Podstawy uczenia maszynowego

23.05.2024

Laboratorium 6

Selekcja cech

Łukasz Stępień, Kacper Fus

1. Cel zadania

Celem tego laboratorium było zapoznanie się z algorytmami selekcji cech przy użyciu zbioru danych o ostrym białaczce (leukemia1). Zbiór ten zawiera informacje dotyczące 72 pacjentów, a każdy pacjent jest opisany przez 7070 cech. Zadaniem było przewidzenie, czy dana osoba cierpi na ostrą białaczkę.

2. Implementacja

Najpierw wczytaliśmy zbiór danych za pomocą funkcji scipy.io.loadmat. Następnie usunęliśmy cechy o niskiej wariancji, stosując metodę VarianceThreshold. Porównaliśmy skuteczność predykcji na oryginalnym zbiorze danych i po usunięciu cech o niskiej wariancji. Z powstałego zbioru cech wybraliśmy m najlepszych cech, korzystając z algorytmu rekursywnej eliminacji cech (RFE) dla regresji logistycznej i lasów losowych. Przeanalizowaliśmy wpływ dwóch metryk - dokładności i metryki AUC - na dokładność klasyfikacji przy użyciu regresji logistycznej i lasów losowych. Wykorzystaliśmy cztero- lub sześciokrotną walidację krzyżową, aby uzyskać wiarygodne wyniki. W ostatnim kroku porównaliśmy skuteczność naszego podejścia z wbudowanymi metodami selekcji cech.

3. Wyniki

- Po usunięciu cech o niskiej wariancji:
    - Regresja Logistyczna:
        a) Średnia dokładność: 0.9199
        b) Średnia metryka AUC: 0.9929

    - Lasy Losowe:
        a) Średnia dokładność: 0.9391
        b) Średnia metryka AUC: 0.9857

- Porównanie ze wbudowanymi metodami selekcji cech:
    - Regresja Logistyczna (L1):
        a) Średnia dokładność: 0.9599
        b) Średnia metryka AUC: 0.9938

    - Lasy Losowe (Feature Importances):
        a) Średnia dokładność: 0.9792
        b) Średnia metryka AUC: 1.0000

4. Wnioski

W naszej analizie stwierdziliśmy, że wbudowane metody selekcji cech, w szczególności regularyzacja L1 dla regresji logistycznej i ważność cech dla lasów losowych, osiągnęły lepsze wyniki w porównaniu z podejściem opartym na RFE. Wbudowane metody selekcji cech mogą uwzględniać zależności między cechami i wybierać te, które mają największy wpływ na model klasyfikacyjny. Regularyzacja L1 dla regresji logistycznej pozwala na wyzerowanie współczynników nieistotnych cech, co prowadzi do bardziej skondensowanego i bardziej efektywnego zestawu cech. Dodatkowo wbudowane metody selekcji cech, zwłaszcza regularyzacja L1, pomagają w redukcji nadmiernego dopasowania poprzez regularyzację współczynników cech. Dzięki temu modele są bardziej odporne na szum w danych i mają lepszą zdolność do generalizacji na nowe dane. Warto też zwrócić uwagę, że wbudowane metody selekcji cech korzystają z domyślnych parametrów, które zostały zoptymalizowane dla danego algorytmu klasyfikacji. Dzięki temu mogą działać bardziej efektywnie w wielu przypadkach niż podejścia, które wymagają ręcznego dostrajania parametrów.