Projekt 4.

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

Paweł Majewski, Kacper Marciniak

26.08.2024

1 Zadanie

Korzystając ze zbioru danych **breast_cancer**, wytrenuj trzy modele klasyfikacyjne: model regresji logistycznej, SVM oraz KNN. Przeprowadź ewaluację z wykorzystaniem walidacji krzyżowej. Dla wybranych parametrów modeli przeprowadź strojenie parametrów z GridSearch. Sugerowane kolejne etapy rozwiązania:

- wczytaj dane,
- zidentyfikuj kolumnę wskazującą na klasę oraz liczbę klas,
- sprawdź czy zbiór danych zawiera brakujące wartości,
- sprawdź czy zbiór danych jest zbalansowany,
- * przeprowadź EDA (exploratory data analysis) - pokaż zmienność wartości cech na określonym wykresie z wyszczególnieniem klas np. histogram, wykres pudełkowy,
- przeprowadź walidację krzyżową dla domyślnych parametrów modeli,
- przeprowadź **GridSearch** dla **SVM** oraz **KNN** (walidacja krzyżowa dla kolejnych zestawów wartości parametrów), w przypadku KNN weź pod uwagę:
  1. liczbę sąsiadów
  2. rodzaj metryki odległości,

  w przypadku SVM:
  1. rodzaj jądra przekształcenia,
  2. wartość marginesu,
- * umieść kolejne etapy przetwarzania we funkcjach oraz opracuj jedną klasę łączącą wszystkie etapy przetwarzania.

* - dla chętnych (na ocenę celującą)

Sprawozdanie powinno zawierać kod źródłowy. Kod źródłowy może być różnież 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 klasyfikacyjne, rozwiń skróty.
2. Objaśnij metryki precision, recall oraz F1-score na postawie wzorów, bazujących na liczbie określonych predykcji (TP, TN, FP, FN).
3. Wyjaśnij problem z metryką accuracy w przypadku danych niezbalansowanych (może być na przykładzie).
4. Dlaczego model KNN wymaga skalowania cech przed użyciem ich do trenowania modelu? Wyjaśnij, odwołując się do określonej metryki odległości w przestrzeni cech (np. euklidesowej).
5. Wymień podstawową różnicę pomiędzy klasyfikacją i regresją. Podaj po dwa przykłady problemów dla klasyfikacji oraz regresji.
6. Czym jest GridSearch w kontekście strojenia parametrów modelu?
7. Podaj dwa hiperparametry, które możemy dostroić dla modelu KNN?
8. Podaj dwa hiperparametry, które możemy dostroić dla modelu SVM?
9. Dlaczego liczba sąsiadów dla modelu KNN powinna być nieparzysta?
10. Dlaczego model KNN nazywa się leniwym ("lazy learner")? Na czym polega trening modelu KNN?
11. Wskaż powiązanie pomiędzy modelem regresji liniowej a modelem regresji logistycznej, nawiązując to pojęcia funkcji sigmoidalnej.
12. Jak możemy estymować prawdopodobieństwo predykcji w przypadku modelu regresji logistycznej.
13. Omów krótko problem niezbalansowania danych w kontekście klasyfikacji oraz wymień dwie strategie radzenia sobie z tym problemem.
14. Omów krótko problem brakujących danych (ang. missing data) oraz wymień dwie strategie radzenia sobie z tym problemem.
15. Narysuj przykładową macierz pomyłek dla problemu klasyfikacji binarnej. Zaznacz liczbę predykcji TP, TN, FP oraz FN.