Wprowadzenie do uczenia maszynowego

Lista 8

Janusz Szwabiński

1. **Wczytanie i eksploracja danych:** Wczytaj publicznie dostępny zbiór danych do analizy wydźwięku (np. zbiór recenzji filmowych z `IMDB`). Zapoznaj się z danymi: sprawdź liczbę recenzji, rozkład klas, i długość tekstu.

2. **Przygotowanie danych tekstowych:** Przeprowadź pre-processing danych tekstowych, który powinien obejmować co najmniej:

   - usunięcie znaków interpunkcyjnych i cyfr,
   - konwersję tekstu na małe litery,
   - usunięcie stopwords (słów-wypełniaczy).

3. **Wektoryzacja tekstu:** Przekształć dane tekstowe na format numeryczny, który może być użyty przez model uczenia maszynowego. Zastosuj jedną z następujących technik:

   - **Bag-of-Words (Bag of words):** Użyj `CountVectorizer` lub `TfidfVectorizer` z `scikit-learn`.
   - **wektoryzacja słów (Word Embeddings):** Użyj gotowych wytrenowanych wektorów (np. `GloVe`) lub wytrenuj własne, jeśli czas na to pozwala.

4. **Budowa modelu klasyfikacji:** Wytrenuj klasyfikator na przygotowanych danych. Możesz użyć:

   - Klasycznego modelu ML (`LogisticRegression`, `Naive Bayes`) lub,
   - Prostej sieci neuronowej z listy 7 (dla wektoryzacji Bag-of-Words).

5. **Ocena i interpretacja wyników:** Oceń wydajność wytrenowanego modelu na zbiorze testowym, używając metryk takich jak dokładność, precyzja i czułość. Dodatkowo, przeanalizuj macierz pomyłek.

6. **Predykcja na nowych danych:** Stwórz kilka własnych, krótkich recenzji (jedną pozytywną, jedną negatywną, jedną neutralną). Przetwórz je w taki sam sposób jak dane treningowe i użyj wytrenowanego modelu do predykcji ich wydźwięku. Sprawdź, czy model poprawnie je sklasyfikował.