SYSTEMY UCZĄCE SIĘ

WYKŁAD 15. ANALIZA DANYCH – WYKRYWANIE OBSERWACJI ODSTAJĄCYCH, UZUPEŁNIANIE BRAKUJĄCYCH DANYCH

Dr hab. inż. Grzegorz Dudek
Wydział Elektryczny
Politechnika Częstochowska

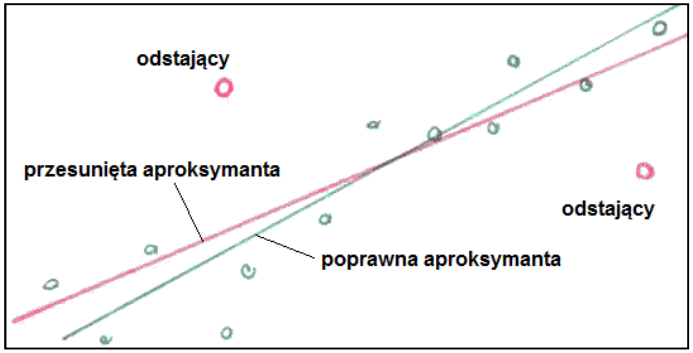
Częstochowa 2014

WYKRYWANIE OBSERWACJI ODSTAJĄCYCH

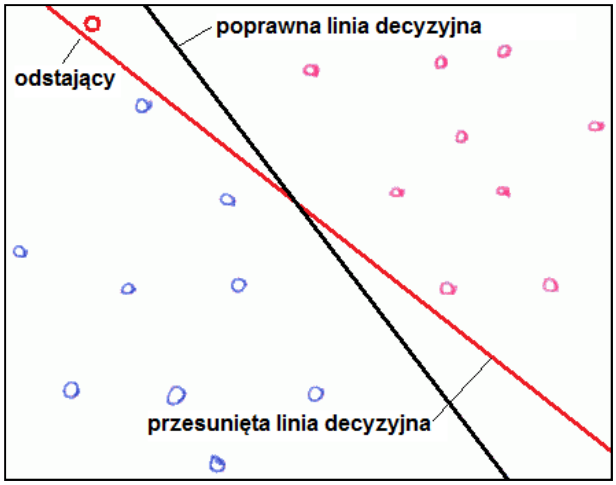
Obserwacja odstająca (*outlier*) jest to obserwacja znacząco różniąca się od pozostałych obserwacji.

Obserwacje odstające zwykle zakłócają obraz procesu i wpływają na jakość modelu (klasyfikatora, aproksymatora, …).

Zakłócenie aproksymacji:



Zakłócenie klasyfikatora:



WYKRYWANIE OBSERWACJI ODSTAJĄCYCH

Obserwacje odstające mogą mieć różne źródła, np. błędy w układzie pomiarowym lub zmiany w mierzonym procesie wywołane działaniem zakłóceń, nietypowych zdarzeń itp.

Te nietypowe zdarzenia mogą być przedmiotem zainteresowania, np.:

- Detekcja intruzów – nietypowe zachowanie
- Oszustwa przy użyciu karty kredytowej – nietypowe wzorce użycia karty
- Uszkodzenia czujników – nietypowe wzorce odczytów
- Diagnozy medyczne – odstające od typowych wyniki badań
- Wykrywanie uszkodzeń i awarii systemów technicznych – parametry przekraczają wartości graniczne
- Wykrywanie anomalii pogodowych, trzęsień ziemi, zmian klimatycznych – na podstawie obrazów satelitarnych, nietypowych wartości mierzonych parametrów
- …

W powyższych sytuacjach dane opisujące obiekty, procesy mają nienormalne, odstające od typowych wartości (obrazy).

WYKRYWANIE OBSERWACJI ODSTAJĄCYCH

Sposób postępowania z obserwacjami odstającymi:

- eliminacja obserwacji ze zbioru danych
- zastępowanie średnią arytmetyczną (lub inną) obserwacji sąsiednich lub reprezentujących podobne cechy
- potraktowanie obserwacji odstających jako brakujących danych i uzupełnienie tych danych odpowiednimi metodami

WYKRYWANIE OBSERWACJI ODSTAJĄCYCH - METODY

Metody identyfikacji obserwacji odstających:

1. Analiza wartości poszczególnych atrybutów

Dla każdego atrybutu oblicza się pierwszy (Q1) i trzeci kwartyl (Q3) oraz rozstęp międzykwartylowy RQ = Q3–Q1 (*interquartile range IQR*).

Kwartyle dzielą wszystkie nasze obserwacje na cztery równe co do ilości obserwacji grupy (w teorii).

Kwartyl pierwszy (Q1) dzieli obserwacje w stosunku 25% - 75%, co oznacza, że 25% obserwacji jest niższa bądź równa wartości Q1, a 75% obserwacji jest równa bądź większa niż wartość Q1

Kwartyl drugi (Q2), inaczej zwany medianą dzieli obserwacje na dwie części w stosunku 50%-50%

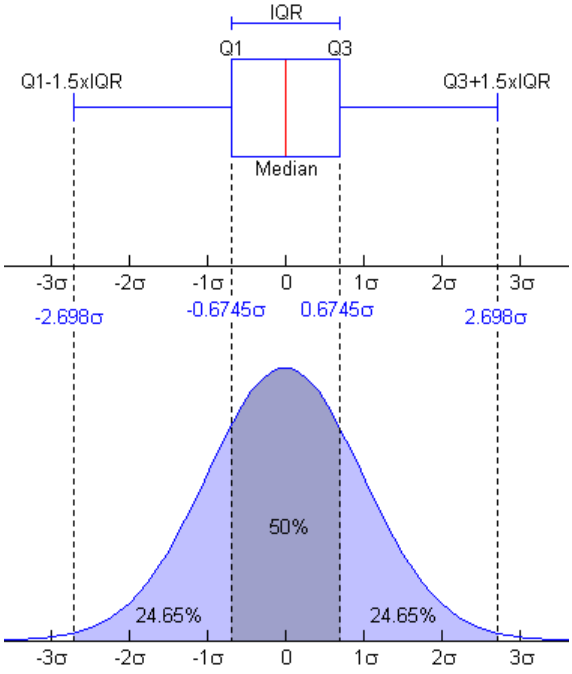
Kwartyl trzeci (Q3) dzieli obserwacje w stosunku 75% - 25%, co oznacza, że 75% obserwacji jest niższa bądź równa wartości Q3, a 25% obserwacji jest równa bądź większa niż wartość Q1

Za obserwacje, które można podejrzewać, że są odstające, uważa się te, których atrybuty wykraczają poza przedział (Q1–1,5RQ, Q3+1,5RQ)

Za obserwacje ekstremalnie odstające uznaje się te których atrybuty wykraczają poza przedział (Q1–3RQ, Q3+3RQ)

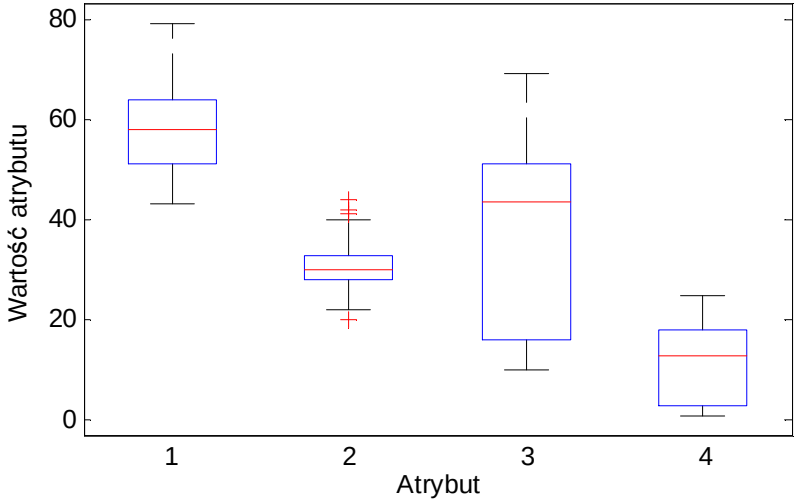
WYKRYWANIE OBSERWACJI ODSTAJĄCYCH - METODY

Ten sposób identyfikacji obserwacji odstających można zobrazować wykresem pudełkowym (*boxplot*):



WYKRYWANIE OBSERWACJI ODSTAJĄCYCH - METODY

Wykres pudełkowy dla zbioru Iris (4 atrybuty):

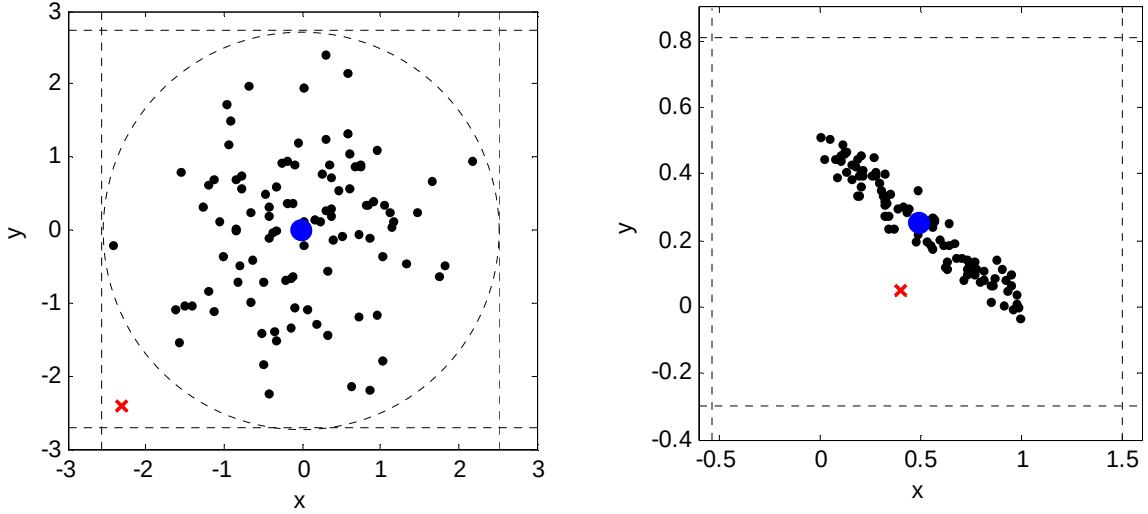


Czerwone krzyżyki oznaczają wartości atrybutów wykraczające poza zakres ±1,5RQ

Liczba atrybutów przykładu **x**, których wartości znajdują się poza ww. przedziałami, może być miernikiem nietypowości przykładu.

WYKRYWANIE OBSERWACJI ODSTAJĄCYCH - METODY

Ten sposób detekcji obserwacji nietypowych, polegający na niezależnej analizie poszczególnych atrybutów, nie zawsze prowadzi do dobrych rezultatów (obserwacje odstające nie zawsze widziane są jako takie w analizie jednowymiarowej – patrz rys.).



Granice przedziałów kwartylowych (Q_1–1,5R_Q, Q_3+1,5R_Q) oraz granice detektora z metryką euklidesową oznaczono liniami przerywanymi

2. Najprostszym sposobem detekcji obserwacji odstających, opartym na analizie wielowymiarowej, jest analiza odległości euklidesowych pomiędzy obserwacjami a ich środkiem (wektorem średnich **m**):

$$d_{Ei} = \sqrt{(\mathbf{x}_i - \mathbf{m})^{\mathsf{T}}(\mathbf{x}_i - \mathbf{m})}$$

W metryce d_E atrybuty są skumulowane i nawet jeśli nie wykraczają one indywidualnie poza przedziały kwartylowe j.w., obserwacja może być zidentyfikowana jako odstająca, jeśli tylko jej odległość jest dostatecznie duża od środka skupiska **m**. Metoda jest uprawniona, gdy dane charakteryzują się rozkładem o symetrii radialnej.

3. W przypadku rozkładu eliptycznego należy skorzystać z metryki Mahalanobisa, która uwzględnia informacje o wariancjach poszczególnych składowych i korelacjach pomiędzy nimi:

$$d_{Mi} = \sqrt{(\mathbf{x}_i - \mathbf{m})^{\mathsf{T}}\mathbf{S}^{-1}(\mathbf{x}_i - \mathbf{m})}$$

gdzie **S** jest macierzą kowariancji

Obserwacje odstające cechuje duża odległość Mahalanobisa od środka **m**.

Identyfikacja obserwacji odstających jest zakłócona w przypadku, gdy ich udział w zbiorze danych jest znaczący. Skupisko obserwacji odstających przyciąga estymowany centroid i fałszuje macierz kowariancji. Powoduje to dwa niekorzystne efekty: zmniejszenie odległości dla obserwacji ostających (*masking effect*) i zwiększenie odległości dla obserwacji typowych, ulokowanych po przeciwnej stronie środka **m** w stosunku do skupiska danych odstających (*swamping effect*), co sprawia, że obserwacje te mogą zostać rozpoznane jako nietypowe. Efekty te redukuje algorytm Gnanadesikana--Ketteringa, w którym środek **m** i macierz kowariancji estymuje się po wykluczeniu obserwacji najbardziej odległych.

4. Kolejny sposób identyfikacji obserwacji odstających opiera się na statystyce h_i nazywanej dźwignią (*leverage*) lub wpływem *i*-tej obserwacji. Metoda ta pozwala zidentyfikować tzw. obserwacje wpływowe. Obserwację uznaje się za wpływową, jeśli w wyniku nieznacznej zmiany jej wartości lub usunięcia z danych znacznie zmieniają się oszacowane parametry modelu.

WYKRYWANIE OBSERWACJI ODSTAJĄCYCH - METODY

5. Źródłem informacji o obserwacjach odstających jest też wykonana a posteriori diagnostyka błędów (reszt) modelu. Błędy odzwierciedlają niezgodność pomiędzy wartościami obserwowanymi i przewidywanymi przez model. Dobrze dopasowany model charakteryzuje się małymi resztami dla obserwacji typowych i dużymi dla obserwacji odstających. Identyfikację obserwacji odstających na podstawie reszt modelu można wykonać, wykorzystując standaryzowane wartości resztowe, analizę odpowiedzi modelu po usunięciu "podejrzanej" obserwacji (metoda DFFITS – *difference of fits*) oraz odległości Cooka zależne od błędów modelu i ich wariancji oraz wpływów obserwacji.

6. Obserwacje odstające można też wykryć metodami grupowania opartymi na gęstościach (patrz wykład 9, str. 17).

W przypadku brakujących danych mamy do wyboru trzy strategie:

1. Pominięcie obserwacji z brakującymi wartościami.
2. Zastosowanie obserwacji niekompletnych w procesie konstrukcji modelu i/lub w trybie pracy odtworzeniowej.
3. Uzupełnienie (imputacja) brakujących danych.

Pierwszy sposób może być stosowany, gdy liczba niekompletnych obserwacji jest ograniczona, tzn. gdy pozostałe, kompletne, obserwacje przenoszą niezbędne informacje potrzebne do konstrukcji modelu dobrej jakości.

Drugie podejście zależne jest od specyfiki modelu, np. w modelach minimalnoodległościowych obliczenie odległości pomiędzy obserwacjami może zachodzić z pominięciem brakujących składowych wektora wejściowego, a obliczenie odpowiedzi jest możliwe bez znajomości wszystkich składowych wszystkich wektorów odpowiedzi w zbiorze uczącym. Także model wykorzystujący drzewo regresyjne dobrze radzi sobie z brakującymi danymi. Inaczej jest np. w przypadku, gdy model opiera się na perceptronowej sieci neuronowej, która wymaga

kompletnej informacji wejściowej i wyjściowej w procesie uczenia i pełnej informacji wejściowej w trybie odtworzeniowym.

Brakujące wartości atrybutów przykładu obciążeń $\mathbf{x}_i$ można estymować z obserwacji najbliższych w sensie geometrycznym. Wartość brakującej *j*-tej składowej jest średnią z wartości tej składowej w *k* najbliższych sąsiadach wektora $\mathbf{x}_i$:

$$x_{i,j} = \frac{1}{k} \sum_{l \in \Theta_k(\mathbf{x}_i)} x_{l,j}$$

gdzie $\Theta_k(\mathbf{x}_i)$ – zbiór indeksów *k* najbliższych sąsiadów wektora $\mathbf{x}_i$.

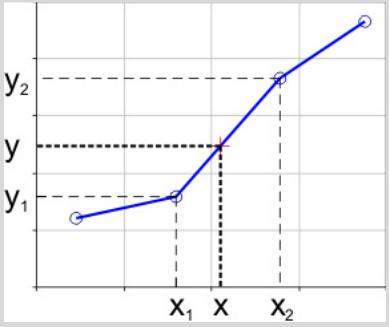
Do zbioru najbliższych sąsiadów zalicza się wektory najbliższe do $\mathbf{x}_i$ w sensie odległości euklidesowej wyznaczanej z pominięciem brakujących składowych.

Inna grupa metod imputacji brakujących danych opiera się na interpolacji.

Zadaniem interpolacji jest utworzenie funkcji, która przebiega przez zadane punkty. Stosuje się różne klasy funkcji do interpolowania – wielomiany algebraiczne, funkcje sklejane, funkcje trygonometryczne.

Zadanie interpolacji możemy sformułować następująco:

W przedziale [a,b] mamy danych n+1 punktów x0,x1,...,xn (węzły interpolacji) oraz wartości funkcji f(x) w tych punktach f(x0)=y0, f(x1)=y1, ..., f(xn)=yn. Znaleźć funkcję g(x), która w węzłach interpolacji ma te same wartości co f(x) i przybliża tę funkcję poza węzłami.

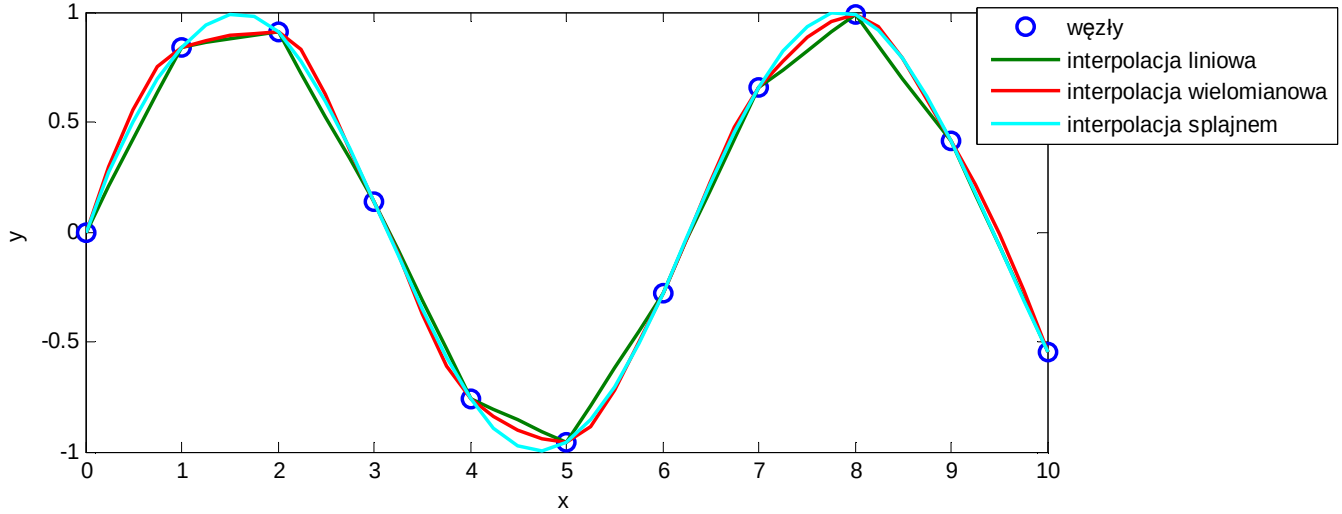


Funkcja interpolacyjna często przyjmuje postać wielomianu *m*-tego stopnia, rozpiętego na *m*+1 węzłach. Często wystarczy aby był to wielomian 3-stopnia.

Interpolacja wielomianem pozwala uwzględniać lokalne nieliniowości. Wielomianową funkcję interpolacyjną *m*-tego stopnia opartą na węzłach (x_i, y_i), $i = 1, 2, \ldots, m+1$ wyraża wzór Lagrange'a:

$$g(x) = \prod_{i=1}^{m+1}(x - x_k)\sum_{j=1}^{m+1}\frac{y_j}{(x - x_j)\prod_{\substack{i=1 \\ i \neq j}}^{m+1}(x_j - x_i)}$$

Lepsze właściwości interpolacyjne posiadają funkcje sklejane (splajny).



Do imputacji danych można użyć metod aproksymacyjnych. Zalety metod aproksymacyjnych ujawniają się, gdy dane obarczone są szumem.