Klasyfikacja urządzeń na podstawie poboru prądu

Kamila Ciężabka

Kwiecień 2024

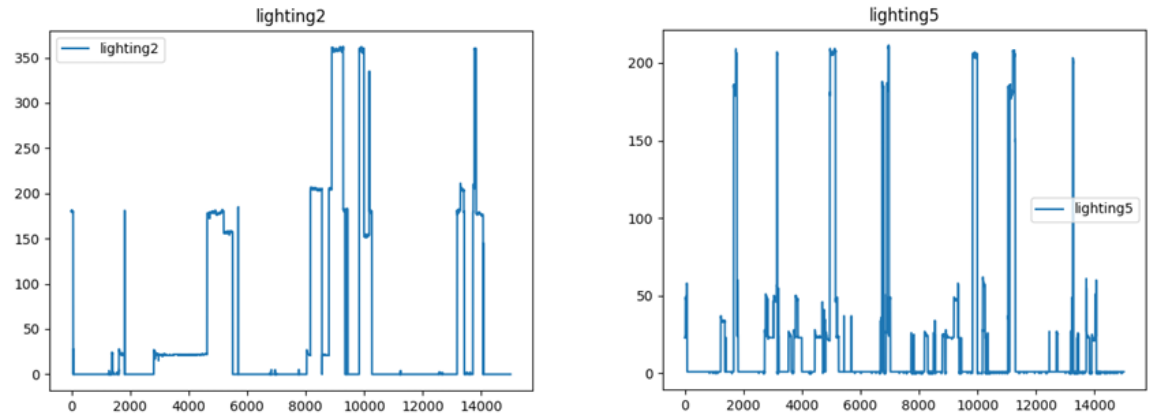
Spis treści

1 Wstęp

Celem niniejszego projektu jest klasyfikacja urządzeń domowych na podstawie danych na temat ich poboru prądu.

Dane, które będziemy analizować, pochodzą z pliku `house3_5devices_train.csv`, zawierającego pomiary poboru mocy dla pięciu urządzeń (`lighting2, lighting5, lighting4, refrigerator, microwave`) w pewnych odstępach czasowych. Pierwsza kolumna zbioru danych to timestamp, z którego można odczytać dokładną datę i godzinę - zmienną tę zignorujemy go w procesie uczenia i testowania, a kolejne wiersze potraktujemy jako kolejne punkty czasowe. Pobór mocy każdego z urządzeń podany jest w pięciu kolumnach.

Przykładowe wykresy emisji prądu dla dwóch pierwszych urządzeń prezentują się następująco:



Zadaniem jest odpowiednia klasyfikacja urządzenia, dla nowych danych bądź wykresów poboru prądu.

Projekt został wykonany przy użyciu ukrytych modeli Markowa (Hidden Markov Models) z wykorzystaniem wbudowanej biblioteki `HMMlearn`. Zadaniem jest stworzenie modelu HMM dla każdego urządzenia, a następnie użycie tych modeli do klasyfikacji nowych pomiarów.

2 Ukryte modele Markowa

2.1 Schemat działania

Ukryte modele Markowa można ogólnie opisać jako pewien losowy spacer po grafie, którego liczba wierzchołków stanowi hiperparametr, który należy ustalić. Proces ten rozpoczyna się w wierzchołku początkowym, a następnie zgodnie z ustalonymi parametrami, na każdym wierzchołku emitowany jest sygnał odpowiadający ilości zużytej energii. Emisja energii pochodzi z pewnego jednowymiarowego rozkładu Gaussowskiego, który jest przypisany do każdego wierzchołka.

W kontekście naszego projektu, stworzyliśmy pięć różnych modeli HMM, z których każdy nauczył się charakterystycznych wzorców emisji sygnałów dla konkretnego urządzenia. Każdy model jest w stanie generować sekwencje danych odpowiadające zużyciu energii przez dane urządzenie.

Do trenowania HMM używany jest algorytm Baum-Welch (BW), którego celem jest optymalizacja wag i parametrów w taki sposób, aby prawdopodobieństwo otrzymania wykresu było jak najbardziej maksymalne.

Gdy dysponujemy wykresem poboru prądu pochodzącym od nieznanego urządzenia, możemy obliczyć prawdopodobieństwo, z jakim każdy z pięciu modeli HMM mógłby wygenerować taką sekwencję. Model, który daje najwyższe prawdopodobieństwo uzyskania obserwowanego kształtu wykresu, jest przypisywany do danego urządzenia.

2.2 Algorytm Baum-Welch

Algorytm Baum-Welch, będący szczególnym przypadkiem algorytmu EM (ang. Expectation-Maximization), iteracyjnie maksymalizuje prawdopodobieństwo zaobserwowanych danych poprzez naprzemienne wykonywanie kroków E i M, aż do osiągnięcia zbieżności lub maksymalnej liczby iteracji (`n_iter`). Każda iteracja algorytmu prowadzi do poprawy parametrów modelu, zbliżając je do wartości optymalnych, które najlepiej opisują zaobserwowane dane.

Etap oczekiwania (E-step) oblicza prawdopodobieństwa przebywania w poszczególnych stanach oraz przejść między stanami, natomiast etap maksymalizacji (M-step) aktualizuje parametry modelu na podstawie tych prawdopodobieństw. W ten sposób algorytm dostosowuje model, aby lepiej pasował do danych, co prowadzi do bardziej dokładnych i wiarygodnych predykcji.

W efekcie, algorytm Baum-Welch jest kluczowym elementem trenowania ukrytych modeli Markowa, umożliwiającym efektywne modelowanie sekwencji danych z ukrytymi stanami.

Jego celem jest oszacowanie parametrów modelu, takich jak macierz przejść między stanami oraz parametry emisji, które maksymalizują prawdopodobieństwo zaobserwowanych danych.

Algorytm Baum-Welch składa się z dwóch głównych etapów, które są iteracyjnie powtarzane: etap oczekiwania (E-step) oraz etap maksymalizacji (M-step).

Etap oczekiwania (E-step): Na tym etapie, dla danych wejściowych, obliczane są wartości oczekiwane zmiennych ukrytych. Główne obliczenia obejmują:

- Forward algorithm - obliczanie prawdopodobieństwa dotarcia do danego stanu w danym czasie, zaczynając od stanu początkowego.
- Backward algorithm - obliczanie prawdopodobieństwa dotarcia do końcowego stanu, zaczynając od danego stanu w danym czasie.
- γ - prawdopodobieństwo przebywania w danym stanie w danym czasie, biorąc pod uwagę całą obserwację.

- χ - prawdopodobieństwo przejścia z jednego stanu do drugiego w danym czasie, biorąc pod uwagę całą obserwację.

Etap maksymalizacji (M-step): Na tym etapie aktualizowane są parametry modelu, aby maksymalizować wartości oczekiwane obliczone w etapie E. Aktualizacje obejmują:

- Macierz przejść między stanami - aktualizacja prawdopodobieństw przejść między stanami, na podstawie wartości ξ.
- Parametry emisji - aktualizacja parametrów emisji (średnie i wariancje rozkładów Gaussa) na podstawie wartości γ.

3 Budowa modeli

3.1 Opis funkcji

Do realizacji tego celu, dla każdego z pięciu urządzeń wytrenowano osobny model HMM korzystając z funkcji `GaussianHMM` z biblioteki `HMMlearn`. Funkcja ta używana jest do tworzenia i trenowania ukrytych modeli Markowa z rozkładami normalnymi dla emisji.

Kluczowe argumenty przyjmowane przez funkcję to między innymi:

- `n_components` - liczba stanów ukrytych w modelu HMM. Jest to podstawowy parametr określający złożoność modelu. Liczbę ukrytych stanów dobrano eksperymentalnie - przetestowane zostały różne różne wartości od 3 do 15,
- `covariance_type` - zmienna określająca, jak modeluje się wariancję i kowariancję między różnymi cechami w danych; w tym przypadku, gdzie dla każdego urządzenia trenujemy osobny model, został wybrany typ `diag`, jako odpowiedni wybór dla jednowymiarowych danych, ponieważ modeluje tylko wariancję jednej cechy,
- `n_iter` - określa maksymalną liczbę iteracji algorytmu Baum-Welch, który jest używany do trenowania modelu HMM. Większa liczba iteracji może prowadzić do lepszego dopasowania modelu, ale po pewnym punkcie zyski mogą być minimalne. Zbyt mała liczba iteracji może prowadzić do niedotrenowania modelu. Wybrana została maksymalna liczba 00 iteracji, która zapewni odpowiednią zbieżność dla większości przypadków.

3.2 Wybór liczby ukrytych stanów

Wybór odpowiedniej liczby stanów ukrytych w ukrytym modelu Markowa jest kluczowy dla uzyskania dokładnego i skutecznego modelu. Liczba stanów wpływa na zdolność modelu do reprezentowania danych.

W celu wyboru odpowiedniej liczby ukrytych stanów, zastosowane zostały kryteria informacyjne Akaike (AIC) oraz Bayesa (BIC), które są kluczowymi narzędziami w

ocenie i wyborze modeli statystycznych. Pomagają one zbalansować zarówno złożoność modeli jak i ich dopasowanie do danych, minimalizując ryzyko niedouczenia bądź przeuczenia.

3.2.1 Kryteria AIC i BIC

Kryterium AIC jest miarą używaną do oceny jakości modeli statystycznych, biorąc pod uwagę zarówno dopasowanie modelu do danych, jak i jego złożoność. Niższa wartość AIC wskazuje na lepszy model. Wzór na AIC przedstawia się następująco:

$$\mathrm{AIC} = 2k - 2\log(L),$$

gdzie k to liczba parametrów modelu, a L to wartość logarytmu funkcji wiarygodności.

Kryterium BIC to podobne kryterium, jednak bardziej surowo karze za złożoność modelu. Niższa wartość BIC również wskazuje na lepszy model. Wzór na BIC przedstawia się następująco:

$$\mathrm{BIC} = \log(n)k - 2\log(L),$$

gdzie n to liczba obserwacji, k to liczba parametrów modelu, a L to wartość logarytmu funkcji wiarygodności.

3.2.2 Liczba parametrów w modelu HMM

Zarówno kryterium AIC jak i BIC zawiera we wzorze liczbę parametrów modelu. W modelach HMM liczba parametrów (k) zależy od:

1. Liczby stanów ukrytych n,
2. Macierzy przejść między stanami - dla każdego z n stanów, mamy $n-1$ parametrów przejścia (nie uwzględniamy przejść do samego siebie), stąd liczba parametrów w macierzy przejść to $n \times (n-1)$,
3. Wektora początkowych prawdopodobieństw: Liczba parametrów to $n-1$, ponieważ ostatnie prawdopodobieństwo jest określone przez pozostałe, gdyż suma musi wynosić 1,
4. Parametrów emisji- w przypadku modelu z rozkładem Gaussowskim, dla każdego stanu ukrytego mamy dwa parametry: średnią (μ) i wariancję (σ^2), stąd liczba parametrów emisji to $2 \times n$.

Podsumowując, liczba parametrów w modelu HMM z rozkładami Gaussowskimi jest obliczana jako:

$$k = n \times (n-1) + (n-1) + 2 \times n,$$

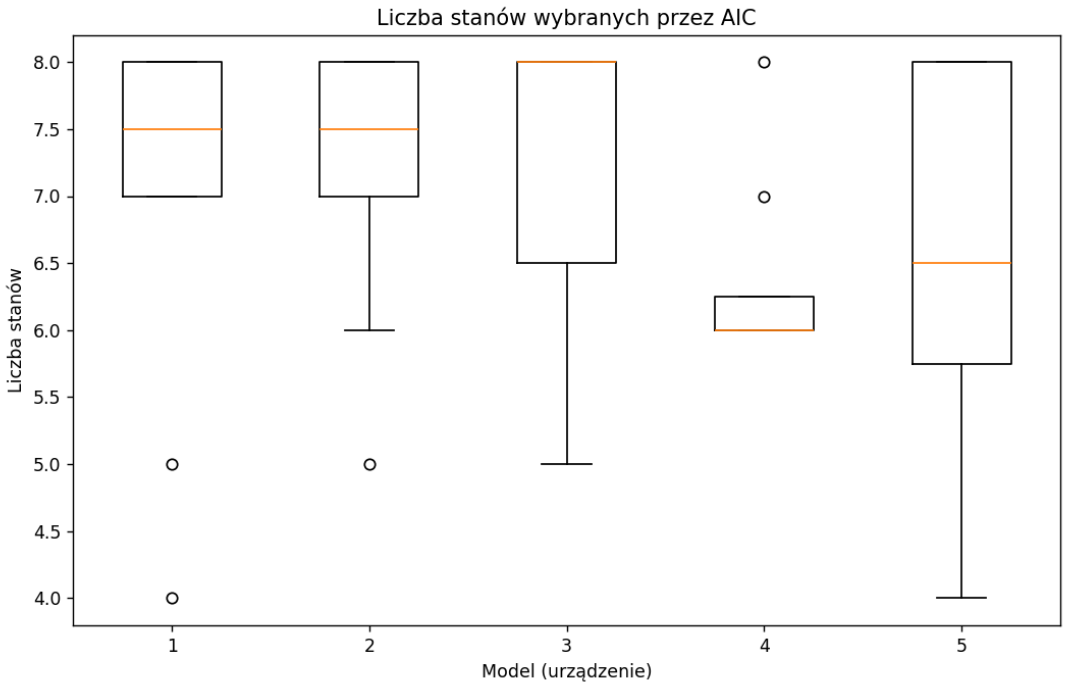
gdzie:

- $n \times (n-1)$ to liczba parametrów w macierzy przejść,
- $n-1$ to liczba parametrów w wektorze początkowych prawdopodobieństw,
- $2 \times n$ to liczba parametrów emisji dla rozkładów Gaussowskich.

3.2.3 Zastosowanie kryteriów informacyjnych

W celu znalezienia optymalnej liczby ukrytych stanów dla każdego urządzenia domowego, przetestowane zostały modele HMM z różnymi liczbami ukrytych stanów (od 2 do 10). Dla każdego modelu obliczane były wartości kryteriów informacyjnych AIC i BIC, a następnie wybierany był model z najniższą wartością tych kryteriów.

Doświadczenie zostało powtórzone 20 razy, a wyniki przedstawione są na wykresach pudełkowych:



Kierując się wartościami dla ukrytych stanów, które były wybierane przez kryteria, wybrane zostały:
- 8 dla modelu dla `lighting2`,
- 8 dla modelu `lighting5`,

- 8 dla modelu `lighting4`,
- 6 dla modelu `refrigerator`,
- 6 dla modelu `microwave`.

Wybrane wartości są optymalne w kontekście rozważanych modeli.

Skutkiem wyboru zbyt małej liczby stanów mogłoby być niedopasowanie modelu (ang. underfitting), gdzie model nie jest w stanie dokładnie przewidywać i reprezentować danych, natomiast zbyt duża liczba stanów może powodować przeuczenie (ang. overfitting). Można napotkać też na sytuację, kiedy niektóre stany mogą nigdy nie być osiągane podczas trenowania - a HMM zakłada, że przejścia między stanami są możliwe i muszą mieć określone prawdopodobieństwa.

4 Klasyfikacja urządzeń

Klasyfikacja urządzeń polega na przypisaniu nowych, nieznanych danych testowych do jednego z wytrenowanych modeli HMM, reprezentujących różne urządzenia. W celu zwiększenia niezawodności klasyfikacji, dla każdego zestawu danych testowych przeprowadzamy proces klasyfikacji kilkukrotnie - wykonujemy 7 powtórzeń. Następnie wybieramy urządzenie, które zostało zaklasyfikowane najwięcej razy spośród powtórzeń. Taka technika pozwala maksymalizować szansę na poprawną klasyfikację urządzenia.

4.1 Kroki klasyfikacji

Aby poprawić dokładność i stabilność klasyfikacji, zastosowane zostało podejście polegające na wielokrotnym powtarzaniu procesu klasyfikacji dla każdego zestawu danych testowych.

- Wczytanie danych testowych: Wczytujemy dane testowe z folderu, które zawierają pobór mocy dla różnych urządzeń.
- Powtarzanie klasyfikacji: Dla każdego pliku z danymi testowymi, powtarzamy proces klasyfikacji 7 razy. W każdym powtórzeniu obliczamy log-likelihood dla każdej sekwencji danych testowych względem wytrenowanych modeli HMM.
- Wybór urządzenia: Dla każdego powtórzenia klasyfikacji, wybieramy urządzenie z modelem HMM, który osiągnął najwyższe prawdopodobieństwo (log-likelihood). Po 7 powtórzeniach klasyfikacji, wybieramy urządzenie, które zostało zaklasyfikowane najwięcej razy.

4.2 Jakość klasyfikacji

W celu sprawdzenia jak jakość klasyfikacji zależy od długości danych testowych, zastosowano własne zbiory testowe, które wyodrębnione zostały z dostarczonych danych. Rozważone zostały następujące długości pomiarów: [200, 500, 1000, 2000, 3000, 4000, 5000, 6000, 7000, 8000, 10000].

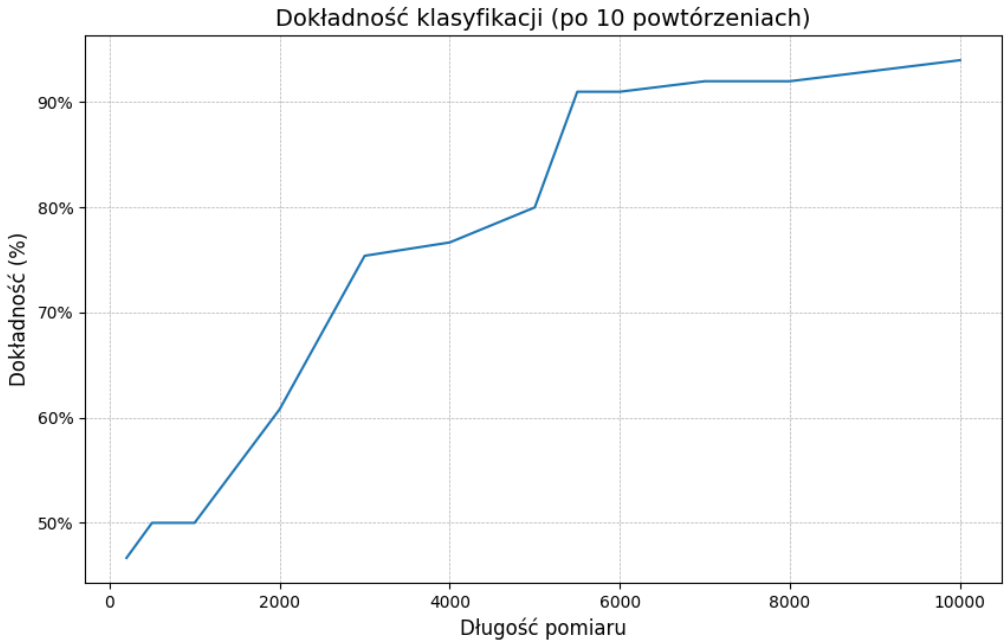
Każda próbka danych testowych pochodziła z jednego z pięciu urządzeń domowych. Zostały one wcześniej oznaczone rzeczywistymi etykietami, które identyfikowały,

z którego urządzenia pochodziły dane. Rzeczywiste etykiety były przechowywane, aby można było później porównać je z wynikami klasyfikacji.

4.3 Zależność od długości pomiaru

Dokładność klasyfikacji została obliczona jako procent poprawnie zaklasyfikowanych próbek testowych. Każda przewidywana klasa (urządzenie) była porównywana z rzeczywistą etykietą próbek danych testowych. Jeżeli przewidywana klasa zgadzała się z rzeczywistą etykietą, próbka była uznawana za poprawnie zaklasyfikowaną.

Dokładność klasyfikacji była obliczana jako stosunek liczby poprawnie zaklasyfikowanych próbek do całkowitej liczby próbek testowych. Wynik był wyrażany jako procent.



Wyniki klasyfikacji pokazują, że jakość klasyfikacji zależy długości danych testowych. Dla krótszych zestawów testowych klasyfikacja była mniej dokładna, co wynika z mniejszej ilości dostępnych danych do analizy.

5 Podsumowanie

Przeprowadzone analizy wykazały, że ukryte modele Markowa są skutecznym narzędziem do klasyfikacji szeregów czasowych, w naszym przypadku zadanych poprzez pobór prądu przez urządzenia domowe. Do skutecznego działania HMM wymagana jest optymalna liczba stanów ukrytych, a kryteria informacyjne AIC i BIC są pomocne w wyborze ich wyborze. Liczba stanów oraz długość danych testowych ma istotny wpływ na dopasowanie modelu do danych, a co za tym idzie, na dokładność klasyfikacji.