Praca domowa 2

Termin oddania: 22.11.2023

1 Wstęp

W tej pracy domowej przyjrzymy się modelom regresji logistycznej (wraz z regularyzacją) oraz modelom SVM (maszynom wektorów podpierających). Celem pracy jest sprawdzenie jak radzą sobie z zadaniem klasyfikacji na danych rzeczywistych.

2 Zbiór danych

W tym celu posłużymy się zbiorem danych `credit-g`. Więcej o danych i znaczeniu kolumn mnoża przeczytać w OpenML. Do dalszej pracy należy podzielić zbiór danych na treningowy i testowy.

```
# Kod do pobrania danych

from sklearn.datasets import fetch_openml

df = fetch_openml(data_id = 31)
y = df.target
X = df.data
```

3 Część 1

Dla zbioru treningowego przygotuj trzy jak najlepsze modele (można skorzystać w tym celu z kroswalidacji):

1. model regresji logistycznej,
2. model regresji logistycznej z regularyzacją $L1$,
3. model regresji logistycznej z regularyzacją $L2$.

Dla każdego z wytrenowanych modeli podaj miarę na zbiorze treningowym oraz testowym. Oblicz dokładność, czułość, precyzję, wartość AUC. Narysuj krzywą ROC.

Zinterpretuj otrzymane wyniki dla poszczególnych modeli. Czy da się wskazać, które zmienne są istotne w modelu a które nie? Jeżeli tak, to proszę je podać.

4 Część 2

Bazując na wiedzy z **Części 1** przygotuj model wektorów podpierających dla zbioru treningowego. Na wejściu można ograniczyć liczbę zmiennych oraz liczbę obserwacji, jednak należy uzasadnić wybór.

Dla wytrenowanego modelu podaj miarę na zbiorze treningowym oraz testowym. Oblicz dokładność, czułość, precyzję, wartość AUC. Narysuj krzywą ROC.

5 Szczegóły rozwiązania

Rozwiązanie powinno zawierać pliki:

- folder `Kody` zawierający wszystkie potrzebne kody do przygotowania rozwiązania zadania domowego,
- plik `NUMERINDEKSU_raport.pdf` opisujący wyniki analiz (maksymalnie 3 strony).

6 Ocena

Łączna liczba punktów do zdobycia jest równa 10.

Część 1 + 2 (*5 + 5 punktów)*

- jakość kodu (porządek, czytelność) - 1 punkt,
- jakość modeli - 1 punkt,
- wnioski - 1 punkt,
- raport - 2 punkty.

7 Oddanie pracy domowej

Wszystkie punkty z sekcji *Szczegóły rozwiązania* należy umieścić w katalogu ZIP o nazwie `NUMERINDEKSU_GR_PD2`, gdzie

$$\texttt{GR} = \begin{cases} 1 & \text{dla środa, 12:15,} \\ 2 & \text{dla środa, 16:15.} \end{cases}$$

Tak przygotowany katalog należy przesłać na adres *anna.kozak@pw.edu.pl* do dnia 22.11.2023 do godziny 23:59. Tytuł wiadomości: *[WUM][PD2] Nazwisko Imię, Numer grupy: GR.*