfikowana do najliczniejszej klasy z pośród k obserwacji najbliszych punktowi x.

Szacowane prawdopodobieństwo przynaleności obserwacji x do danej klasy wsród x najbliszych sąsiadów, zapisujemy jako:

$$\hat{j}|\mathbf{x} = \frac{1}{k}\sum_{i=1}^{n} l\left(\rho(\mathbf{x},\mathbf{x}_i) \leq \rho(\mathbf{x},\mathbf{x}^{(k)})\right) l(y_i = j), \quad j = 1,\ldots,g \tag{1}$$

gdzie:
$x^{(k)}$ - jest k-tym co do odległości x punktem z próby uczącej
ρ - jest pewną odległością, określaną jako miara niepodobieństwa.

3.1.2. Naiwny Klasyfikator Bayesa

Jest klasyfikatorem probablistycznym, który opiera się na uyciu twierdzenia.

$$P(C|F_1,\ldots,F_n) \tag{2}$$

gdzie:
C - oznacza zmienną zależną, będącą zbiorem etykiet klas
$F_1,\ldots,F_n$ - cechami opisującymi zbiór przypadków.

Literatura

[1] https://www.researchgate.net/publication/291331282_The_International_Classification_of_Headache_Disorders_3rd_edition_beta_version
[2] https://www.ehf-headache.com/
[3] https://link.springer.com/article/10.1186/s10194-018-0909-4
[4] https://codeocean.com/capsule/1269964/tree/v1
[5] https://docs.python.org/3/license.html
[6] https://github.com/pandas-dev/pandas/blob/main/LICENSE
[7] https://github.com/numpy/numpy/blob/main/LICENSE.txt
[8] https://github.com/mwaskom/seaborn/blob/master/LICENSE.md
[9] https://github.com/tensorflow/tensorflow/blob/master/LICENSE
[10] Van Rossum, Guido and Drake, Fred L.
"Python 3 Reference Manual", 2009
ISNB 1441412697
[11] https://zenodo.org/records/10957263
[12] https://doi.org/10.1038/s41586-020-2649-2
[13] https://joss.theoj.org/papers/10.21105/joss.03021
[14] https://zenodo.org/records/10798587