Wydział Elektroniki i Technik Informacyjnych
Politechnika Warszawska

Wstęp do Sztucznej Inteligencji
Semestr 24L
Sprawozdanie z ćwiczenia nr 4

Klasyfikator SVM oraz Drzewa Decyzyjne

Mikołaj Wewiór

Warszawa,
7 V 2024

1. Opis problemu

Celem ćwiczenia było wykorzystanie gotowych implementacji z biblioteki `scikitlearn` klasyfikatora SVM oraz drzewa decyzyjnego do zadania klasyfikacji irysów. Należało zbadać wpływ takich parametrów jak siła regularyzacji, funkcja jądra, liczba iteracji dla SVM oraz kryterium oceny, technika podziału węzła i maksymalna głębokość drzewa w przypadku drzewa decyzyjnego. Do oceny jakości wykoorzystano metryki: *accuracy*, *precision*, *recall* oraz *F1*.

2. Ładowanie zbioru danych

Klayfikowane były kwiaty irysów, mianowicie trzy gatunki - *Setosa*, *Versicolor* i *Virginica*. Dane były w postaci numerycznych parametrów odpowiadającym fizycznym właściwościom kwiatów. Pobrany zbiór danych zawierał 150 próbek. Pobrane dane zostały wymieszane i podzielone na 5 równolicznych podzbiorów. Wykorzystując pięciokrotną walidację krzyżową, cztery z podzbiorów służyły do treningu modelu, a piąty był zbiorem walidacyjnym, z którego ewaluowano model.

W celu powtarzalności przeprowadzanych eksperymentów skorzystano z ziarna generacji liczb pseudolosowych. Wykorzystano je również w dalszej części jako `random_state` w klasyfikatorach. Wykorzystane ziarna to 318 407, 271 102, 231 219.

3. SVM

Do zbadania parametrów SVM wykorzystano kombinacje z wartości parametrów podanych poniżej:

— funkcja jądra: *linear*, *poly*, *rbf*, *sigmoid*;

— siła regularyzacji: *0.5, 1, 2, 5, 10*;

— maksymalna liczba iteracji: *1, 5, 20*.

Co daje w sumie 60 różnych wyników dla każdego ziarna liczb pseudolosowych.

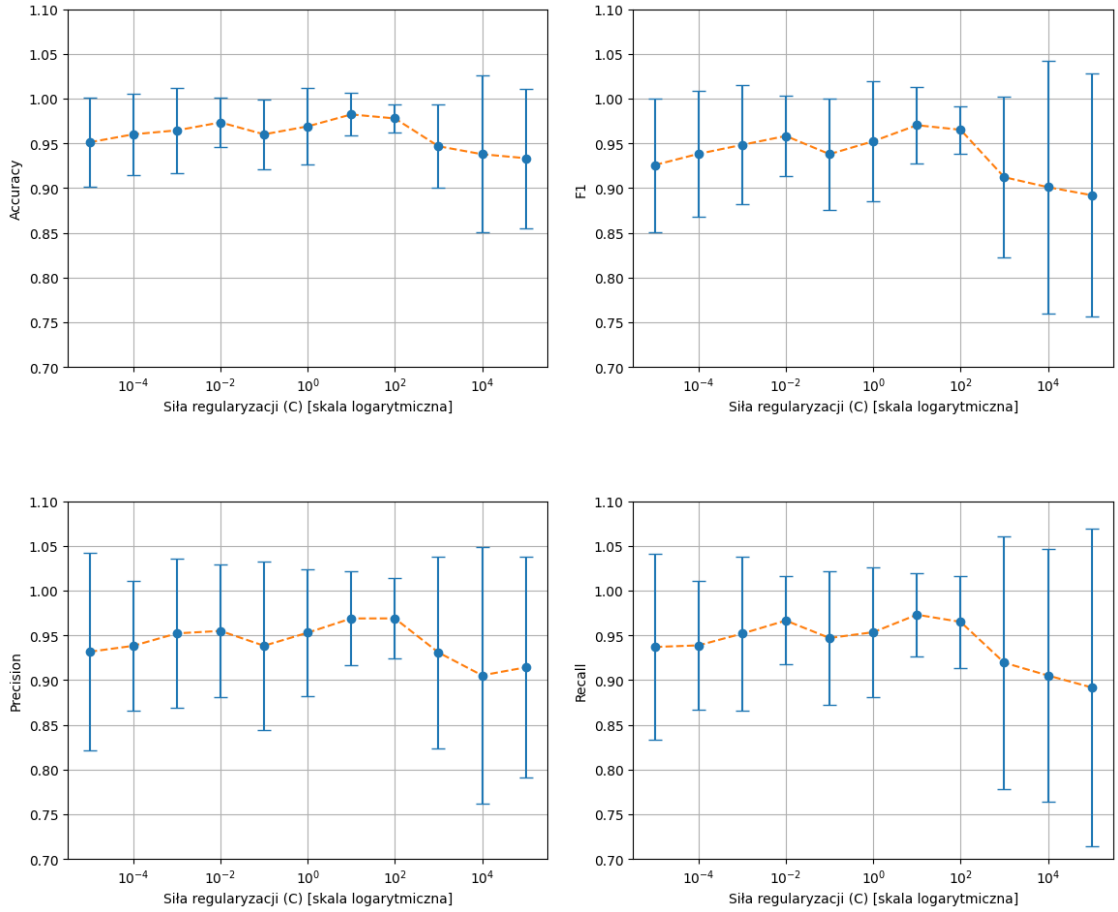
3.1. Porównanie funkcji jądra

Miara Jakości		Linear	Poly	Rbf	Sigmoid
accuracy	średnia	0.9484444	0.9066667	0.968	0.4011852
	odchylenie	0.079186	0.1252356	0.0449219	0.2292846
precision	średnia	0.930995	0.8671106	0.954297	0.0525703
	odchylenie	0.1379298	0.2109123	0.0880823	0.1063037
recall	średnia	0.9223185	0.8635546	0.9517347	0.1026506
	odchylenie	0.1641571	0.2288739	0.1048024	0.2009535
F1	średnia	0.9159746	0.8508525	0.9480961	0.0635927
	odchylenie	0.1417046	0.213157	0.0849382	0.1158556

3.2. Porównanie siły regularyzacji

Miara Jakości		0.5	1	2	5	10
accuracy	średnia	0.8085185	0.8151852	0.8092593	0.7962963	0.8011111
	odchylenie	0.2731418	0.2695291	0.2702549	0.2777531	0.271482
precision	średnia	0.7010338	0.7156826	0.7082498	0.6918149	0.6894348
	odchylenie	0.4026747	0.3973835	0.3991353	0.4039716	0.4081046
recall	średnia	0.710405	0.72433	0.7182087	0.695496	0.7018832
	odchylenie	0.3965225	0.3908119	0.3913128	0.400056	0.3995573
F1	średnia	0.6969442	0.709199	0.7000104	0.6831521	0.6838392
	odchylenie	0.3961845	0.3913926	0.3906589	0.3953592	0.3974142

Przedstawienie wykresów wartości metryk w zależności od siły regularyzacji (pozostałe parametry dobrano tak, aby wyniki były jak najlepsze):



3.3. Porównanie Maksymalnej liczby iteracji

Miara Jakości		1	5	20
accuracy	średnia	0.7866667	0.8077778	0.8237778
	odchylenie	0.2350886	0.2825785	0.2949661
precision	średnia	0.6629305	0.7087749	0.7320243
	odchylenie	0.3863512	0.4039658	0.4133171
recall	średnia	0.6797154	0.713909	0.7365694
	odchylenie	0.3805572	0.3976059	0.4067496
F1	średnia	0.6483343	0.7034449	0.7321077
	odchylenie	0.3694988	0.3983288	0.4094913

4. Drzewa decyzyjne

Do zbadania parametrów drzewa decyzyjnego wykorzystano kombinacje z wartości parametrów podanych poniżej:

— kryterium: *entropy*, *gini*, *log loss*;

— technika podziału: *best*, *random*;

— maksymalna głębokość drzewa: *1, 2, 3, 4, 5*.

Co daje w sumie 30 różnych wyników dla każdego ziarna liczb pseudolosowych.

4.1. Porównanie Kryterium podziału

Miara Jakości		Entropy	Gini	Log loss
accuracy	średnia	0.9129966	0.9138047	0.9113805
	odchylenie	0.1250611	0.1306472	0.1268393
precision	średnia	0.850932	0.8556235	0.8509732
	odchylenie	0.2669234	0.2686448	0.2673233
recall	średnia	0.8793122	0.8811799	0.877294
	odchylenie	0.2486162	0.2522585	0.2504735
F1	średnia	0.8511442	0.85397	0.8495485
	odchylenie	0.2461024	0.250175	0.2477537

4.2. Porównanie techniki podziału

Miara Jakości		Best	Random
accuracy	średnia	0.9262551	0.9043621
	odchylenie	0.1175392	0.1275931
precision	średnia	0.8674996	0.8506597
	odchylenie	0.2528563	0.2629627
recall	średnia	0.8973355	0.8664426
	odchylenie	0.2304646	0.2512992
F1	średnia	0.8736119	0.8400449
	odchylenie	0.2348138	0.2398793

4.3. Porównanie maksymalnej głębokości drzewa

Miara Jakości		1	2	3	4	5
accuracy	średnia	0.7264198	0.9380247	0.9444444	0.9587654	0.9654321
	odchylenie	0.1863709	0.0631365	0.0553329	0.0475243	0.0370535
precision	średnia	0.4684542	0.9229399	0.9272135	0.9433972	0.9490761
	odchylenie	0.4100103	0.1171203	0.1167416	0.0877664	0.0720799
recall	średnia	0.6396065	0.9082575	0.9175046	0.9394712	0.9470236
	odchylenie	0.4656539	0.1435472	0.1329226	0.0940292	0.076798
F1	średnia	0.5082242	0.9027654	0.9105906	0.9368372	0.9457158
	odchylenie	0.397942	0.1027876	0.0973268	0.0720015	0.0611195

5. Podsumowanie

W przypadku klasyfikatora SVM najlepszą kombinacją jego parametrów okazała się funkcja Rbf, z siłą regulacji równą 1 oraz odpowiednio długą liczbą iteracji, co w przypadku klasyfikacji irysów przełożyło się na 20 lub więcej. Zwiększanie liczby iteracji nie przynosiło widocznych efektów - wybór dwudziestu iteracji lub miliona dawały takie same wyniki z dokładnością do 5 liczb po przecinku. Co do jądra funkcji klasyfikatora niewiele gorsza od Rbf okazała się funkcja liniowa a zaraz za nią wielomianowa. Najgorsze wyniki, które okazały się niezadowalające dawała funkcja sigmoidalna. Siła regularyzacji była trudna do oceny, ponieważ z powyższych danych nie widać tego tak dobrze. Jednak przeprowadzając dalsze testy dla pozostałych dobrze dobranych parametrów okazuje się że mniejsza siła regularyzacji wpływa pozytywnie na wyniki.

Klasyfikacja drzewem decyzyjnym najlepsze dawała najlepsze wyniki dla kryterium podziału według wskaźnika Giniego, techniki "best" podziału węzła oraz w przypadku irysów głębokości drzewa wynoszącym 4 lub 5. Tak naprawdę wszystkie kryteria podziału dawały bardzo podobne wyniki - różnica średniej dokładności sięgała niecałe 0,2% oraz około 0,5% w przypadku odchylenia standardowego. Dla dobrze dobranego zestawu parametrów lepiej działało kryterium entropii jednak dla szerokiego spektrum ich kombinacji lepszy okazało się kryterium Giniego. Jeżeli chodzi o porównanie Techniki podziału i głębokości drzewa, to dla techniki best wystarczyło dać 4 - czyli tyle ile parametrów opisujących dane, a w przypadku techniki losowej o jeden więcej. Dla dowolnej głębokości klasyfikator zawsze niemal bezbłędnie klasyfikował gatunek Iris Setosa. Niedokładności pojawiały się przy pozostałych dwóch gatunkach, których parametry były znacznie bardziej do siebie podobne i odpowiednio różne od pierwszego z nich.

W bezpośrednim porównaniu nieco skuteczniejszy okazał się klasyfikator SVM.