Klasyfikacja

Uczenie Maszynowe
uczeniemaszynowe@cs.uni.wroc.pl

04.11.2025, 10:00
pracowania

Zadanie 1. [1 pkt]

Powtórz zadanie 3 z listy 2 dla modelu kNN. Użyj danych `advertising.csv`. Zbuduj modele kNN na danych, w których skala zmiennych jest zachowana i na danych znormalizowanych. Porównaj te dwa modele, zapisz obserwacje i wnioski.

Zadanie 2. [2 pkt]

- Zbuduj model regresji logistycznej na danych `bank.csv`[1]. Porównaj wyniki dla progów `[0.1, 0.25, 0.5, 0.75, 0.9]`.
- Dla wybranego progu zaznacz obszary, które klasyfikowane są do obu klas. Co charakterystycznego dostrzegasz? Jak granica między klasami ma się do wektora parametrów modelu?
- Następnie wylicz TPR i FPR dla progów `[0, 0.01, 0.02, ..., 0.98, 0.99, 1]` i wyrysuj krzywą ROC.
- Wybierz optymalny próg zakładając, że koszt błędu pierwszego rodzaju (False Positive) wynosi 10, a koszt błęgu drugiego rodzaju (False Negative) wynosi 3.

Zadanie 3. [2 pkt]

W tym zadaniu przyjrzymy się, jak podobieństwa między produktami są kodowane w zanurzeniach produktów uzyskanych z systemu rekomendacyjnego.

W pliku `embeddings.pkl` znajduje się macierz zanurzeń produktów, gdzie każdy wiersz macierzy odpowiada jednemu produktowi. Zanurzenia te uzyskano z systemu rekomendacyjnego[2]. W pliku `item_list.txt` znajdują się oryginalne identyfikatory produktów. W pliku `meta_Books.json.gz`[3] znajdują się metadane produktów – nasze embeddingi dotyczą produktów z kategorii `Books`. Identyfikator produktu znajduje się w polu `asin`.

Podziel dane na treningowe i testowe; w razie potrzeby możesz też wydzielić dodatkowy zbiór walidacyjny. Użyj modelu kNN do klasyfikacji kategorii i podkategorii produktów, które znajdziesz w metadanych. Sprawdź, jaka metryka jest odpowiednia dla tego zadania.

[1] https://www.kaggle.com/datasets/janiobachmann/bank-marketing-dataset/data
[2] https://github.com/kuandeng/LightGCN
[3] informacje i dodatkowe instrukcje: https://cseweb.ucsd.edu/ jmcauley/datasets/amazon/links.html

Wskazówki:

- Zwróć uwagę na różne formaty danych wejściowych. Mamy plik zapisany binarnie z wykorzystaniem biblioteki `pickle`, plik tesktowy oraz skompresowany plik w którym każda linijka jest `json`em.
- Informacje o metrykach dostępnych w pakiecie sklearn: https://docs.scipy.org/doc/scipy/reference/spatial.distance.html.
- Zwróć uwagę czy produkt ma jedną czy więcej klas, zastanów się jak należałoby taką sytuację obsłużyć. Być może będzie Ci wygodnie zaproponować własną implementację modelu.

Zadanie 4. [2.5 pkt]

Zbuduj model regresji liniowej na danych `adult.csv`[4] przewidujący, czy dana osiąga przychody powyżej $50k. Skorzystaj z biblioteki `statsmodels`[5]. Przeprowadź procedurę wyboru zmiennych używając następującej procedury:

- Zbuduj model bazowy.
- Sprawdź rezultaty włączenia każdej z nieużytych zmiennych.
- Wytypuj zmienną, ktora daje największy wzrost jakości modelu.
- Za pomocą wybranego kryterium informacyjnego (AIC, BIC) sprawdź, czy ma sens włączenie tej zmiennej do modelu.
- Jeżeli tak, włącz zmienną do modelu i wytypuj kolejny najlepszy predykator. W przeciwnym wypadku zakończ nie włączaj wytypowanej zmiennej, zakończ procedurę.

Przeprowadź ewaluację uzyskanego modelu.
Wskazówki:

- Możesz pobrać dane bezpośrednio z Kaggle'a:

```
import kagglehub

# Download latest version
path = kagglehub.dataset_download("uciml/adult-census-income")

print("Path to dataset files:", path)
```

- Przydatne metody i atrybuty: `model.summary()`, `model.aic`, `model.bic`.
- Aby dopasować intercept użyj `statsmodels.tools.add_constant`.

[4] https://www.kaggle.com/datasets/uciml/adult-census-income
[5] https://www.statsmodels.org/stable/generated/statsmodels.discrete.discrete_model.Logit.html

Zadanie 5. [2.5 pkt]

Zapoznaj się z danymi `Titanic Dataset.csv`[6]. Uzupełnij puste wartości na dwa sposoby, zgodnie z instrukcją poniżej. Następnie porównaj modele regresji liniowej przewidujące zmienną `survived`.

- Sprawdź jaki odestek brakujących wartości w kolejnych kolumnach.
- Sprawdź czy istnieje zależność między zmiennymi czynnikowymi a brakującymi wartościami w kolumnach numerycznych?
- Rozwiązanie baselinowe:
  - usuń kolumny o wysokim udziale pustych wartości,
  - uzupełnij brakujące wartości numeryczne średnią danej cechy,
  - w przypadku zmiennych kategorycznych najczęstszą wartością.
- Rozwiązanie zaawansowane:
  - zmienną `Cabin` zamień na zmienną kategoryczną informującą czy wartość była pusta,
  - wyekstrahuj tytuły ("Mr", "Mrs", ...) z imion pasażerów, użyj ich do uzupełniania wieku pasażerów uzupelniając wiek medianą dla osób o danym tytule,
  - pozostałe jak wyżej.

Dyskusja na pracownię

- Jakie są inne możliwe procedury wyboru zmiennych do modelu?
- Przeanalizuj sens metod uzupełaniania brakujących wartości zaproponowanych dla kolejnych kolumn.

[6] https://www.kaggle.com/code/sakshisatre/titanic-s-missing-data-visualizing-null-values/input