Dryf danych i pojęcia – detekcja oraz adaptacja

Stanisław Kaźmierczak

Agenda

1. Motywacja
2. Dryf danych
3. Dryf pojęcia
4. Projektowanie systemu predykcyjnego
5. Metody adaptacji
6. ARMA
7. ARIMA

Wprowadzenie (1)

„Dziwne, u mnie działa.”

„Dziwne, na zbiorze testowym jakość modelu jest dobra.”

Istotnymi przyczynami, dla których jakość modelu działającego na produkcji jest istotnie gorsza niż podczas testów są dryf pojęcia (ang. *Concept drift*) oraz dryf danych (ang. *Data drift*).

[1]

Wprowadzenie (2)

Dryf danych

- Rozkład danych wejściowych się zmienia.
- Ważnym typem są klasy, których instancje są rzadkie lub wręcz nieobecne w zbiorze treningowym.
- Przykład 1: system do rozpoznawania mowy w języku angielskim osiągający generalnie wysoką skuteczność słabo radzi sobie z brytyjskim akcentem (zbiór treningowy i testowy zawierał mało próbek mowy Brytyjczyków).
- Przykład 2: model predykujący zużycie energii na podstawie historycznych; ale w między czasie zmiany klimatu spowodowały zmiany w charakterystyce pogody

Wprowadzenie (3)

Dryf pojęcia

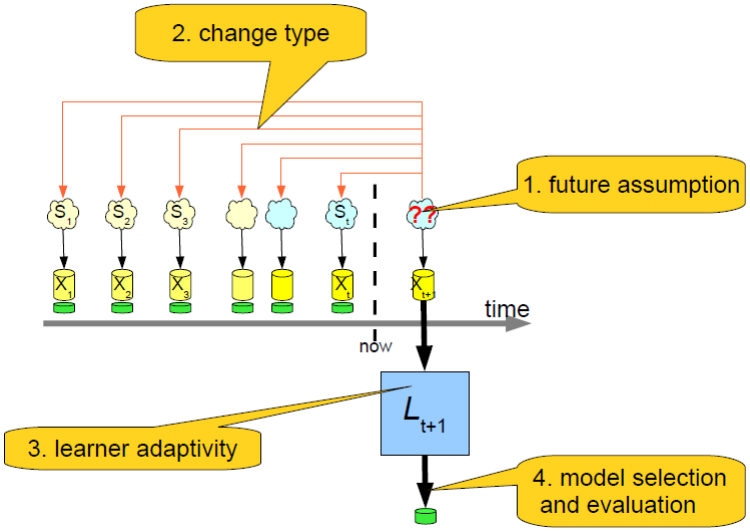
- Mapowanie x -> y zmienia się
- Przykład: model predykuje cenę mieszkania, a inflacja sukcesywnie podnosi ceny

Założenia

- Szum nie jest uznawany za dryf (o ile rozkład danych się nie zmienia).
- Sezonowość generalnie nie jest również traktowana jako dryf (o ile jesteśmy jej świadomi)
  - Nie zawsze jest oczywiste, kiedy zmiany związane z sezonowością będą miały miejsce (np. popularność lodów jest zależna od pogody).
- Identyfikując rozkład danych, z którego pochodzi analizowana instancja, możemy zidentyfikować źródło tych danych.
  - Znając źródło danych możemy zastosować osobny dla tego źródła model

Projektowanie systemu predykcyjnego

- Celem systemu jest jak najlepsza predykcja w chwili *t+1*
- Aby zbudować taki system, należy podjąć 4 zasadnicze kwestie:
  - Założenie odnośnie rozkładu danych w chwili *t+1*
  - [Estymacja zmiany rozkładu]
  - Wybór metody adaptacji
  - Selekcja modelu



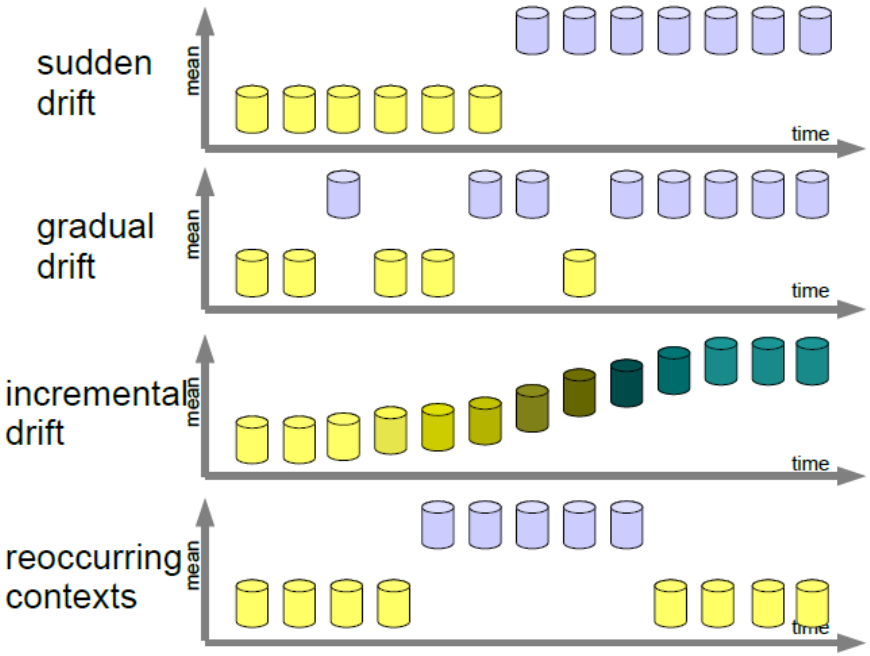
[2]

Założenia odnośnie rozkładu danych

Istnieją trzy główne podejścia:

- Założenie, że rozkład danych w chwili $t+1$ będzie taki sam, jak w chwili t
  - Najprostsze, a jednocześnie najrzadziej stosowane
- Estymacja rozkładu na podstawie wartości cech instancji w chwili $t+1$
  - Ustalenie źródła odbywa się poprzez porównanie wartości cech instancji $t+1$ z wcześniejszymi instancjami, których źródła znamy
- Predykcja rozkładu na podstawie dostępnych instancji

Rodzaje zmian



[2]

- Powtarzający się kontekst nie musi być periodyczny (jest to zasadnicza różnica względem sezonowości).
- Określenie rodzaju zmiany jest kluczowe w kontekście zaprojektowania mechanizmu adaptacji.

Metody adaptacji (1)

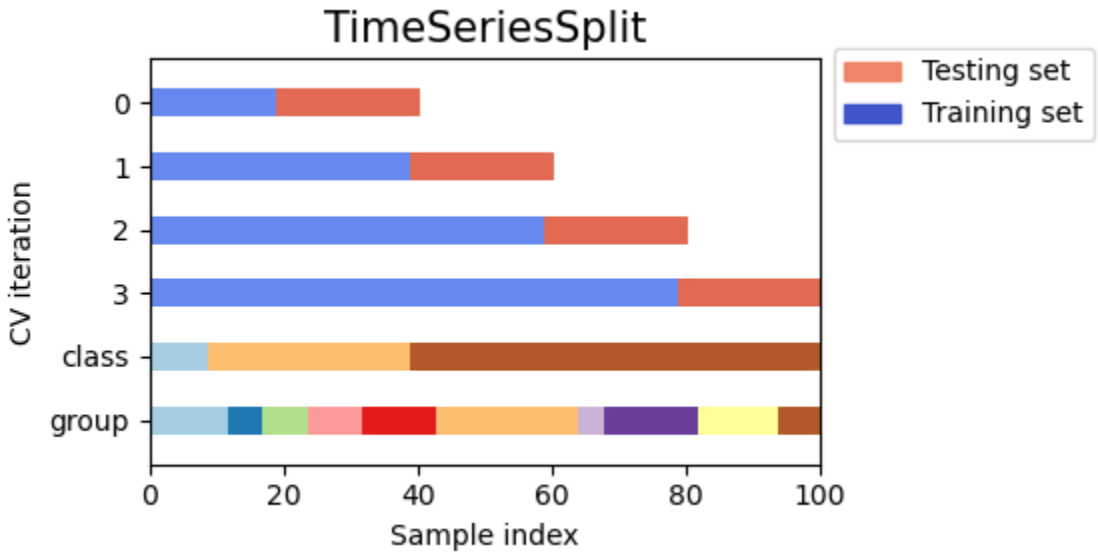
- Uczenie online
  - Aktualizacja modelu po każdej nowej instancji lub niewielkiej grupie nowych instancji
- Okresowe dotrenowywanie
  - Triggery:
    - stałe odstępy czasu
    - duża różnica między instancjami w kolejnych oknach czasu
    - spadek jakości modelu poniżej ustalonego progu
  - Wariant: dotrenowywanie tylko na tych instancjach, na których obecna wersja modelu nie radzi sobie dobrze
  - Kiedy opłaca się uczyć model od początku tylko na najnowszych danych?
  - Przykładami modeli, które można dotrenowywać są sieci neuronowe, adaptacyjne drzewa decyzyjne [3], lub zmodyfikowany SVM [4].

Metody adaptacji (2)

- Uczenie zespołowe (ang. *Ensemble learning*)
  - Szczególnie przydatny jest tu wariant, w którym dla każdej nowej grupy danych tworzony jest nowy model, a następnie dodawany do zespołu
  - Waga modelu może być zależna od świeżości danych, na których model był trenowany
- Usuwanie cech
  - Tworzone modele budowane są na pojedynczych cechach
  - Cechy, na których budowane były modele, których jakość się pogorszyła, mogą być uznane za dryfujące i warto rozważyć ich usunięcie.
- Przygotowanie danych
  - Cel: likwidacja/zmniejszenie systematycznych zmian (trendy, sezonowość) w danych w dziedzinie czasu, np. Model ARIMA

Selekcja modelu

- Kroswalidacja dla szeregów czasowych



[5]

- Alternatywą jest stałej długości zbiór treningowy

Detekcja dryfu – podsumowanie

Dwie główne grupy metod służące do detekcji dryfu:

- Testy statystyczne dla wartości cech instancji pochodzących z kolejnych okien czasowych.
- Porównanie wyniku klasycznej kroswalidacji z wynikiem kroswalidacji dla szeregów czasowych.

Model AR (1)

- Model autoregresyjny przewiduje przyszłe wartości szeregu czasowego na podstawie **minionych wartości**
- Wymaga, aby szereg czasowy był stacjonarny
- Stacjonarny szereg czasowy to szereg, którego właściwości nie zmieniają się w czasie
- Model autoregresji bazujący na jednej obserwacji wstecz [6]:

$$y_t = c + \phi_1 y_{t-1} + \epsilon_t$$

Where:

y_t Is the value at time step t, c is a constant, ϕ_1 is a coefficient, and ϵ_t is a white noise error term with $\epsilon_t \sim N\ (0, \sigma^2)$.

Model AR (2)

- Ogólny model autoregresji bazujący na p obserwacjach wstecz:

$$y_t = c + \phi_1 y_{t-1} + \phi_2 y_{t-2} \ldots + \phi_p y_{t-p} + \epsilon_t$$

$$y_t = c + \sum_{i=1}^{p} \phi_i \, y_{t-i}$$

Where:

ϕ_i Is the corresponding coefficient for each respective prior time step y_{t-i}

- Zatem wektor współczynników do estymacji ma postać:

$$\phi = (\phi_1, \phi_2, \ldots \phi_i)$$

Model MA (1)

- Model średniej kroczącej próbuje przewidywać przyszłe wartości na podstawie wcześniejszych błędów predykcji
- Model ten zakłada, że model AR może przybliżać ten ciąg błędów
- Nie należy mylić z pojęciem średniej kroczącej, która jest procesem wygładzania, a nie modelem predykcyjnym
- Model średniej kroczącej bazujący na jednej obserwacji wstecz:

$$y_t = c + \theta_1 \epsilon_{t-1}$$

Where:

y_t Is the value at time step t, c is a constant, θ_1 is a coefficient, and ϵ_{t-1} is a previous white noise error term.

- Jest to zatem regresja liniowa, w której zmienną niezależną jest wartość białego szumu z wcześniejszego punktu czasowego

Model MA (2)

- Ogólny model średniej kroczącej bazujący na q obserwacjach wstecz:

$$y_t = c + \theta_1 \epsilon_{t-1} + \theta_2 \epsilon_{t-2} \ldots + \theta_q \epsilon_{t-q} + \epsilon_t$$

$$y_t = c + \sum_{j=1}^{q} \theta_j \, \epsilon_{t-j}$$

Where:

θ_j Is the corresponding coefficient for each respective prior error ϵ_{t-j}

- Zatem wektor współczynników do estymacji ma postać:

$$\theta = (\theta_1, \theta_2, \ldots \theta_j)$$

Model ARMA

Model ARMA stanowi połączenie modelu autoregresyjnego oraz średniej kroczącej:

$$y_t = c + \sum_{i=1}^{p} \phi_i y_{t-i} + \sum_{j=1}^{q} \theta_j \epsilon_{t-j} + \epsilon_t$$

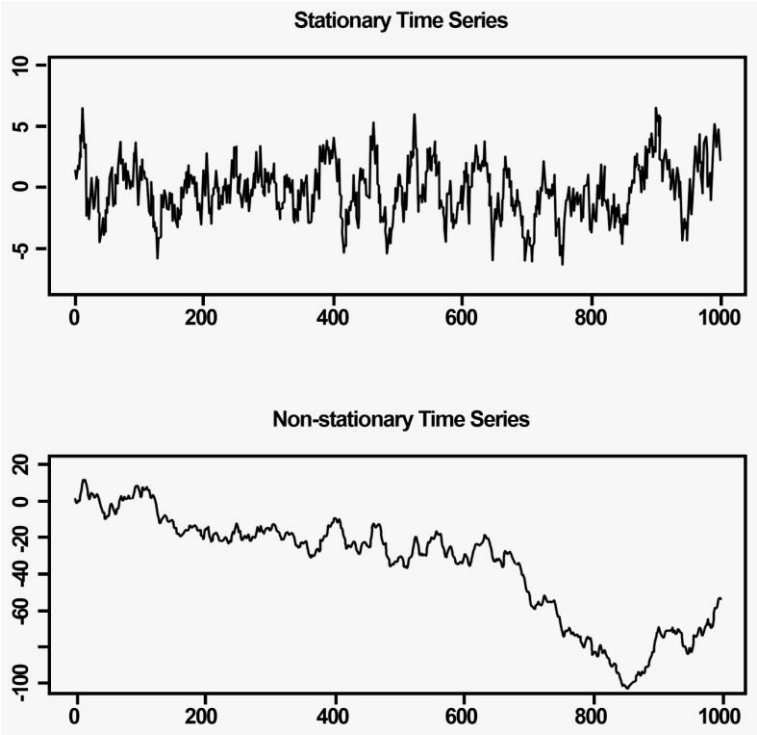
Where:

p and q are the orders of the AR and MA models, respectively.

Wadą modelu ARMA pozostaje założenie o stacjonarności predykowanego szeregu czasowego.

Model ARIMA (1)

- ARIMA – *Auto Regressive Integrated Moving Average*
- Stacjonarność szeregu czasowego implikuje, że statystyczne właściwości takie jak średnia lub odchylenie standardowe nie zmieniają się w czasie.
- Rzeczywiste szeregi czasowe są z reguły niestacjonarne
- Należy przekształcić je do postaci stacjonarnej



Model ARIMA (2)

- Proces ten nazywa się różnicowaniem szeregu czasowego (ang. Differencing).
- Różnica liczona jest d razy aż do momentu otrzymania szeregu stacjonarnego.
- Model ten może być zatem wyrażony formułą:

$$\text{ARIMA}(p, d, q)$$

Where:

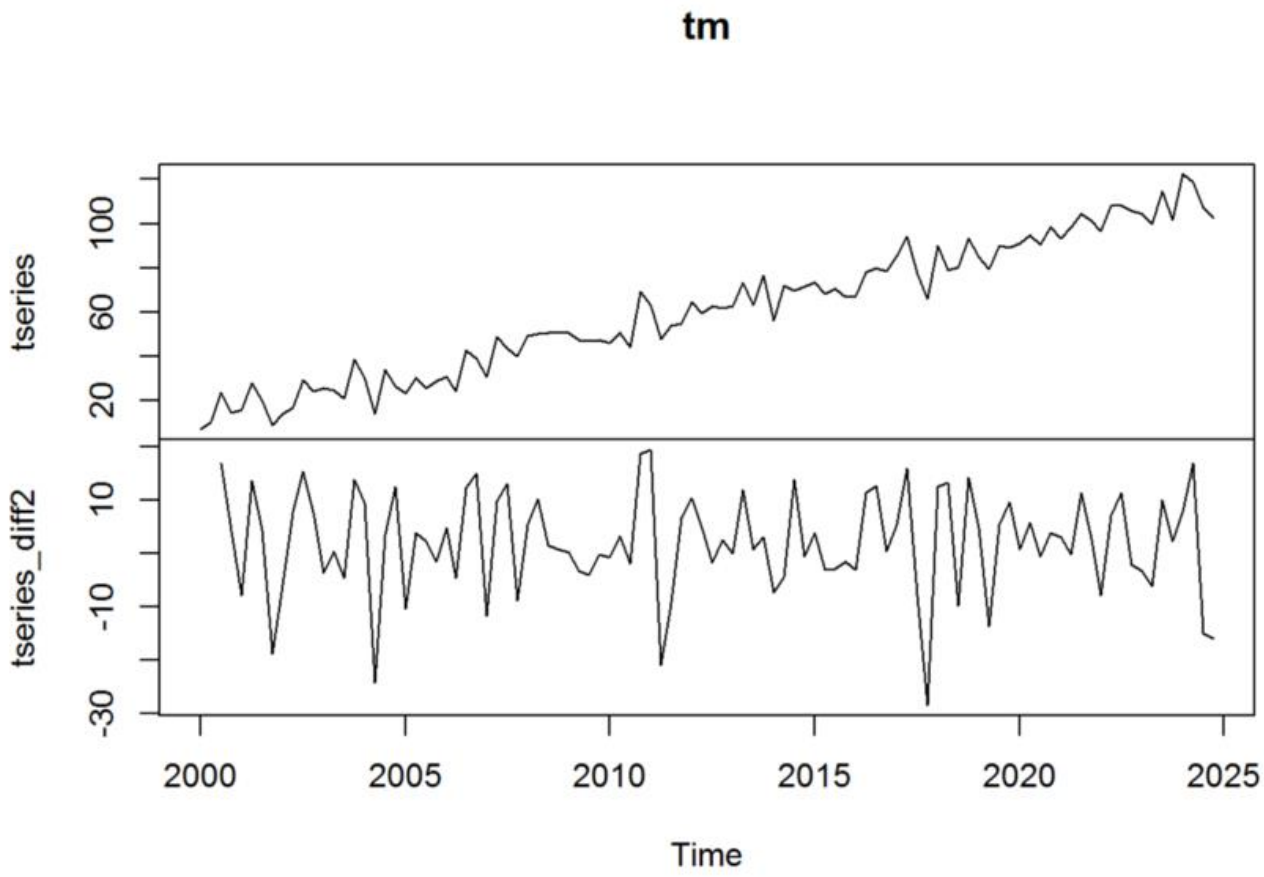
p is the order of the AR model component, d is the number of differences to conduct on the time series and q is the order of the MA model component.

- Różnicowanie:

$$y'_t = y_t - y_{t-1}$$

$$\begin{aligned} y^*_t &= y'_t - y'_{t-1} \\ &= (y_t - y_{t-1}) - (y_{t-1} - y_{t-2}) \\ &= y_t - 2y_{t-1} + y_{t-2} \end{aligned}$$

Model ARIMA (3)

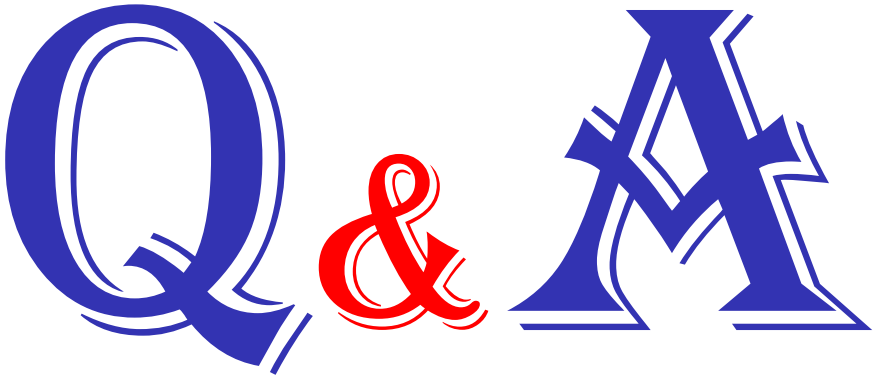


Źródła (1)

1. https://ipvm.com/reports/face-masks
2. Žliobaitė, I. (2010). Learning under concept drift: an overview. *arXiv preprint arXiv:1010.4784*.
3. Hulten, G., Spencer, L., & Domingos, P. (2001, August). Mining time-changing data streams. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 97-106).
4. Klinkenberg, R., & Joachims, T. (2000, June). Detecting concept drift with support vector machines. In *ICML* (pp. 487-494).
5. https://scikit-learn.org/stable/modules/cross_validation.html#time-series-split

Źródła (2)

6. https://medium.com/analytics-vidhya/a-thorough-introduction-to-arima-models-987a24e9ff71


Q&A