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
- Policzyć wynik dla przynajmniej 3 różnych sposobów przygotowania danych, na przykład usuwajac niektóre kolumny, dodajac normalizacje 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 cześć uczaca (75) i testowa (25) w celu przeprowadzenia eksperymentów. Cztery różne metody przygotowania danych zostana 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 usuniecie miektórych kolumn.

Jako miary jakości zostana użyte:

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

3 Przygotowanie środowiska i danych

Skrypt można uruchomić przez terminal za pomoca 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 beda wykorzystane do porównania w nastepnych metodach jako parametr "Przed zmiana"

Accurancy	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

	Accurancy	F1	AUROC
Przed normalizacja	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

Usuniety były kolumny "radius1", "texture1", "perimeter1", "area1", "smoothness1", "compactness1", "concavity1", "concave$_p oints1$","$symmetry1$","$fractal_d imension1$"

	Accurancy	F1	AUROC
Przed usunieciem niektórych kolumn	0.94	0.93	0.97
Po usunieciu niektórych kolumn	0.83	0.82	0.87

Table 3: Metoda 3: Usuwanie niektórych kolumn

Usuniety były kolumny "radius1", "radius2", "radius3"

	Accurancy	F1	AUROC
Przed usunieciem niektórych kolumn	0.94	0.93	0.97
Po usunieciu niektórych kolumn	0.93	0.91	0.95

Table 4: Metoda 3: Usuwanie niektórych kolumn

4.4 Metoda 4: Normalizacja wartości atrybutów oraz usuniecie miektórych kolumn

Usuniety były kolumny "radius1", "texture1", "perimeter1", "area1", "smoothness1", "compactness1", "concavity1", "concave$_p oints1$", "$symmetry1$", "$fractal_d imension1$"

	Accurancy	F1	AUROC
Przed normalizacja i usunieciem kolumn	0.94	0.93	0.97
Po normalizacji i usunieciu kolumn	0.99	0.98	1.00

Table 5: Metoda 4: Normalizacja wartości atrybutów oraz usuniecie miektórych kolumn

5 Wnioski

5.1 Wnioski z eksperymentu

Najlepsze wyniki uzyskano przy połaczeniu obu metod: usuniecia niektórych kolumn oraz normalizacji wartości. Jednak po usunieciu kolumn bez normalizacji algorytm działa tym gorzej, im wiecej usunietych kolumn.

Normalizacja cech znaczaco wpłyneła na poprawe 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 potwierdzaja teze, że przygotowanie danych jest kluczowe dla jakości modelu. Dobrze przygotowane dane moga znaczaco poprawić wyniki klasyfikacji, co jest ważne zarówno dla algorytmów klasyfikacyjnych, jak i dla regresji logistycznej.

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