Praca domowa 2

Jan Kwiecień 320565

1 Wstęp

Celem pracy domowej numer 2 było stworzenie trzech modeli logistycznych (bez kary, z regularyzacją ℓ_1, z regularyzacją ℓ_2) oraz modelu SVM. Następnie mieliśmy sprawdzić jak radzą sobie z zadaniem klasyfikacji na danych rzeczywistych. Korzystaliśmy z informacji o kredytobiorcach z nimieckiego banku Deutsche Mark.

2 Część 1

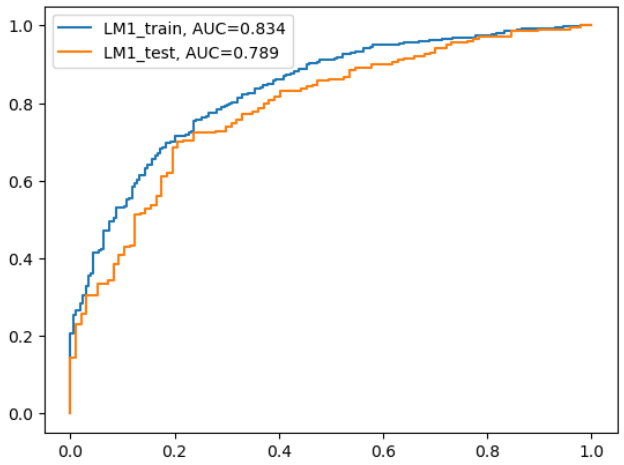
2.1 Przygotowanie danych

W pierwszej kolejności podzieliłem zbiór trenignowy i testowy w proporcji 7:3 oraz zamieniłem wszystkie dane na liczbowe (*pd.get_dummies*). W każdym z rozważanych modeli ustawiałem parametr *max_iter* = 1000, *random_state* = 320565 a w przypadku regularyzacji ℓ_1 i ℓ_2 testowałem parametr α od 0.1 do 10 z krokiem 0.1. W dwóch ostatnich modelach użyłem również pięciokrotnej kroswalidacji.

2.2 Model regresji logistycznej

Jako pierwszy zaimplementowałem model regresji logistycznej bez regularyzacji. Miary oraz krzywa ROC dla zbiorów treningowych i testowych przedastawiają się nastepująco:

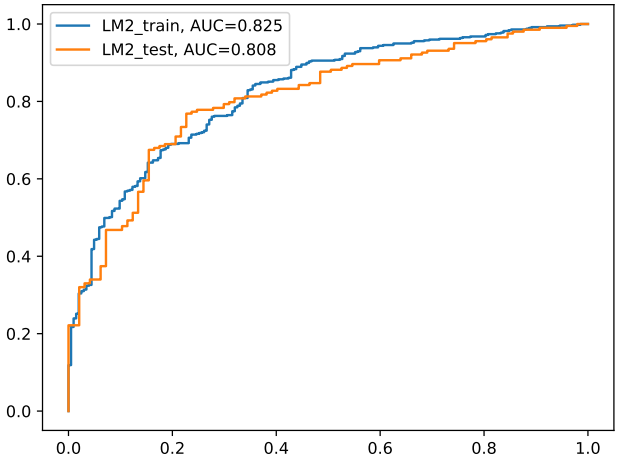
	precyzja	czułość	dokładność	wartość AUC
treningowy	0.823	0.907	0.796	0.834
testowy	0.775	0.882	0.747	0.789



2.3 Model regresji logistycznej z regularyzacją ℓ_1

Jako drugi zaimplementowałem model regresji logistycznej z regularyzacją ℓ_1. W tym przypadku najlepszym parametrem α okazała się $\alpha = 0.4$. Miary oraz krzywa ROC dla zbiorów treningowych i testowych przedastawiają się nastepująco:

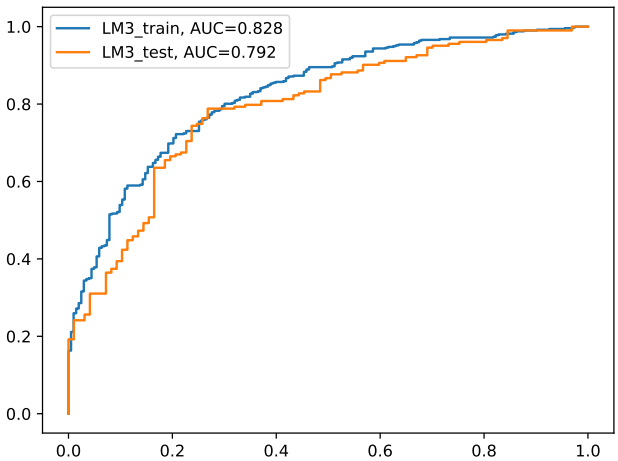
	precyzja	czułość	dokładność	wartość AUC
treningowy	0.809	0.918	0.787	0.825
testowy	0.782	0.882	0.753	0.808



2.4 Model regresji logistycznej z regularyzacją ℓ_2

Jako trzeci zaimplementowałem model regresji logistycznej z regularyzacją ℓ_2. W tym przypadku najlepszym parametrem α okazała się $\alpha = 0.2$. Miary oraz krzywa ROC dla zbiorów treningowych i testowych przedastawiają się nastepująco:

	precyzja	czułość	dokładność	wartość AUC
treningowy	0.809	0.909	0.783	0.828
testowy	0.768	0.882	0.74	0.792



2.5 Interpretacja wyników część 1

Wszsytkie modele dawały podobne rezultaty (z niewielką przewagą regularyzacji ℓ_1). Dla każdego z nich, największą wartość przyjmowała czułość, najniższą dokładność. W każdym przypadku lepsze predykcje uzsykaliśmy na zbiorze treningowym. Ponadto jesteśmy w stanie określić, które zmienne w modelu są nieistotne. W tym celu musimy sprawdzić, które współczynniki w zadaniu regresji logistycznej z regularyzacją ℓ_1 zerują się. Świadczy

to o niewielkim wpływie zmiennych na zachowanie modelu. Były to kolumny: 'residence_since', 'num_dependents', 'checking_status_0 <=X< 200', 'credit_history_delayed previously', 'purpose_domestic appliance', 'purpose_repairs', 'purpose_vacation', 'purpose_retraining', 'purpose_business', 'purpose_other', 'savings_status_100 <=X< 500', 'savings_status_500 <=X< 1000', 'employment_unemployed', 'employment_< 1', 'employment_1 <=X< 4', 'employment_>= 7', 'personal_status_female div/dep/mar', 'personal_status_female single', 'other_parties_none', 'other_parties_co applicant', 'property_magnitude_car', 'property_magnitude_no known property', 'other_payment_plans_bank', 'other_payment_plans_stores', 'housing_rent', 'housing_for free', 'job_unemp/unskilled non res', 'job_unskilled resident', 'job_high qualif/self emp/mgmt', 'own_telephone_none', 'foreign_worker_yes'.

3 Część 2

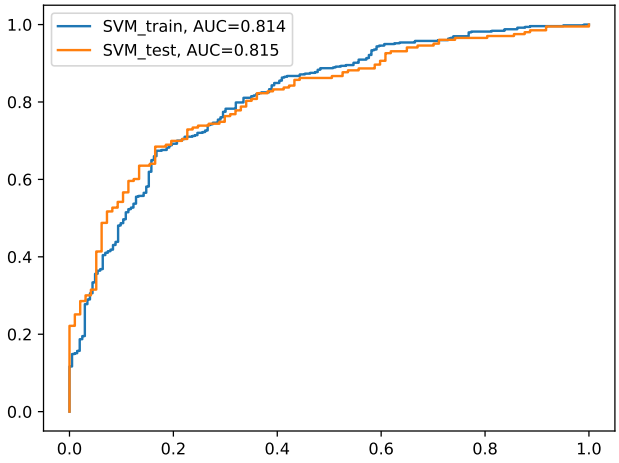
3.1 Przygotowanie danych

Bazując na wynikach uzyskanych z części 1, usunąłem te kolumny z ramki danych, dla których współczynniki najlepszego modelu w regularyzacji ℓ_1 wynosiły 0. Następnie dla pomniejszonego zbioru X ponownie podzieliłem zbiór trenignowy i testowy w proporcji 7:3.

3.2 Model SVM

Jako ostatni zaimplemetowałem model SVM. Miary oraz krzywa ROC dla zbiorów treningowych i testowych przedastawiają się nastepująco:

	precyzja	czułość	dokładność	wartość AUC
treningowy	0.837	0.865	0.784	0.814
testowy	0.806	0.857	0.763	0.815



4 Wnioski

Wybrane przeze mnie modele wypadły dobrze na zbiorach testowych. Niezależnie od wyboru modelu regresji, każdy z nich dawał podobne rezultaty. Można jednak stwierdzić, że najlepszym z nich był ten z regularyzacją ℓ_1, gdyż prowadzi on do pominięcia zmiennych nieistotnych. Co więcej, prawdopodobnie bylibyśmy w stanie uzyskać lepsze wyniki w modelu SVM, używając kroswalidacji oraz znajdując odpowiedni parametr C.