

Raport zad4

Alesia Filinkova

336180

1 Treść polecenia

- Zaimplementować algorytm regresji logistycznej.
- Sprawdzić jakość działania algorytmu dla klasyfikacji na zbiorze danych Breast Cancer Wisconsin Diagnostic.
<https://archive.ics.uci.edu/dataset/17/breast+cancer+wisconsin+diagnostic>
- Policzyc wynik dla przynajmniej 3 różnych sposobów przygotowania danych, na przykład usuwając niektóre kolumny, dodając normalizację wartości.

2 Cel i opis eksperymentów

Celem tego ćwiczenia jest implementacja algorytmu regresji logistycznej oraz przetestowanie jego skuteczności na zbiorze danych Breast Cancer Wisconsin Diagnostic. Zbiór danych jest podzielony na część ucząca (75) i testową (25) w celu przeprowadzenia eksperymentów. Cztery różne metody przygotowania danych zostaną przetestowane:

- Metoda 1: Bez normalizacji oraz bez usuwania kolumn.
- Metoda 2: Normalizacja wartości atrybutów bez usuwania kolumn.
- Metoda 3: Usuwanie niektórych kolumn.
- Metoda 4: Normalizacja wartości atrybutów oraz usunięcie niektórych kolumn.

Jako miary jakości zostaną użyte:

- Celność (accuracy)
- F1
- AUROC (Area Under the Receiver Operating Characteristic Curve).

3 Przygotowanie środowiska i danych

Skrypt można uruchomić przez terminal za pomocą polecenia:

1. `git clone https://gitlab-stud.elka.pw.edu.pl/afilinko/wsi.git`
2. `python3 -m venv venv`
3. `source venv/bin/activate`
4. `cd /lab4`
5. `pip install -r requirements.txt`
6. `python3 main.py`

4 Wyniki

4.1 Metoda 1: Bez normalizacji oraz bez usuwania kolumn

Te wyniki będą wykorzystane do porównania w następnych metodach jako parametr "Przed zmianą"

Accuracy	F1	AUROC
0.94	0.93	0.97

Table 1: Metoda 1: Bez normalizacji oraz bez usuwania kolumn

4.2 Metoda 2: Normalizacja wartości atrybutów bez usuwania kolumn

	Accuracy	F1	AUROC
Przed normalizacją	0.94	0.93	0.97
Po normalizacji	0.97	0.96	1.00

Table 2: Metoda 2: Normalizacja wartości atrybutów bez usuwania kolumn

4.3 Metoda 3: Usuwanie niektórych kolumn

Usunięte były wszystkie kolumny z indeksem 1 na końcu

	Accuracy	F1	AUROC
Przed usunięciem niektórych kolumn	0.94	0.93	0.97
Po usunięciu niektórych kolumn	0.83	0.82	0.87

Table 3: Metoda 3: Usuwanie niektórych kolumn

Usunięty były wszystkie kolumny "radius" oraz "perimeter"

	Accuracy	F1	AUROC
Przed usunięciem niektórych kolumn	0.94	0.93	0.97
Po usunięciu niektórych kolumn	0.94	0.93	1.00

Table 4: Metoda 3: Usuwanie niektórych kolumn

4.4 Metoda 4: Normalizacja wartości atrybutów oraz usunięcie niektórych kolumn

Usunięty były wszystkie kolumny z indeksem 1 na końcu

	Accuracy	F1	AUROC
Przed normalizacją i usunięciem kolumn	0.94	0.93	0.97
Po normalizacji i usunięciu kolumn	0.99	0.98	1.00

Table 5: Metoda 4: Normalizacja wartości atrybutów oraz usunięcie niektórych kolumn

5 Wnioski

5.1 Wnioski z eksperymentu

Najlepsze wyniki uzyskano przy połączeniu obu metod: usunięcia niektórych kolumn oraz normalizacji wartości. Jednak po usunięciu kolumn bez normalizacji algorytm działał tym gorzej, im więcej usuniętych kolumn.

Normalizacja cech znacząco wpłynęła na poprawę wyników, co wskazuje na znaczenie przygotowania danych przed treningiem. Eksperymenty potwierdziły, że algorytm regresji logistycznej działa lepiej na znormalizowanych danych.

5.2 Własna interpretacja wyników

Wyniki potwierdzają tezę, że przygotowanie danych jest kluczowe dla jakości modelu. Dobrze przygotowane dane mogą znacząco poprawić wyniki klasyfikacji, co jest ważne zarówno dla algorytmów klasyfikacyjnych, jak i dla regresji logistycznej.

Metryki takie jak AUROC są bardzo czułe na przygotowanie danych, co pokazuje, jak istotne jest usunięcie niepotrzebnych kolumn i normalizacja.