MAT4-Projekt-2425

jm

2024-12-08

Klasyfikacja Bayesowska

Problem

Wiadomość tekstowa **message** została zakodowana w jednym z trzech języków:

- *wakandyjskim(W)*,
- *latveriańskim(L)*,
- *symkariańskim(S)*.

Rozszyfrowanie wiadomości wymaga wskazania z jakiego języka ona pochodzi. Zaproponuj rozwiązanie problemu ustalenia języka, w którym napisano wiadomość, w oparciu o aktualizacje Bayesowskie (ang. Bayesian updating), na podstawie kolejnych symboli występujących w wiadomości **message**.

Rozwiąż powyższy problem w następujących sytuacjach:

I. Przyjmij, że przed odebraniem wiadomości zakładano, że może pochodzić z każdego z trzech języków W, L, S z jednakowym prawdopodobieństwem.

II. Przyjmij, że przed odebraniem wiadomości zakładano, że komunikaty tekstowe Latverianie przesyłają trzykrotnie częściej niż mieszkańcy pozostałych dwóch krain. (Wykonaj to samo dla Wakandian i Symkarian).

Zasobami, którymi dysponujesz są teksty źródłowe:
dwak (j.wakandyjski), **dlatver** (j.latveriański), **dsymk** (j.symkariański). Na ich podstawie możesz ocenić częstość występowania poszczególnych znaków w danym języku.

Podstawy teoretyczne

Zdefiniuj powyższy problem w języku klasyfikacji Bayesowskiej.

- Zdefiniuj badane hipotezy.
- Wyjaśnij pojęcie prawdopodobieństwa a priori i a posteriori.
- Zdefiniuj funkcję wiarogodności, jaka jest jej interpretacja?

Rozwiązanie problemu - prezentacja wyniku

1. W każdym z przypadków I i II zaprezentuj zmianę prawdopodobieństw *a priori*/*a posteriori* jaka następuje po kolejno ana-li-zo-wa-nych symbolach wiadomości **message**. Przykładowe możliwości prezentacji:

- tabela dla kolejnych prawdopodobieństw *a posteriori*
- wykres zmian prawdopodobieństw *a posteriori* (dla każdego z języków) w zależności od liczby prze-ana-li-zo-wa-nych kolejno symboli wiadomości **message**
- "stacked bar plot" - wykres słupkowy przedstawiający rozkład prawdopodobieństw *a posterioi* po każdej aktualizacji
- dodatkowe własne propozycje wizualizacji prawdopodobieństw zmieniających się wraz z kolenymi aktualizacjami

2. Czy przyjęte początkowo rozkłady *a priori* mają wpływ na ostateczną klasyfikację?

- Porównaj wyniki uzyskane w przypadkach I i II

3. Zaproponuj i uzasadnij sensowną metodę stopu umożliwiającą zakończenie procedury aktualzacji bez czytania całej wiadomości **message**.
4. Spośród liter alfabetu $\{A, B, C, D, E, F\}$ wybierz dwie, a następnie tylko dwie litery pozostaw na ich miejscach w otrzymanej wiadomości **message**, a pozostałe symbole zastąp symbolem N - oznaczającym dowolną z pozostałych liter. Dla tak zmienionej wiadomości przeprowadź procedurę aktualizacji Bayesowskiej. Opisz uzyskane wnioski i spostrzeżenia.

Uwaga

Zasadniczą częścią zaliczenie projektu, które polegać będzie na indywidualnej rozmowie z wykładowcą, będzie sprawdzenie zrozumienia teoretycznych aspektów opisywanych problemów (a nie tylko kwestie implementacyjne).

Literatura:

1. https://ocw.mit.edu/courses/18-05-introduction-to-probability-and-statistics-spring-2022/resources/mit18_05_s22_class10-prep-a_pdf
2. https://ocw.mit.edu/courses/18-05-introduction-to-probability-and-statistics-spring-2022/resources/mit18_05_s22_class10-prep-b_pdf/
3. https://ocw.mit.edu/courses/18-05-introduction-to-probability-and-statistics-spring-2022/resources/mit18_05_s22_class11-prep_pdf/df
4. Bayesian Data Analysis (Rozdział 1), http://www.stat.columbia.edu/~gelman/book/BDA3.pdf
5. Oliver Dürr, Beate Sick, Elvis Murina, "Probabilistic Deep Learning With Python, Keras and TensorFlow Probability", ISBN 9781617296079, Manning.