Praca domowa 2

Termin oddania: 26.11.2025

1 Wstęp

W tej pracy domowej przyjrzymy się regularyzowanym modelom regresji logistycznej oraz modelom SVM (maszynom wektorów podpierających). Celem pracy jest porównanie jakości tych modeli w rozwiązywaniu problemu klasyfikacji.

2 Zbiór danych i ich przygotowanie

W tym celu posłużymy się zbiorem danych $\mathcal{D} = (X, y)$, gdzie X `= df_X.csv,` y `= df_y.csv`. Aby przygotować dane do dalszej pracy należy podzielić zbiór $\mathcal{D}$ na treningowy i testowy w proporcji 8:2 ustawiając parametr `random_state = NUMER_INDEKSU`.

W kodzie należy uwzględnić proces podstawowej eksploracji danych, przygotowania danych i transformacji niezbędnych do wytrenowania modeli opisanych w dalszych częściach. W raporcie należy krótko podsumować wykonane kroki i przeprowadzone przekształcenia danych.

3 Część 1

Dla zbioru treningowego przygotuj cztery jak najlepsze modele:

1. model regresji logistycznej,
2. model regresji logistycznej z regularyzacją $L1$,
3. model regresji logistycznej z regularyzacją $L2$,
4. model regresji logistycznej z karą *elasticnet* (jest to technika regresji liniowej, która łączy regularyzację $L1$ i $L2$).

Jakość modeli powinna być oceniona:

- na zbiorze treningowym za pomocą walidacji krzyżowej (kroswalidacji)
- oraz na zbiorze testowym.

Ponadto dla każdego z modeli:

- podaj wielkości współczynników dla każdego z modeli (tabela może być podana na końcu raportu na oddzielnej stronie),
- podaj wartości hiperparametru C w przypadku modeli regularyzowanych z karą $L1$, $L2$, a w przypadku modelu z karą *elastic net* hiperparametrów $l1\_ratio$ i C,
- podaj miary jakości modeli na zbiorze treningowym oraz testowym. Oblicz metryki: dokładność, czułość, precyzję, wartość AUC. Na jednym wspólnym wykresie narysuj krzywe ROC dla każdego modelu.

Zinterpretuj otrzymane wyniki dla poszczególnych modeli, a także potencjalne przyczyny dlaczego dany model osiągnął najlepszą jakość predykcyjną (w ujęciu danej metryki). 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 (SVM) dla zbioru treningowego. Na wejściu można ograniczyć liczbę zmiennych oraz liczbę obserwacji, jednak należy uzasadnić ten 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 + 1 strona z tabelą ze współczynnikami modeli opisanych w Części 1).

6 Ocena

Łączna liczba punktów do zdobycia jest równa 15.

Przygotowanie danych (*1 punkt*)

Część 1 (*10 punktów*)

- jakość kodu (porządek, czytelność) - 1 punkt,
- jakość modeli - 4 punkty,
- wnioski - 3 punkty,
- raport - 2 punkty.

Część 2 (*4 punkty*)

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

7 Oddanie pracy domowej

Wszystkie punkty z sekcji *Szczegóły rozwiązania* należy umieścić w katalogu ZIP o nazwie `NAZWISKO_IMIE_GR_PD2` (bez polskich znaków), gdzie

$$\texttt{GR} = \begin{cases} 1 & \text{dla środa, 12:15 (AK),} \\ 2 & \text{dla środa, 12:15 (KW),} \\ 3 & \text{dla środa, 14:15 (AK),} \\ 4 & \text{dla czwartek, 14:15 (AK).} \end{cases}$$

Tak przygotowany katalog należy przesłać na adres *katarzyna.woznica@pw.edu.pl* do dnia 26.11.2025 do godziny 06:00. Prace przesłane po tym terminie będą miały odjęte 2 punkty za każdy rozpoczęty dzień spóźnienia. Tytuł wiadomości: *[WUM][PD2] Nazwisko Imię, Numer grupy: GR.*
Prace, które nie będą przygotowane w odpowiednim formacie nie będą oceniane.