Projekt 3.

Modele regresyjne z biblioteki scikit-learn. Trenowanie modeli regresyjnych dla wybranego zbioru danych. Dobór parametrów modelu. Ewaluacja modeli regresyjnych.

Paweł Majewski, Kacper Marciniak

26.08.2024

1 Zadanie

Korzystając ze zbioru danych **california_housing**, wytrenuj dwa modele regresyjne: model regresji liniowej oraz drzewo decyzyjne. Przeprowadź ewaluację, wykorzystując podział zbioru danych na zbiory treningowy, walidacyjny oraz testowy. Dla wybranych parametrów modeli przeprowadź strojenie parametrów z GridSearch. Sugerowane kolejne etapy rozwiązania:

1. wczytaj dane,
2. zidentyfikuj kolumnę wskazującą na wartości docelowe dla regresji,
3. sprawdź czy zbiór danych zawiera brakujące wartości,
4. * przeprowadź **EDA (exploratory data analysis)** - pokaż zmienność wartości zmiennych niezależnych na określonym wykresie,
5. podziel zbiór danych na zbiory treningowy, walidacyjny oraz testowy według stosunku 60%, 20% oraz 20%.
6. przeprowadź skalowanie wartości zmiennych niezależnych, zachowując niezależność pomiędzy zbiorami,
7. wytrenuj modele regresyjne i przeprowadź ewaluację na zbiorze testowym dla domyślnych hiperparametrów modeli, wykorzystując metrykę **RMSE**,
8. przeprowadź strojenie wartości hiperparametrów dla modelu drzewa decyzyjnego na zbiorze walidacyjnym, uwzględniając **maksymalną głębokość** drzewa decyzyjnego,
9. * umieść kolejne etapy przetwarzania we funkcjach oraz opracuj jedną klasę łączącą wszystkie etapy przetwarzania,
10. * przeprowadź selekcję cech, wykorzystując metodę **LASSO**.

* - dla chętnych (na ocenę celującą)

Sprawozdanie powinno zawierać kod źródłowy. Kod źródłowy może być również udostępniony na Github. W kodzie źródłowym należy wskazać na funkcje lub sekcje, związane z określonymi etapami rozwiązaniami. Sprawozdanie nie musi zawierać wprowadzenia teoretycznego.

2 Pytania kontrolne

1. Wymień przynajmniej trzy modele regresyjne, rozwiń skróty.
2. Napisz wzory dla metryk MAE oraz RMSE wraz z objaśnieniami zmiennych. Która z tych zmiennych jest bardziej czuła na wartości odstające?
3. Wyjaśnij rolę zbioru treningowego, walidacyjnego oraz testowego.
4. Czym jest walidacja krzyżowa? Wymień główną zaletę walidacji krzyżowej, względem arbitralnego podziału na zbiór treningowy, walidacyjny oraz testowy.
5. Narysuj przykładowe drzewo decyzyjne dla regresji o głębokości 3 i z maksymalną liczbą węzłów decyzyjnych. Przyjmij dowolne wartości cech we węzłach.
6. Dlaczego predykcje modelu drzewa decyzyjnego dla regresji mają wartości dyskretne a dla modelu regresji liniowej mają wartości ciągłe.
7. Podaj przykładowy problem regresji oraz wskaż zmienną zależną i przynajmniej trzy zmienne niezależne.
8. Czym jest regresja wielomianowa. Podaj przykład hipotezy regresji wielomianowej dla stopnia wielomianu dwa.
9. Co to jest współczynnik determinacji R^2 i jak go interpretować?
10. Jakie metryki stosuje się do oceny jakości modelu regresji? Wymień przynajmniej trzy.
11. Jakie techniki regularizacji można zastosować w regresji? Wymień przynajmniej dwa.
12. Czym jest strojenie (tuning) hiperparametrów? Podaj przykładowe hiperparametry (przynajmniej dwa) dla wybranego modelu regresyjnego.
13. Dlaczego usuwanie wartości odstających może być ważne w modelach regresyjnych?
14. Co to jest normalizacja i standaryzacja danych?
15. Wymień trzy wybrane założenia modelu regresji liniowej.