Wprowadzenie do uczenia maszynowego

Lista 4

Janusz Szwabiński

1. **Wstęp do drzew decyzyjnych:** Wczytaj zbiór danych `Iris` i zwizualizuj go. Zbuduj model drzewa decyzyjnego, używając klasy `DecisionTreeClassifier` z biblioteki `scikit-learn`. Przedstaw strukturę drzewa na wykresie za pomocą funkcji `plot_tree` lub `export_graphviz` i zinterpretuj warunki podziału.

2. **Problem przeuczenia (overfitting):** Wytrenuj dwa modele drzewa decyzyjnego na danych `Iris`: jeden bez ograniczeń (domyślne parametry) i drugi z ograniczoną maksymalną głębokością (`max_depth`). Porównaj ich dokładność na zbiorze treningowym i testowym. Wyjaśnij zjawisko przeuczenia.

3. **Lasy losowe (Random Forest):** Zbuduj model lasu losowego, używając klasy `RandomForestClassifier`. Porównaj jego dokładność i stabilność z modelem pojedynczego drzewa decyzyjnego. Wyjaśnij, w jaki sposób losowość wpływa na poprawę wyników.

4. **Znaczenie cech (Feature Importance):** Dla wytrenowanego modelu lasu losowego, uzyskaj informację o ważności poszczególnych cech. Zwizualizuj ważność cech na wykresie słupkowym i zinterpretuj, które zmienne mają największy wpływ na klasyfikację.

5. **Zastosowanie AdaBoost:** Użyj klasy `AdaBoostClassifier` z biblioteki `scikit-learn`. Wytrenuj model z różną liczbą estymatorów (`n_estimators`) i przeanalizuj, jak wpływa to na wydajność.

6. **Wprowadzenie do Gradient Boosting:** Wytrenuj model wzmacniania gradientowego (`GradientBoostingClassifier`) i porównaj jego wyniki z modelem lasu losowego. Omów, jaka jest główna różnica w sposobie budowania modelu przez te dwa algorytmy.

7. **Zaawansowane biblioteki - XGBoost:** Zainstaluj bibliotekę `xgboost` i użyj jej do wytrenowania modelu na zbiorze danych do klasyfikacji. Porównaj jego dokładność i czas potrzebny na wyuczenie z modelami z `scikit-learn`. Wyjaśnij, dlaczego XGBoost jest często używany w konkursach Data Science.