Dokumentacja – laboratorium nr 4
Wprowadzenie do Sztucznej Inteligencji

Dominika Wyszyńska 318409

14 grudnia 2023

1 Opis treści zadania

Celem zadania było zaimplementowanie algorytmu SVM. Algorytm ten operuje na zasadzie znajdowania optymalnej hiperpłaszczyzny separującej dane, maksymalizującej odległość między klasami. Zbiorem danych w zadaniu mało być Wine Quality Data. Aby dostosować zbiór danych do problemu klasyfikacji binarnej, należało zmienną objaśnianą (jakość wina) zmodyfikować na podstawie ustalonej wartości minimalnej jakości wina. Zakładamy, że wino o jakości powyżej 6 włącznie jest dobrej jakości.

Należało także zbadać wpływ hiperparametrów na działanie implementowanego algorytmu. Do badań wybrano dwie funkcje jądrowe poznane na wykładzie - wielomianowa oraz RBF.

Implementacja programu została umieszczona na gitlabie wydziałowym.

2 Opis planowanych eksperymentów numerycznych

Eksperymenty zostały zaprojektowane w następujący sposób:

1. Funkcja jądrowa - wielomianowa $f(\mathbf{x}_i, \mathbf{x}_j) = (\mathbf{x}_i^T \cdot \mathbf{x}_j + c)^d$
   - Badanie wpływu parametrów wielomianu - współczynnika c oraz stopnia wielomianu - parametr d
   - Badanie wpływu hiperparametru C
2. Funkcja jądrowa - RBF $f(\mathbf{x}_i, \mathbf{x}_j) = \exp\left(-\gamma \|\mathbf{x}_i - \mathbf{x}_j\|^2\right)$
   - Badanie wpływu parametru γ
   - Badanie wpływu hiperparametru C

3 Opis uzyskanych wyników

3.1 Funkcja jądrowa - wielomianowa

- Badanie wpływu parametrów c oraz d wielomianu w funkcji jądrowej
  Na potrzeby eksperymentów ustawiono wartość hiperparametru C na 1.

Współczynnik c	Potęga wielomianu d	Dokładność modelu
1	1	73.4%
4	1	73.4%
2	2	76.9%
2	3	76.9%
2	4	79.1%
6	4	80.9%
1	6	77.2%
2	6	76.9%
1	10	73.8%

Stopień d wielomianu można wykorzystać do kontrolowania złożoności modelu. Im wyższy stopień, tym bardziej skomplikowana i nieliniowa staje się funkcja jądrowa. Wysoki stopień d zaowocuje bardziej złożonym modelem, który może naddopasować dane. Podczas gdy niski stopień d spowoduje prostszy model, który może niedostatecznie dopasować dane. Parametr c kontroluje, jak bardzo funkcja jądrowa jest nieliniowa. W praktyce, dobór wartości c i d zależy od charakterystyki danych i problemu klasyfikacji. Po wykonaniu wielu eksperymentów najlepsze nastawy - najwyższy procent - dokładność modelu jest przy $c = 6$ oraz $d = 4$. Dokładność modelu wynosi wtedy ok. 80.9%.

- Badanie wpływu hiperparametru C Na potrzeby eksperymentów jako parametry w funkcji jądrowej przyjęto $c = 6$ oraz $d = 4$, czyli z poprzedniego podpunktu, gdy dokładność modelu była największa.

Wartość parametru C	Dokładność modelu
0.001	77.5%
0.1	77.5%
0.5	79.7%
1	80.9%
10	76.3%
100	77.5%
1000	73.4%

Dobór optymalnej wartości parametru C w SVM jest istotny dla uzyskania równowagi między dopasowaniem do danych treningowych a zdolnością modelu do efektywnej generalizacji na nowe dane. Wartość, podobnie jak dla innych parametrów, należy wzynaczać eksperymentalnie. Mała wartość C prowadzi do większego marginesu decyzyjnego, co oznacza, że model jest bardziej tolerancyjny na błędy treningowe i bardziej zgeneralizowany.Natomiast duża wartość C skutkuje niskim błędem treningowym, ponieważ model jest bardziej skłonny do dopasowywania się do każdej pojedynczej próbki z treningowego zbioru danych (może prowadzić do nadmiernego dopasowania). Stąd w podanym przypadku najlepszą wartość dokładności modelu - testowanie na danych testujących - 80.9% otrzymana dla $C = 1$.

3.2 Funkcja jądrowa - RBF (Radial Basis Function)

- Badanie wpływu parametru γ
  Na potrzeby eksperymentów ustawiono wartość hiperparametru C na 1.

Wartość parametru γ	Dokładność modelu
0.1	75.3%
1	75.3%
5	75.9%
10	76.3%
16	80.0%
20	78.4%
100	70.6%

Jak widać, najlepszy procent dokładności modelu otrzymujemy, gdy $\gamma = 16$. Wartość tą należy także dobierać eksperymentalnie. Mała wartość γ prowadzi do bardziej płaskiego modelu, który jest bardziej elastyczny i ma większy margines decyzyjny. Może to prowadzić do gorszego dopasowania do skomplikowanych danych. Natomiast duża wartość γ skutkuje bardziej skomplikowanym modelem, który jest bardziej dopasowany do danych treningowych.

- Badanie wpływu hiperparametru C Na potrzeby eksperymentów jako parametr γ w funkcji jądrowej przyjęto 16, czyli z poprzedniego podpunktu, gdy dokładność modelu była największa.

Wartość parametru C	Dokładność modelu
0.01	64.4%
0.1	70.3%
0.5	78.1%
1	80.0%
10	77.8%
100	76.6%

Wpływ parametru C jest podobny, jak w przypadku funkcji jądrowej - wielomianowej. W przypadku funkcji jądrowej RBF, najlepszym rozwiązaniem jest $C = 1$, gdyż wtedy dokładność modelu jest dokładnie równa 80%, czyli najwięcej z podanych eksperymentów.

4 Podsumowanie i wnioski

Celem tych eksperymentów było zrozumienie wpływu różnych parametrów na zachowanie się algorytmu SVM. Parametr γ w funkcji jądrowej RBF używanej w SVM ma istotny wpływ na kształt decyzyjnej hiperpłaszczyzny. Wartość parametru C wpływa na dopasowanie. Wysoka wartość C - mniejszy margines, nadmierne dopasowanie, niska wartość C - większy margines, lepsza generalizacja. Ogólnie obie funkcje jądrowe dość dobrze sobie poradziły. Przy najlepszych doborach parametrów - dokładność modelu znajduje się w okolicach 80%, co jest dość zadawalającym wynikiem.