Zadanie 2 – Braki w danych

Eksploracja danych PS1

Artur Tagisow

Tomasz Kołtun

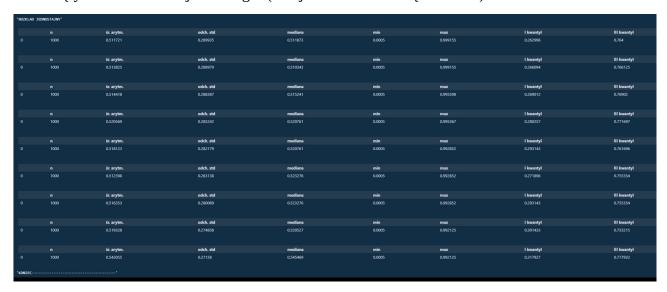
1. Oblicz dla próbki podstawowe statystyki - średnią, odchylenie standardowe, min, max, medianę, kwartyle



- 2. Na podstawie oryginalnej próbki stwórz serię próbek usuwając 10%, 20%, 30%, 40%, 50%, 60%, 70%, 80% danych (w każdej próbce powinny być usuwane te same elementy, tylko rozszerzamy zakres usuniętych obserwacji)
- 3. Oblicz podstawowe statystyki dla tych próbek (tak jak w pkt. 2.)
- Z usuniętymi dla rozkładu normalnego: (kolejno od 10% usunięć do 80%)



Z usuniętymi dla rozkładu jednolitego: (kolejno od 10% usunięć do 80%)



4. Wypełnij nowe próbki

- wartością średnią (z pkt. 4 odpowiednią dla każdej kolumny, np. dla kolumny z usuniętymi 20% danymi, liczymy średnią z wartościami pustymi i następnie tą średnią uzupełniamy braki w tej kolumnie)
- mediana
- wylosowanymi w sposób losowy nowymi wartościami (najlepiej z tego samego rozkładu jak dane oryginalne z pierwszej próbki)
- 5. Oblicz podstawowe statystyki dla wypełnionych danych i przeanalizuj
 - jak wpływa na statystyki brak danych (zwiększający się procent braków)
 - jak wpływa wypełnianie braków średnią, medianą
 - jak wpływa wypełnianie braków losowymi wartościami

Tabel danych jest tyle (3x8), że trudno byłoby je czytać, dlatego proszę o spojrzenie w załączony PDF z Jupyter Notebooka.

Wnioski

Widać, że między wszystkimi badanymi statystykami, najbardziej wrażliwe na zmiany przy wypełnianiu braków w danych są statystyki dotyczące rozkładu (I kwantyl, III kwantyl, odch. std)

Jak wpływa na statystyki brak danych (zwiększający się procent braków):

We wszystkich sprawdzanych statystykach (oprócz min i max) wyniki różnią się o tak małą ilość, że prawdopodobnie mieści się w granicy błędu statystycznego.

Podobnie w przypadku rozkładu jednolitego.

Jak wpływa wypełnianie braków średnią, medianą

Wypełnianie średnią sprawia że odchylenie standardowe staje się bardzo małe eg. $2.22 * 10^{-16}$ (zarówno dla rozkładu normalnego i jednolitego). Uzupełnianie medianą nie ma takiego efektu. Jest

to zrozumiałe, skoro składnikiem wzoru na odchylenie standardowe jest średnia arytmetyczna w całym zbiorze. Wypełnienie braków w danych właśnie średnią arytmetyczną sprawia, że dla wypełnionych próbek zachodzi ((średnia arytmetyczna) – (średnia arytmetyczna))^2, czyli 0.

Wypełnianie medianą wpływa najbardziej na wartość I i III kwantyla – sprawia, że wraz z zwiększeniem ilości ubytków (a zatem liczby wartości, które wypełniamy medianą) I i III kwantyl zbliżają się do siebie.

Jak wpływa wypełnianie braków losowymi wartościami

Wypełnianie braków losową wartością między min a max w danym zbiorze sprawia, że uzyskiwane podstawowe statystki są bardzo oddalone od przypadku bazowego.

Różnica między zbiorem jednostajnym a normalnym jest najbardziej widoczna w odchyleniu standardowym.

Najbardziej drastycznie zmienia się odchylenie standardowe w rozkładzie normalnym. Dzieje się tak, ponieważ losowe liczby nie respektują rozkładu normalnego.

Co ciekawe, mediana w przypadku rozkładu jednostajnego jest również mocno zmieniona. Po sporządzeniu sprawozdania zauważyliśmy, że wartości, którymi uzupełnialiśmy zbiór jednostajny mogły przyjmować tylko wartość 0 lub 1 (błąd w kodzie – `math.random(int(min) – int(max))`). Cu automatycznie sprawia, że uzyskany zbiór nie jest w rozkładzie jednostajnym, ponieważ wartości 0 i 1 mają o wiele większe prawdopodobieństwo wystąpienia..

Wraz z wrostem ubytku średnia arytmetyczna maleje, co wskazywałoby na to, że funkcja math.random w języku Python faworyzuje liczbę zero, gdy stoi przed wyborem między zero a jeden.