Metody imputacji danych

Bąkała, Pastuszka, Pysiak

Rozkład jazdy

1. Przedstawienie problemu
2. Losowość braku danych
3. Rozwiązania pomijające imputację
4. Podstawowe techniki imputacji
5. Zaawansowane techniki imputacji
6. Bardzo bardzo zaawansowane techniki imputacji

Czy braki danych to problem?

- Obniżenie poprawności modeli
- Odrzucenie danych przez niektóre modele
- Większa trudność analizy danych
- ...

“Tak, no tak, nie wiem co to niby zmienia, ale tak.” ~ Otis, ok. 30 r. p.n.e.

Losowość braku danych

- MCAR (Missing Completely at Random) - braki danych nie zależą od znanych ani brakujących wartości

- MAR (Missing at Random) - braki danych zależą tylko od znanych wartości

- MNAR (Missing Not at Radom) - braki danych zależą od znanych oraz brakujących wartości

Co jeśli nie chcemy imputować?

Możliwe rozwiązania:

- Pozostawienie wszystkich obserwacji.
  Algorytm sam zajmie się obsługą braków danych.
- Usunięcie wszystkich obserwacji z brakami danych.
  Zmniejsza jakość predykcji i wprowadza obciążenie.
- Usunięcie obserwacji z brakami danych w analizowanych kolumnach.
  Ulepszona wersja poprzedniego rozwiązania.

Imputacja średnią

- Zastąpienie wszystkich wystąpień braków danych średnią wartością danej zmiennej.
- Możliwa dowolna funkcja uśredniająca, zwykle jednak średnia arytmetyczna bądź mediana.
- Prosta konceptualnie i obliczeniowo.
- Działanie wyłącznie na zmiennych ilościowych.
- Ignorowanie korelacji pomiędzy zmiennymi.

Imputacja modą lub stałą

- Ewolucja poprzedniego rozwiązania.
- Zwykle stosowaną stałą jest jakiś rodzaj zera.
- Możliwe stosowanie również na zmiennych kategorycznych.
- Także ignorowanie korelacji między zmiennymi.
- Prawdopodobne wprowadzenie (niechcianego) obciążenia.

- Imputacja braku danych w obserwacji wartością innej, losowo wybranej spośród podobnych obserwacji.
- Dobre działanie przy mocno skorelowanych zmiennych (ale wtedy można wyrzucić problematyczne zmienne).
- Możliwe wartości ograniczone do występujących w zbiorze.
- Problematyczne przy małym podobieństwie obserwacji.

Imputacja k-NN

- Ewolucja poprzedniej metody.
- Imputowana wartość jako średnia ważona wartości podobnych obserwacji, zależna od podobieństwa między obserwacjami.
- Znaczny wzrost złożoności obliczeniowej względem poprzednich rozwiązań.
- Podatna na pozależe*.

*) outliery [przyp. tłumacza]

Imputacja regresją

- Przewidywanie brakującej wartości na podstawie pozostałych zmiennych
- Pozwala uwzględnić zależność między zmiennymi
- Wariant - stochastyczna imputacja regresją - wprowadzenie losowości
- Zaleta - nie jest ograniczona przez istniejące wartości zmiennej
- Wada - nie jest ograniczona przez istniejące wartości zmiennej

Predictive Mean Matching

- Korzysta z regresji w celu wyznaczenia podobnych obserwacji
- Zabezpiecza przed występowaniem wartości spoza oczekiwanego zakresu poprzez imputacje istniejącymi wartościami
- Uwzględnia element losowości, niedeterministyczna
- Zawarta w pakiecie `mice`

Imputacja MICE

Multivariate imputation by chained equations

- Meta-metoda iteracyjna, wykorzystująca pod spodem model regresyjny, np: predictive mean matching
- Przydatna, gdy braki danych występują w wielu kolumnach
- Niedeterministyczna, pozwala uwzględnić niepewność przewidywanej wartości poprzez wygenerowanie kilku zbiorów
- Dostępna w R w pakiecie `mice`

Imputacja uczeniem głębokim

- W wielu przypadkach daje dobre rezultaty
- Może działać wolno dla dużych zbiorów
- Czarna skrzynka
- Przykład - biblioteka `datawig`

Bibliografia

1. 6 Different Ways to Compensate for Missing Values In a Dataset (Data Imputation with examples), Will Badr
2. Tutorial: Introduction to Missing Data Imputation, Cambridge Spark
3. https://www.youtube.com/watch?v=XnnA9z7lv4Q
4. https://datascienceplus.com/imputing-missing-data-with-r-mice-package/