

# Eksploracja danych

## Laboratorium 5 – Ukryte Modele Markova

Marcin Pietrzykowski

ver. 1.00

### 1 Cel

Celem laboratorium jest zapoznanie się z Ukrytymi Modelami Markova.

### 2 Zadanie

1. Pobrać pakiet *mchmm* <https://github.com/maximtrp/mchmm> implementujący Modele Markova (*pip3 install mchmm*).
2. Zapoznać się z klasą `HiddenMarkovModel`.
3. Wczytać dane z pliku *szczecin\_sep\_2009.txt* pochodzące ze strony *Weather Underground*, zawierające historyczne dane pogodowe dla Szczecina.
4. Zadaniem będzie opracowanie Ukrytego Modelu Markowa traktując wrześnieową pogodę w Szczecinie jako źródło Markowowskie.
5. Zbiór stanów w modelu będzie podyktowany przez wartości zmiennej pogodowej "Condition" pogrupowane do takich możliwości: {Clear, Fog, Light Rain, Rain, Cloudy, Unknown}. Zbiór obserwacji będzie wyznaczony przez produkt kartezjański wybranych zmiennych: temperatura (kolumna 2), wilgotność (kolumna 4), ciśnienie (kolumna 5), prędkość wiatru (po dyskretyzacji) (kolumna 8). Przykładem może być czwórka (4, 6, 4, 7) jako możliwa pojedyncza obserwacja, którą mógł wyprodukować np. stan "Light Rain".
6. Przy użyciu algorytmu Viterbiego należy odkryć ukrytą ścieżkę stanów pogodowych ("Condition") mając daną sekwencję obserwacji.
7. Dane należy podzielić na sekwencje uczące i testujące np. w proporcji 70:30. Długość pojedynczej sekwencji można ustalić np. na  $T = 12, 24, 48$  (zważywszy na fakt, że notowania są co pół godziny).

8. Nastroić model Markova za pomocą reestymacji Bauma-Welcha. Przed uczeniem należy wstępnie nastroić macierz  $A$  na podstawie zmiennej pogodowej "Condition".
9. Przetestować gotowy model dla sekwencji testowych, tj. odkrywać ukryte ścieżki stanów pogodywch na podstawie sekwencji obserwacji. Przedstawić zbiorcze procentowe wyniki poprawności (procent poprawnych stanów, procent poprawnych całych ścieżek, średnia poprawna długość podciągu).
10. Sprawdzić także, czy model po nauczaniu lepiej naśladuje zjawisko niż początkowy model losowy

### 3 Zadanie dodatkowe

1. Pobrać dane dla dwóch dodatkowych lokalizacji ze strony (<https://www.wunderground.com/history>).
2. Dane należy podzielić na sekwencje uczące i testujące np. w proporcji 70:30.
3. Nastroić 3 różne Ukryte Modele Markova każdy dla innej lokalizacji.
4. Dokonać klasyfikacji danych testowych przy użyciu algorytmu forward-backward (klasyfikować z którego miasta pochodzą dane pogodowe.). Przedstawić zbiorcze procentowe wyniki poprawności