Podstawy Przetwarzania Danych

Wykład 6: Braki w danych

dr inż. Marcin Luckner
mluckner@mini.pw.edu.pl

Wydział Matematyki i Nauk Informacyjnych

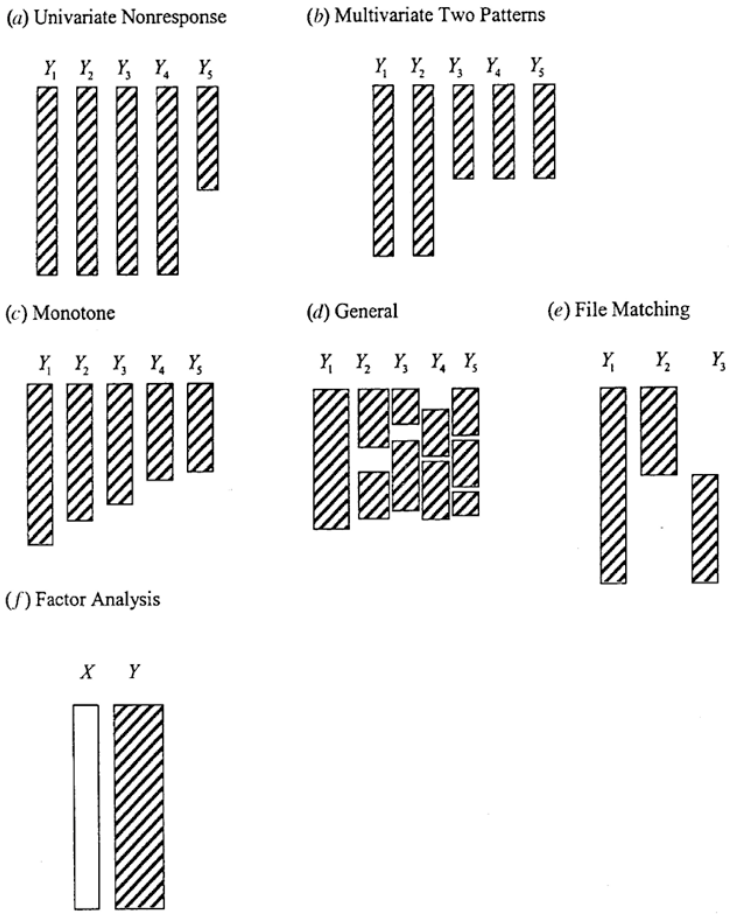
Wersja 1.1
5 marca 2021

Fundusze Europejskie Wiedza Edukacja Rozwój
Rzeczpospolita Polska
Politechnika Warszawska
Unia Europejska Europejski Fundusz Społeczny

Projekt „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca” współfinansowany jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Zadanie 10 pn. „Modyfikacja programów studiów na kierunkach prowadzonych przez Wydział Matematyki i Nauk Informacyjnych”, realizowane w ramach projektu „NERW 2 PW. Nauka – Edukacja – Rozwój – Współpraca”, współfinansowanego jest ze środków Unii Europejskiej w ramach Europejskiego Funduszu Społecznego.

Typy braków danych



Rysunek 1: Przykłady braków [Little and Rubin, 2002]

- Braki w pojedynczej zmiennej
- Braki w wielu zmiennych
  - Jednoczesne braki w kilku zmiennych
- Narastajace
  - Przyrastający brak danych
- Ogólne
  - Rozproszone braki pojedynczych danych
- Brak danych z źródła
  - Brak danych z jednego źródła
- Ukryta zmienna
  - Brak danych dla danej cechy

Powody powstawania braków

- Brak danych dla części obiektów.
  - Dziekanat zapisuje informacje o dacie zaliczenia przedmiotu.
- Brak informacji z części sensorów.
  - Brak pomiaru tętna w części danych pochodzących z aplikacji dla biegaczy
- Wyłączanie sensorów podczas trwania eksperymentu.
  - Dane obserwacyjne sondy kosmicznej.
- Brak odpowiedzi na poszczególne pytania.
  - Respondenci nie są zmuszani do ujawniania informacji.
  - Deklaracja płci.
- Brak wyliczania danych dla części podmiotów.
  - Dane dla aglomeracji Warszawskiej administrowane przez wiele podmiotów.
- Usunięcie wcześniej pozyskanych danych.
  - Rozporządzenie o Ochronie Danych Osobowych,
  - General Data Protection Regulation.

Typy braków

- Całkowicie przypadkowe występowanie braków
  - Missing Completely at Random (MCAR)
- Przypadkowe występowanie braków
  - Missing at Random (MAR)
- Nieprzypadkowe występowanie braków
  - Missing Not at Random (MNAR)

Oznaczenia

- Określmy analizowany zbiór danych jako $Y = (y_{ij})$.
- Macierz $M = (m_{ij})$ opisuje braki w zbiorze danym jako

$$m_{ij} = \begin{cases} 1, & \text{jeżeli brakuje danej } y_{ij} \\ 0, & \text{jeżeli istnieje dana } y_{ij} \end{cases}$$

- Możemy teraz założyć, że zbiór danych składa się z dwóch typów elementów
  - Y_{obs} - zmienne z pełnymi obserwacjami (bez braków)
  - Y_{mis} - zmienne z brakującymi obserwacjami

MCAR

- Braki są określane jako całkowicie przypadkowe (MCAR) kiedy prawdopodobieństwo braków danych nie zależy ani od wartości zmiennych z brakami Y_{mis}, ani od wartości zmiennych bez braków Y_{obs}.
- Dopuszczalne jest jednak, aby braki zależały od jakiś nieznanych parametrów ϕ

$$\Pr(M \mid Y, \phi) = \Pr(M \mid \phi)$$

- Innymi słowy prawdopodobieństwo braku danych nie jest powiązane z żadną rejestrowaną zmienną.
- Najczęściej założenie, że braki są całkowicie przypadkowe nie jest zasadne.

Przykład MCAR

Tabela 1: Brak odczytu wiersza danych

Pomiar	Prędkość [km/h]	X	Y
1	NA	NA	NA
2	50	2	1
3	60	2	2
4	50	2	3
5	50	2	4

- Brak spowodowany niewystąpieniem odczytów podczas dokonywania pomiaru.
- Brak nie jest zależny od jakiejkolwiek zmiennej.

MAR

- Braki są określane jako przypadkowe (MAR) kiedy prawdopodobieństwo braków danych zależy tylko od zmiennych bez braków Y_{obs} i czynników zewnętrznych ϕ.

$$\Pr(M \mid Y, \phi) = \Pr(M \mid Y_{obs}, \phi)$$

- W takim przypadku zakładamy, że nie ma powiązania występowania braków z wartościami danej zmiennej.

Przykład MAR

Tabela 2: Brak odczytu lokalizacji pojazdu

Pomiar	Prędkość [km/h]	X	Y
1	40	1	1
2	500	NA	NA
3	60	2	2
4	50	2	3
5	50	2	4

- Brak spowodowany błędnym, ale istniejącym odczytem prędkości, który nie pozwolił na wyliczenie aktualnej pozycji.
- Brak jest zależny od zmiennej niezawierającej braków.

MNAR

- Braki są określane jako nieprzypadkowe (MNAR) kiedy prawdopodobieństwo braków danych zależy od zmiennych z brakami Y_{mis} i czynników zewnętrznych ϕ.

$$\Pr(M \mid Y, \phi) = \Pr(M \mid Y_{mis}, \phi)$$

- Zakładamy, że występowanie braków nie jest przypadkowe i da się opisać w zależności od wartości zmiennej.

Przykład MNAR

Tabela 3: Brak odczytu zapisu prędkości

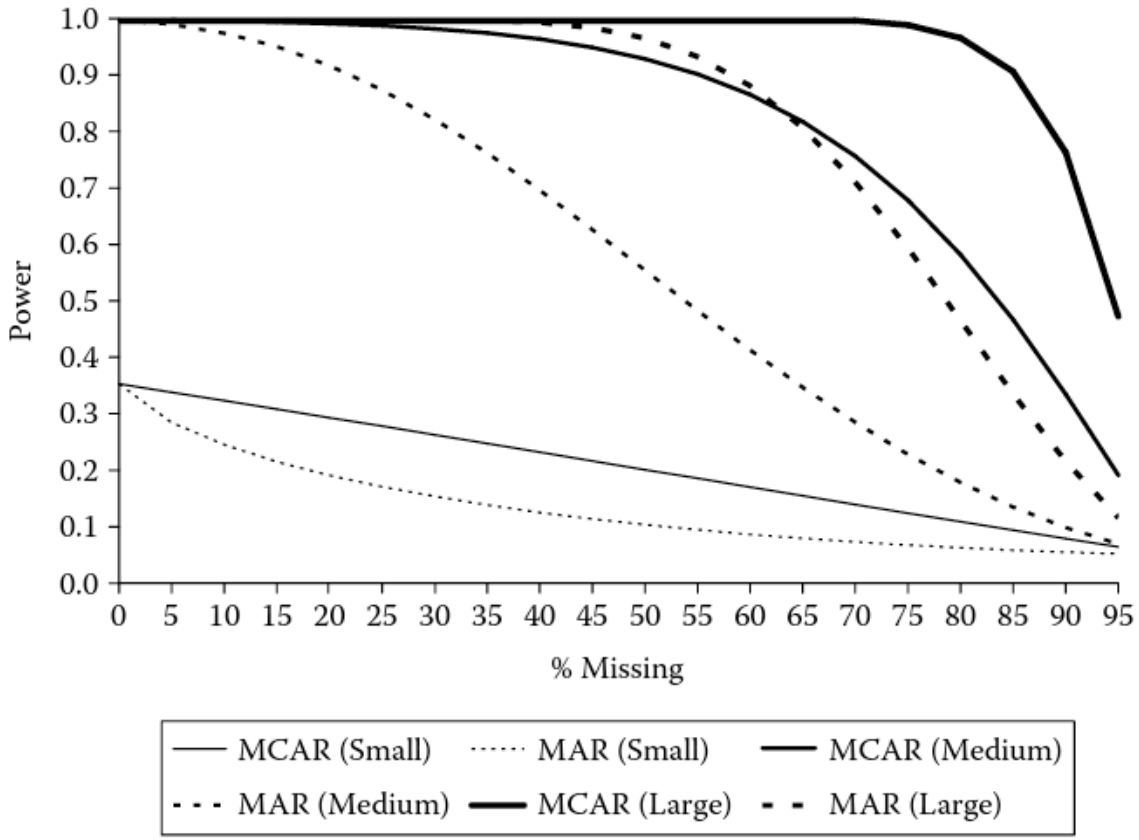
Pomiar	Prędkość [km/h]	X	Y
1	5	1	1
2	2	2	1
3	0	2	2
4	NA	2	3
5	3	2	4

- Brak spowodowany wykroczeniem odczytu prędkości poza dopuszczalną skalę.
- Brak jest zależny od wartości zmiennej w której występuje.

Eksperyment oceniający wpływ braków

- Dla 250 próbek oceniono wpływ braków danych na jakość analizy statystycznej [Davey and Savla, 2009].
- Mierzono moc testu czyli prawdopodobieństwo niepopełnienia błędu nieodrzucenia hipotezy zerowej jeżeli jest ona fałszywa.
- Eksperyment wykonano na danych o różnej korelacji.
- Badano braki typu MCAR i MAR

Wyniki eksperymentu



Rysunek 2: Moc testu dla braków MCAR i MAR [Davey and Savla, 2009]

- Braki MCAR mają mniejszy wpływ na testy niż MAR.
- Silna korelacja uodparnia na straty danych.

Postępowanie z brakami

- Usuwanie
  - Usuwanie rekordów lub zmiennych z brakami
- Imputacja
  - Zastąpienie braków nowymi wartościami
- Pomijanie
  - Stosowanie metod analizy pomijających braki
- Przegląd metod radzenia sobie z brakami można znaleść w [Graham, 2009].

Usuwanie danych

- Usuwanie eliminuje elementy puste z analizy.
- Może odbywać się poprzez usuwanie całych rekordów.
- Niektóre źródła utorzsamiają Usuwanie z Pomijaniem, które ogranicza się do niewykorzystywania brakujących elementów w analizach.

Usuwanie całych rekordów

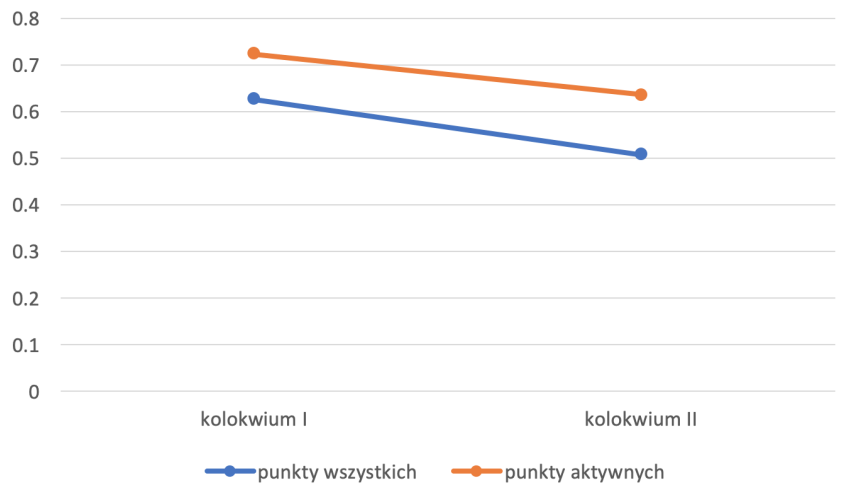
- Jedna z grup metod usuwania zakłada, że w dalszej analizie używamy tylko kompletne rekordy.
- Wszystkie rekordy z brakami są ignorowane.
- Stosowanie tych metod jest dopuszczalne tylko jeżeli braki dotykają stosunkowo niewielkiej liczby rekordów.
- Algorytmy:
  - Usuwanie według listy,
  - Usuwanie według listy z wyważaniem.

Usuwanie według listy

- Usuwanie według listy *List-Wise Deletion*.
- Najpopularniejsza metoda.
- Usuwamy wszystkie rekordy z brakami.
- Podejście jest zasadne tylko jeżeli braki są typu MCAR.
- W innym wypadku metoda może prowadzić do uzyskania tendencyjnych wyników.

Przykład usunięcia rekordów

- Na przedmiocie Teoria Automatów i Języków można było otrzymać od 0 do 12 punktów za kolokwium.
- Jednocześnie można otrzymać do dwóch punktów za aktywność na zajęciach.
- Brak aktywności oznaczany jest jako wartość pusta (NULL).
- Jak na średnie wyniki kolokwiów wpłynie usunięcie rekordów zawierające wartości puste?



Rysunek 3: Średni uzyskany procent punktów z kolokwiów. Wyniki dla wszystkich rekordów i po usunięciu rekordów z pustymi wartościami.

Usuwanie według listy z wyważaniem

- Usuwanie według listy z wyważaniem *List-Wise Deletion With Weighting*.
- Usuwamy wszystkie rekordy z brakami ale stosujemy metody naprawcze mające zapobiegać tendencyjności.
- Jeżeli usunięto wielu przedstawicieli jednej z klas to nadajemy wagi pozostałym rekordom z tej klasy, aby zrównoważyć ją z innymi.
- Inna metoda polega na dodaniu nowej zmiennej do każdej z klas, która określa szacowany brak odpowiedzi [Heckman, 1979]

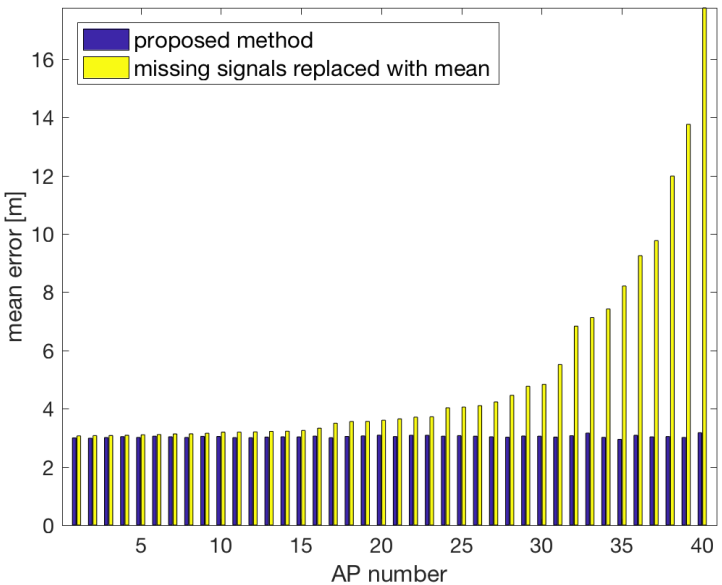
Usuwanie zmiennych

- Jeżeli uważamy, że braki będą szkodliwe dla analizy to możemy usunąć kolumny zawierające braki danych.
- Metodę stosujemy w sytuacjach gdy zmienna zawiera wiele braków.
- Możemy stosować metodę w sposób selektywny, to znaczy stworzyć podzbiory danych poprzez selekcję kolumn w taki sposób, aby w każdym podzbiorze nie było braków.
- Dalszą analizę prowadzimy osobno dla każdego podzbioru.

Analiza braków źródeł sygnałów

- Rozpatrzmy zbieranie danych z 40 punktów dostępowych.
- Dane mają posłużyć do budowy lokalizatora, ale występują w nich braki.
- Braki usuwamy przy pomocy dwóch algorytmów
  - Uzupełnianie wartością średnią wyliczaną dla danego źródła [Saleem and Wyne, 2016].
  - Usuwając kolumnę z brakującymi danymi [Górak and Luckner, 2018].
- Zmodyfikowane dane zostały wykorzystane do budowy lokalizatora.

Porównanie metod



Rysunek 4: Porównanie metod usuwania braków

- Średnia różnica między metodami przekracza 2 m.
- Największa różnica wynosi 14 m.
- Usuwanie kolumny sprawdziło się lepiej niż zastępowanie braków średnią.

Imputacja

- Zadaniem imputacji jest odtworzenie brakujących danych na podstawie podobieństwa do innych danych.
- W przeciwieństwie do pozostałych metod imputacja tworzy nowe dane.
- Z tego powodu, przez dłuższy czas metody imputacji nie cieszyły się zaufaniem badaczy.
- Wyróżnia się:
  - Pojedynczą imputację,
  - Wielokrotną imputację.

Pojedyncza imputacja

- W pojedynczej imputacji brakujące dane są zastępowane wartościami wyliczonymi z istniejących wartości.
- Popularnymi podejściami są:
  - Zastąpienie średnią.
    - Braki uzupełniamy wartością średnią wyliczoną dla zmiennej z brakami.
  - Podstawienie.
    - Jeżeli istnieją dwa wiersze o takich samych (lub zbliżonych) wartościach zmiennych, a jeden z nich ma braki, to zastępujemy je wartościami z drugiego wiersza.
  - Regresja.
    - Wyliczamy wartość brakującej zmiennej modelem regresji zbudowanym na podstawie wartości pozostałych zmiennych i zweryfikowanym na niebrakujących wartościach modelowanej zmiennej.
- W zasadzie żadna z tych metod nie jest polecana.

Wielokrotna imputacja

- Wielokrotna imputacja [Rubin, 1987] działa według następującego schematu.
  1. Zamiast pojedynczej wartości imputacji generujemy m wartości według danego rozkładu.
  2. Tworzymy m kompletnych zbiorów danych uzupełnionych wygenerowanymi wartościami i analizujemy ich właściwości.
  3. Uzyskane wyniki z m zbiorów są konsolidowane, aby utworzyć zbiór wyjściowy.
- Szczegóły działania powyższego schematu zależą od implementacji.
- Przykładem aplikacji jest algorytm MICE (*Multivariate imputation by chained equations*).

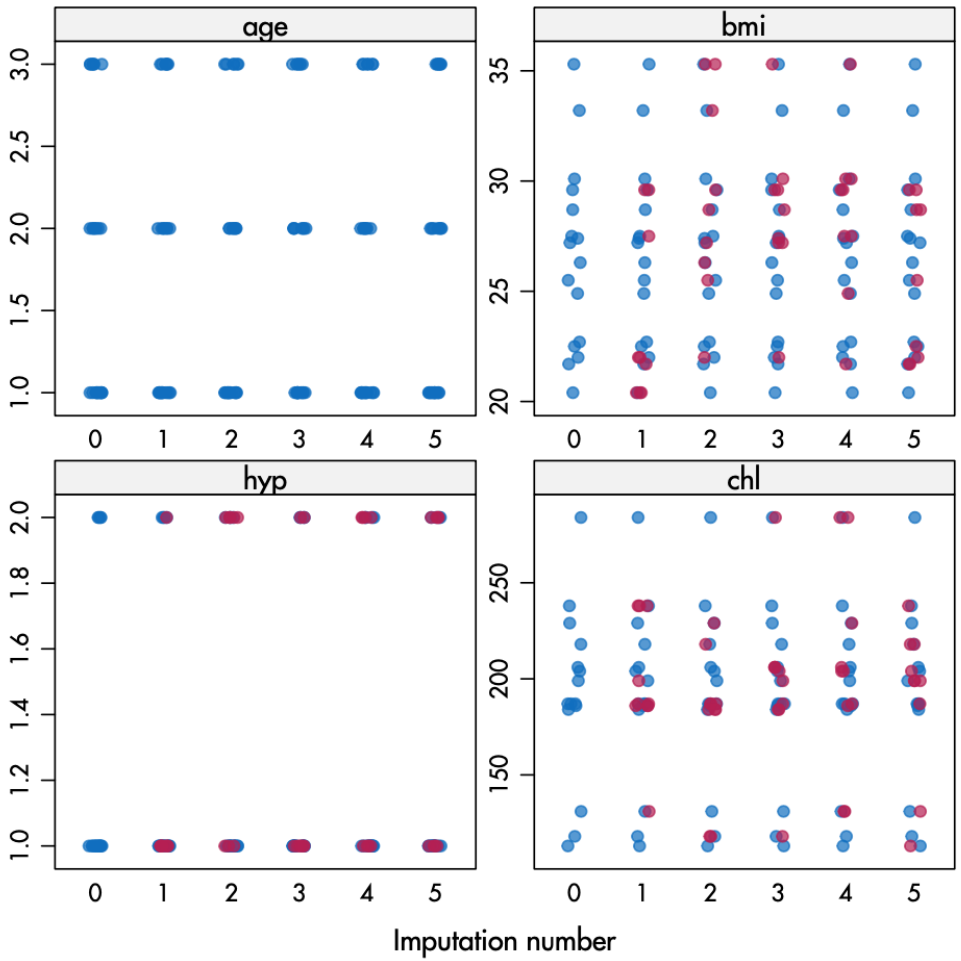
Algorytm MICE

1. Wszystkie brakujące wartości są zastępowane pojedynczą imputacją.
2. Dla wybranej zmiennej imputowane wartości są ponownie zastępowane brakami.
3. Tworzony jest model regresyjny dla wartości wybranej zmiennej w oparciu o wartości pozostałych zmiennych.
4. Braki zmiennej są zastępowane wynikami modelu regresyjnego.
5. Kroki 2-4 są powtarzane raz dla każdej zmiennej z brakami.
6. Kroki 1-5 są powtarzane przez ustaloną liczbę cykli.

Uwagi do algorytmu

- Sugerowaną liczbą cykli jest 10.
- Można także badać stabilność rozwiązań obserwując zmiany parametrów modelu regresyjnego.
- Metoda pozwala na stosowanie różnych modeli regresji.
  - liniowy,
  - logistyczny,
  - Poisona.

Imputacja MICE



Rysunek 5: Imputacja MICE. Czerwone punkty powstały w skutek imputacji [van Buuren and Groothuis-Oudshoorn, 2011]

Analiza danych z brakami

- Kolejna grupa metod zakłada, usuwanie braków poprzez nie włączanie ich do analizy.
- Oznacza to, że możemy korzystać z rekordów zawierających braki o ile nie korzystamy z brakujących elementów.
- Stosowanie tych metod powinno być ograniczone do sytuacji gdy braki nie są liczne.
- W mniejszym stopniu niż poprzednie metody skutkuje tendencyjnością danych.
- Algorytmy:
  - Pair-Wise,
  - Expectation Maximization Algoritm,
  - Full Information Maximum Likelihood.

Usuwanie według par

- Usuwanie według par *Pair-Wise*.
- Popularna metoda radzenia sobie z brakami.
- Wszystkie podstawowe momenty (średnie, wariancje, kowariancje) są wyliczane na podstawie rekordów, dla których istnieje para zmiennych.

Problemy

- Załóżmy, że mamy obliczyć kowariancję między dwoma zmiennymi zawierającymi braki

$$cov(X, Y) = \sum_{i=1}^{n} \frac{(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

- Które elementy będą użyte do wyliczenia kowariancji?

Przykład

$$X = \begin{bmatrix} 1 \\ NULL \\ 3 \\ -7 \\ 5 \\ 6 \end{bmatrix}, \; Y = \begin{bmatrix} 6 \\ -7 \\ 4 \\ NULL \\ 2 \\ 1 \end{bmatrix}$$

- Wyliczmy kowariancję X i Y korzystając z wszystkich dostępnych danych dla każdej ze zmiennych.
  - Otrzymujemy dodatnią kowariancję: $cov(X, Y) = 0.72$.
- Wyliczmy kowariancję X i Y korzystając tylko z wierszy zawierających obydwie zmienne.
  - Otrzymujemy ujemną kowariancję: $cov(X, Y) = -3.69$.

Algorytm EM

- Algorytm EM *Expectation Maximization algorithm* [Dempster et al., 1977] jest dwukrokowym algorytmem iteracyjnym służącym do estymacji podstawowych momentów.
- W pierwszym kroku E (*Expectation*) brakujące wartości są zastępowane przez ich wartości oczekiwane względem innych zmiennych modelu.
- W drugim kroku M (*Maximization*) maksymalizujemy prawdopodobieństwo, że estymowane wartości są brakującymi wartościami.

Opis rozwiązywanego problemu

- Y jest obserwowaną wielowymiarową zmienną losową z rodziną rozkładów prawdopodobieństwa z parametrem θ.
- Interesuje nas odkrycie wartości odpowiadającej maksimum rozkładu prawdopodobieństwa.
- W tym celu zakładamy, że istnieje ukryta zmienna Z od której zależy Y.
- Z rozkładu warunkowego otrzymujemy wyrażenie do maksymizacji

$$\Pr_{\theta}(Y) = \frac{\Pr_{\theta}(Z, Y)}{\Pr_{\theta}(Z \mid Y)}$$

Idea rozwiązania

- Korzystając z funkcji log-warygodności przekształcamy problem

$$l_\theta(Y) = l_{0,\theta}(Z, Y) - l_{1,\theta}(Z \mid Y).$$

gdzie l_θ, $l_{0,\theta}$ i $l_{1,\theta}$ są różne, gdyż dotyczą innych zmiennych
- Nakładając obustronnie wartość oczekiwaną $E(. \mid Y, \theta')$ dla suboptymalnych parametrów θ' otrzymujemy

$$l_\theta(Y) = E(l_{0,\theta}(Z, Y) \mid Y, \theta') - E(l_{1,\theta}(Z \mid Y) \mid Y, \theta')$$

- Upraszczając notację

$$l_\theta(Y) = Q(\theta, \theta') - R(\theta, \theta')$$

- Algorytm EM ogranicza się do maksymalizacji wyrażenia $Q(\theta, \theta')$, aby pośrednio zmaksymalizować $l_\theta(Y)$.
- Można dowieść, że mimo elementu $R(\theta, \theta')$, zwiększenie $Q(\theta, \theta')$ nie zmniejszy $l_\theta(Y)$.

Kroki algorytmu

1. Wybierz początkową wartość wektora parametrów $\hat{\theta}^{(1)}$.
2. Krok E (*Expectation*). Wyznacz warunkową wartość oczekiwaną dla aktualnego $\hat{\theta}^{(j)}$

$$Q(\theta, \hat{\theta}^{(j)}) = E\left(l(\hat{\theta}, Z, Y) \mid Y, \theta^{(j)}\right).$$

3. Krok M (*Maximization*). Wyznacz kolejną wartość $\hat{\theta}^{(j+1)}$, taką która maksymalizuje

$$\hat{\theta}^{(j+1)} = \arg\max_{\theta} Q(\theta, \hat{\theta}^{(j)}).$$

4. Powtarzaj kroki 2-3 do spełnienia określonego warunku stopu.

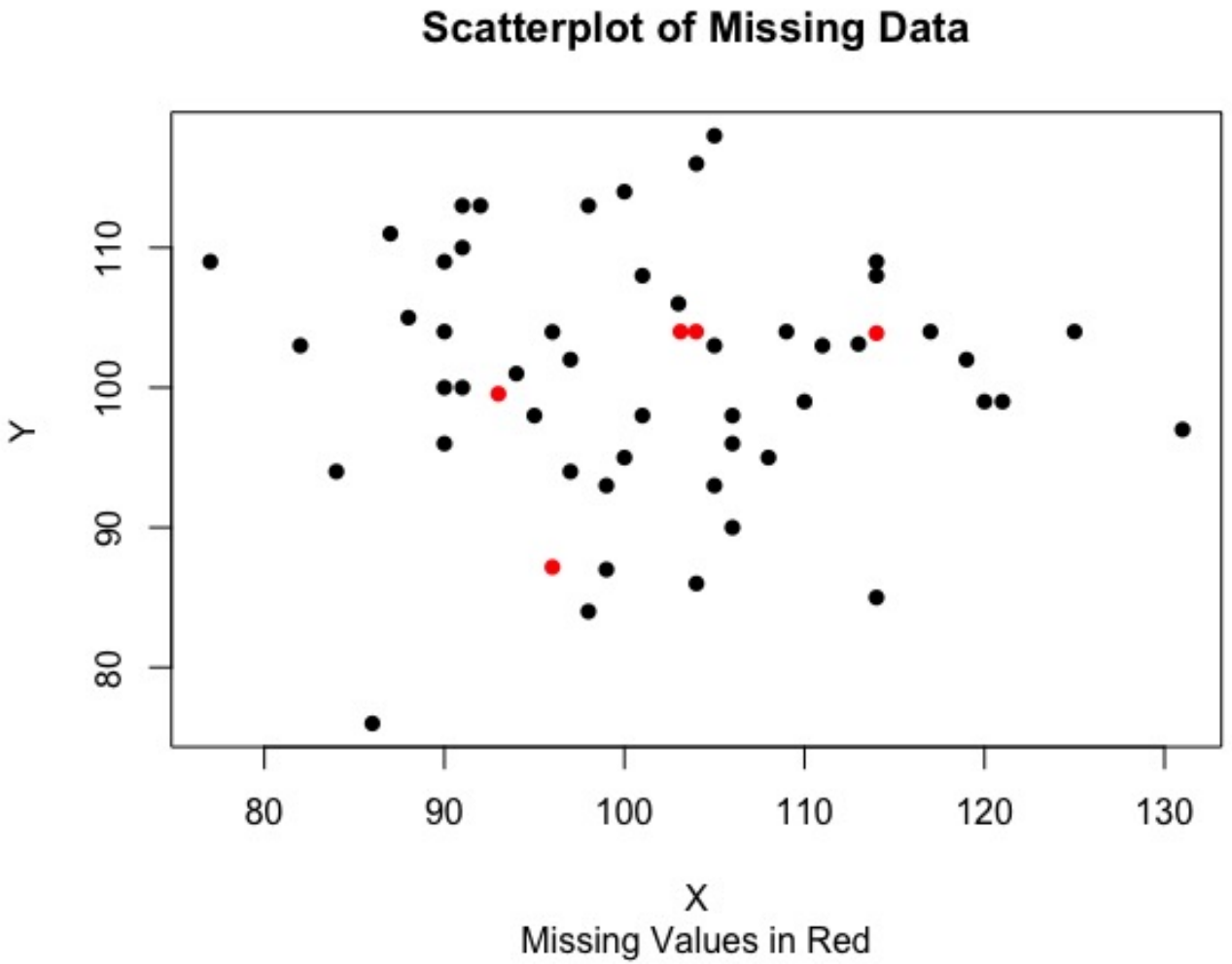
Uwagi do algorytmu

- Warunek stopu jest zależny od zmiany Q

$$Q(\theta, \hat{\theta}^{(j+1)}) - Q(\theta, \hat{\theta}^{(j)}) \leqslant \epsilon.$$

- Algorytm EM ma znacznie szersze zastosowania i jest używany w klasteryzacji i badaniu mieszanin rozkładów.
- W przeciwieństwie do imputacji średnią algorytm może wyznaczać różne wartości w miejsce braków tej samej zmiennej.

Imputacja algorytmem EM



Rysunek 6: Przykład imputacji braków za [Wesley, 2013]

Drzewa decyzyjne, a brakujące wartości

- Drzewa decyzyjne mają wbudowane mechanizmy pracy z brakami danych.
- W zależności od rodzaju drzewa braki są traktowane w inny sposób.
- Odmienne mechanizmy zaimplementowano w algorytmach
  - CHAID,
  - CART,
  - C4.5.

Drzewa CHAID

- Drzewa CHAID budują swoją strukturę na podstawie wartości dyskretnych.
  - Jeżeli zmienna jest ciągła to są one rzutowane na zmienne dyskretne.
- Testy podziału odbywają się na podstawie statystyk.
  - F-teście w wypadku regresji.
  - χ^2 w wypadku klasyfikacji.
- W związku z powyższym mogą traktować brakujące wartości jako osobną wartość, która jest brana pod uwagę przy wybieraniu warunku podziału.

Drzewa CART

- Drzewa CART tworzą dla wartości pustych podziały zastępcze *surrogate split*.
- Podczas oceny jakości podziału względem danej zmiennej obserwacje z wartościami pustymi tej zmiennej są ignorowane.
- Podczas klasyfikacji przypadku, jeżeli trafimy na warunek oparty na zmiennej, której brakuje, zastępujemy decyzję o przypisaniu decyzją zastępczą opartą na zmiennej, która nie jest pusta.
- Spośród dostępnych podziałów zastępczych wybiera się taki, który najlepiej oddaje rozkład podstawowego podziału.

Drzewa C4.5

- Drzewo C4.5 budowane jest na podstawie zachłannego przeszukiwania przestrzeni testów które maksymalizują heurystyczne kryteria podziału [Kohavi and Quinlan, 1999].
- Drzewa zniechęcają do tworzenia testów w oparciu o cechy zawierające braki.
- Jeżeli podczas uczenia dojdzie jednak do testu opartego o wartość pustą obserwacja jest dzielona pomiędzy podzbiory.
- W wyniku tego, podczas czasie klasyfikacji, traktujemy decyzję drzewa jako rezultat mieszaniny prawdopodobieństwa.

Przyrost informacji

- Informację opisaną przez przypadki ze zbioru S, należące do klas C_i gdzie $i = 1 \ldots x$ opisujemy jako

$$I(S) = -\sum_{j=1}^{x} RF(C_j, S) \log(RF(C_j, S)),$$

gdzie $RF(C_j, S)$ jest względną frekwencją klasy C_j w zbiorze S.

- Przyrost informacji po podziale zbioru S na t zbiorów $S_1, S_2, \ldots, S_t$, przez test B jest dany jako

$$G(S, B) = I(S) - \sum_{i=1}^{t} \frac{|S_i|}{|S|} I(S_i).$$

- Drzewo wybiera test B maksymalizując $G(S, B)$

Stosunek przyrostu

- Kryterium przyrostu informacji faworyzuje podział na wiele podzbiorów.
  - Kryterium osiągnie maksymalną wartość, gdy każdy zbiór S_i zawiera dokładnie jeden przypadek.
- Z tego powodu wprowadza się dodatkową heurystykę

$$P(S, B) = -\sum_{i=1}^{t} \frac{|S_i|}{|S|} \log\left(\frac{|S_i|}{|S|}\right).$$

- Drzewo wybiera test B maksymalizując $G(S, B)/P(S, B)$.

Wpływ brakujących wartości

- Niech zbiór $S_0 \subset S$ oznacza obserwacje z pustymi wartościami zmiennej będącej podstawą testu B.
- Informacja otrzymana z podziału B jest mniejsza, bo nie czerpiemy wiedzy z S_0

$$G(S, B) = \frac{|S - S_0|}{|S|} G(S \setminus S_0, B)$$

- Podobnie modyfikujemy $P(S, B)$ uwzględniając wpływ S_0

$$P(S, B) = -\frac{|S_0|}{|S|} \log\left(\frac{|S_0|}{|S|}\right) - \sum_{i=1}^{t} \frac{|S_i|}{|S|} \log\left(\frac{|S_i|}{|S|}\right).$$

- Obydwie zmiany powodują, że testy oparte o zmienną z dużymi brakami danych są mniej atrakcyjne.

Przypisanie elementu z brakami do węzła

- Drzewa C4.5 stosują metodę cząstkowych wystąpień *fractional instances*.
- Jeżeli wybrany zostane test w wartościami pustymi i natrafimy na wartość pustą to traktujemy ten przypadek jako wpadający częściowo do wszystkich potomków węzła.
- Wagi dla takiego podziału dobieramy na podstawie liczności potomków węzła $\frac{|S_i|}{|S-S_0|}$.

Klasyfikacja

- W czasie klasyfikacji nie przeprowadza się przypisania przypadku Y do pojedynczej klasy.
- W zamian wylicza się prawdopodobieństwo $CP(T, Y)$ przynależności Y do klasy w wyniku decyzji drzewa T.
  - Jeżeli T jest liściem $CP(T, Y)$ jest względną frekwencją klasy w zbiorze uczącym.
  - Jeżeli można przeprowadzić test B w korzeniu T (wartość cechy jest znana) to

$$CP(T, Y) = CP(T_i, Y)$$

    gdzie T_i jest poddrzewem wskazanym w teście B.
  - Jeżeli nie można przeprowadzić test B w korzeniu T (wartość cechy nie jest znana) to

$$CP(T, Y) = \sum_{i=1}^{t} \frac{|S_i|}{|S - S_0|} CP(T_i, Y)$$

    - Przypadek Y trafił do wszystkich podzbiorów.

Bibliografia I

[Davey and Savla, 2009] Davey, A. and Savla, J. (2009).

Statistical power analysis with missing data: A structural equation modeling approach.

Routledge Taylor & Francis Group.

[Dempster et al., 1977] Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977).

Maximum likelihood from incomplete data via the em algorithm.

Journal of the Royal Statistical Society. Series B (Methodological), 39(1):1–38.

[Górak and Luckner, 2018] Górak, R. and Luckner, M. (2018).

Automatic detection of missing access points in indoor positioning system.

Sensors, 18(11):3595.

Bibliografia II

[Graham, 2009] Graham, J. W. (2009).
Missing data analysis: Making it work in the real world.
Annual Review of Psychology, 60(1):549–576.
PMID: 18652544.

[Heckman, 1979] Heckman, J. J. (1979).
Sample Selection Bias as a Specification Error.
Econometrica, 47(1):153–161.

[Kohavi and Quinlan, 1999] Kohavi, R. and Quinlan, R. (1999).
Decision Tree Discovery.
in Handbook of Data Mining and Knowledge Discovery, 3(Hunt 1962):267—-276.

[Little and Rubin, 2002] Little, R. J. A. and Rubin, D. B. (2002).
Statistical Analysis with Missing Data.
Wiley.

Bibliografia III

[Rubin, 1987] Rubin, D. B. (1987).

Multiple Imputation for Nonresponse in Surveys.

Wiley.

[Saleem and Wyne, 2016] Saleem, F. and Wyne, S. (2016).

Wlan–based indoor localization using neural networks.

Journal of Electrical Engineering, 67(4):299–306.

[van Buuren and Groothuis-Oudshoorn, 2011] van Buuren, S. and Groothuis-Oudshoorn, K. (2011).

mice: Multivariate imputation by chained equations in r.

Journal of Statistical Software, Articles, 45(3):1–67.

[Wesley, 2013] Wesley (2013).

Imputing missing data with expectation – maximization.