Podstawy Przetwarzania Danych

Wykład 2: Dobór typów i normalizacja danych

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

Wydział Matematyki i Nauk Informacyjnych

Wersja 1.2
18 października 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 zmiennych

- Przy przetwarzaniu danych wyróżniamy następujące typy zmiennych
  - Zmienne jakościowe
    - Przyjmują określoną liczbę wartości
    - Najczęściej wartości nieliczbowe
    - Zwane też wyliczeniowymi, czynnikowymi lub kategorycznymi
  - Zmienne ilościowe
    - Odpowiadają ciągłym zmiennym numerycznym

Typy zmiennych jakościowych

- binarne
  - dwuwartościowe
  - PRAWDA, FAŁSZ
- uporządkowane
  - można określić ich kolejność
  - ZIMA, WIOSNA, LATO, JESIEŃ
- nominalne
  - nieuporządkowane wartości
  - PW, UW, SGH, AWF

Typy zmiennych ilościowych

- zliczenia
  - liczba wystąpień pewnego zjawiska
  - liczba studentów na wykładzie
- ilorazowe
  - zmienne mierzone w skali
  - długość w metrach
- interwały
  - wyznaczanie długości przedziału
  - okres czasu

Określenie typu zmiennych

- Bazodanowe typy danych nie przekładają się jednoznacznie na typy zmiennych.
  - Dane numeryczne
    - W zależności od liczby wartości zmienna ilościowa lub jakościowa, uporządkowana
  - Dane tekstowe
    - Dane jakościowe
    - Jeżeli nie można określić wąskiej kategorii danych to nieprzetworzone dane tekstowe są nieużyteczne.
  - Dane logiczne
    - jakościowa, binarna
  - Data i czas
    - W zależności od kontekstu jakościowa uporządkowana lub ilościowa
    - Szereg czasowy $\longrightarrow$ zmienna ilościowa
    - Wydarzenie jednostkowe $\longrightarrow$ zmienna jakościowa

Progowanie

- Jeżeli chcemy przetwarzać dane ilościowe jako dane jakościowe to możemy spotkać się z problemem zbyt dużej liczby kategorii.
- Niektóre metody – takie jak drzewa decyzyjne – ze względów obliczeniowych ograniczają liczbę dopuszczalnych kategorii.
- Rozwiązaniem jest *progowanie* czyli zastąpienie zbioru wartości etykietą określającą przedział wartości.
- Szczególnym przypadkiem progowania jest *binaryzacja* czyli ograniczenie liczby wartości zmiennej do dwóch.
- Progi do kategoryzacji dobiera się na podstawie wiedzy eksperckiej.

Przykład progowania

- Sygnał w bezprzewodowej sieci Wi-Fi ma siłę z przedziału $[-90, 0)$ dBM.
- Korzystając z wiedzy eksperckiej [Metageek, 2019] możemy dokonać kategoryzacji według przedziałów.

Tabela 1: Kategorie siły sygnału Wi-Fi

Przedział [dBm]	Kategoria	Uwagi
(-90,-80]	Niedostateczna	Brak sygnału
(-80,-70]	Słaba	Brak zastosowań
(-70,-67]	Dobra	Poczta, strony WWW
(-67,-30]	Bardzo dobra	Streaming
(-30,0]	Niesamowita	Dowolne zastosowanie

Przykład binaryzacji

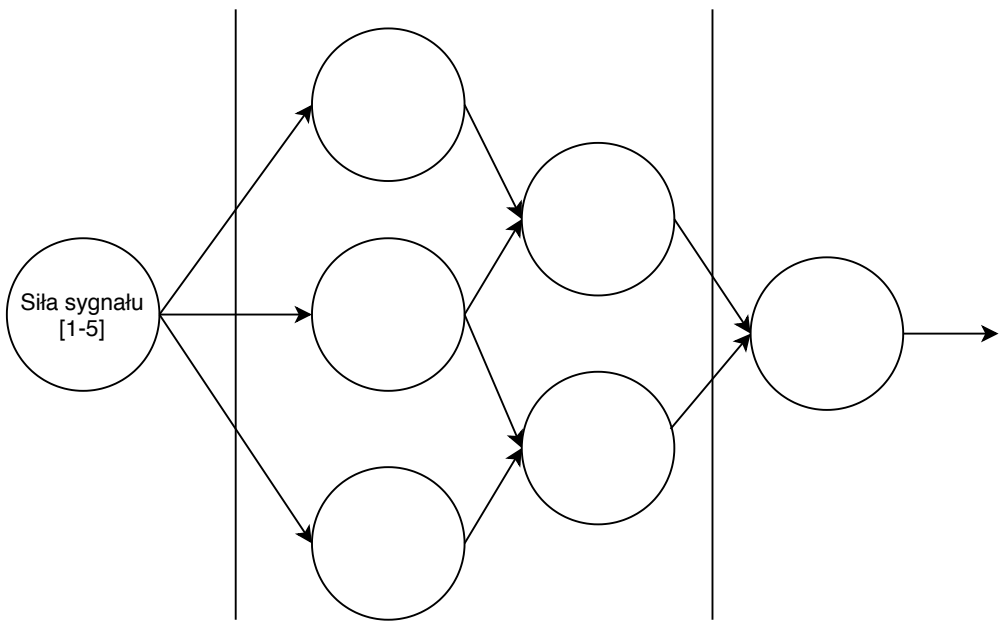
- Bazując na tej samej wiedzy eksperckiej możemy też dokonać binaryzacji siły sygnału Wi-Fi.

Tabela 2: Binaryzacja siły sygnału Wi-Fi

Przedział [dBm]	Kategoria
(-90,-80]	Brak sygnału
(-80,-0)	Sygnał obecny

Przetworzenie danych dla sieci neuronowych

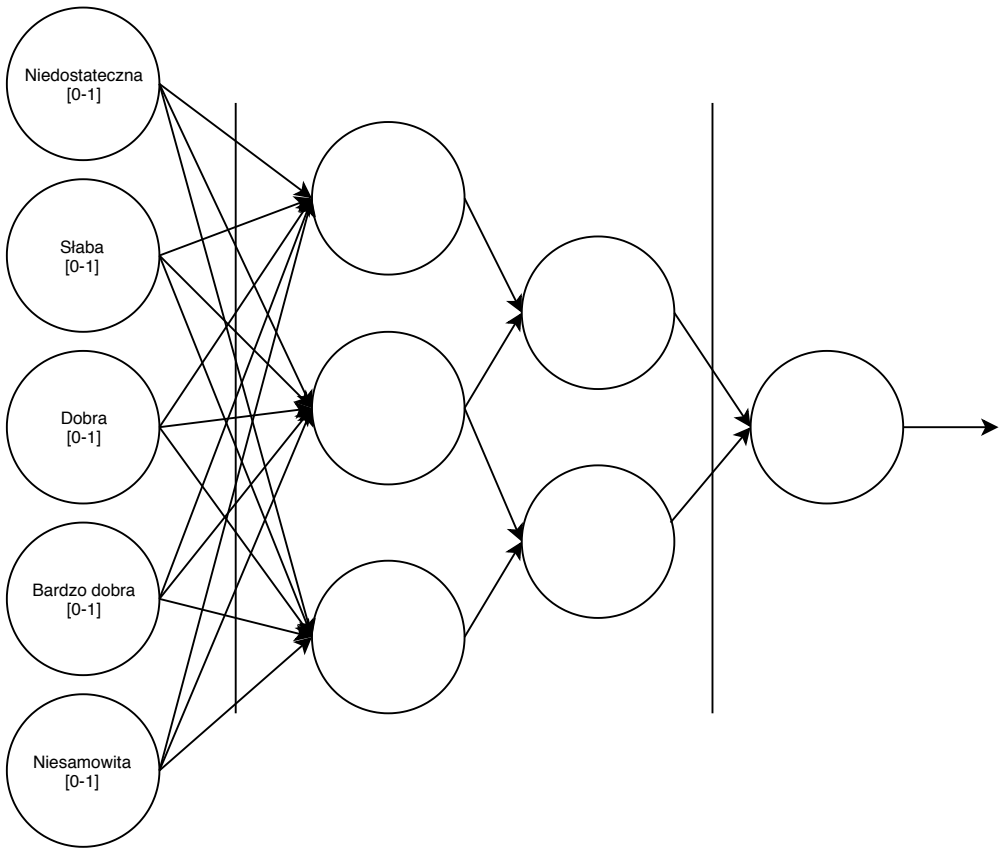
- Zmienna określająca siłę sygnału może przyjmować pięć wartości
  - Niedostateczna, Słaba, Dobra, Bardzo dobra, Niesamowita
- Można utworzyć z niej część wejścia dla sieci neuronowej, reprezentowane przez jeden neuron przyjmujący wartości 1-5.



Rysunek 1: Wejście jedno-neuronowe dla kategorii danych

Przetworzenie danych dla sieci neuronowych II

- Lepsze wyniki może dać stworzenie pięciu neuronów z których tylko jeden przyjmuje wartość 1, a pozostałe 0.
- Jest to kodowanie *one-hot*.



Rysunek 2: Wejście wielo-neuronowe dla kategorii danych

Przetwarzanie danych tekstowych

- Dane tekstowe mogą zostać w różnorodny sposób przetworzone do postaci liczbowej [Liu and Dong, 2018].
- Reprezentacja *Bag of Words* zakłada stworzenie wektora, którego kolumny odpowiadają poszczególnym słowom, a wartości są zliczeniem wystąpień.
- Można zliczać słowa, zapisywać ich częstotliwość (*Term frequency*) lub, po prostu, odnotowywać wystąpienie.
- Popularne jest też zliczanie sekwencji wyrazów (*biterms, triterms*).
- Podejścia strukturalne obejmują elementy gramatyki i rozróżniają sekwencje względem ścieżki wywodu.

Normalizacja

- Zakresy zmiennych na ogół różnią się od siebie.
- Może to powodować, że zmienne z większym zakresem będą miały nadmierny wpływ na wyniki.

Tabela 3: Dane atletów wykonujących test Coopera

Wiek	Wzrost [cm]	Waga [kg]	Test Coopera [m]
40	168	85	1840
13	146	27	2100
15	170	42	1800
41	165	67	1700

- Dane dotyczące wykonywania testu Coopera mają różne jednostki i zakresy.

Normalizacja min-max

- Normalizacja min-max (ang. *min-max normalization*) określa o ile dana wartość jest większa od minimalnej i skaluje tę różnicę przez zakres

$$\overline{X} = \frac{X - min(X)}{max(X) - min(X)}$$

- Znormalizowane wartości będą mieścić się w przedziale $[0, 1]$.

Znormalizowane dane

Tabela 4: Znormalizowane dane atletów wykonujących test Coopera

Wiek	Wzrost	Waga	Test Coopera
0.96	0.92	1.00	0.35
0.00	0.00	0.00	1.00
0.07	1.00	0.26	0.25
1.00	0.79	0.69	0.00

- Normalizacja częściowo utrudnia interpretację danych oderwanych od jednostek i skali.
  - W szczególności nieintuicyjna jest interpretacja zera.
- Ułatwia równocześnie wychwytywanie skrajnych odczytów.

Normalizacja min-max - ograniczenia

- Normalizacja min-max jest bardzo wrażliwa na obserwacje odstające.
- Jeżeli chociaż pojedyncza wartość odbiega znacznie od pozostałych odczytów to wpłynie ona na proces normalizacji i ujednolicenie pozostałych odczytów.
- Dodatkowym problemem jest praca na osobnych zbiorach uczącym i testowym. Jeżeli ustalimy parametry normalizacji na zbiorze uczącym, a zbiór testowy zawiera dane spoza jego zakresu to normalizacja spowoduje uzyskanie wartości spoza przedziału $[0, 1]$.

Wartość odstająca

- Do przeciętnych i słabych wyników testu Coopera dodajemy obserwację wyniku sportowca.

Tabela 5: Wyniki testów Coopera

Dane	1.6	1.4	1.6	1.7	1.4	1.0	1.6	1.7
Normalizacja	0.86	0.57	0.86	1.00	0.57	0.00	0.86	1.00
	1.6	1.4	1.6	1.7	1.4	1.0	1.6	**3.7**
Uczący	0.22	0.15	0.22	0.25	0.15	0.00	0.22	1.00
Testowy	0.86	0.57	0.86	1.00	0.57	0.00	0.86	3.93

- Jeżeli obserwacja była w zbiorze testowym to spowoduje zbliżenie się do siebie średnich wyników.
- Jeżeli obserwacja była w zbiorze uczącym to uzyskamy wyniki wykraczające poza przedział $[0, 1]$

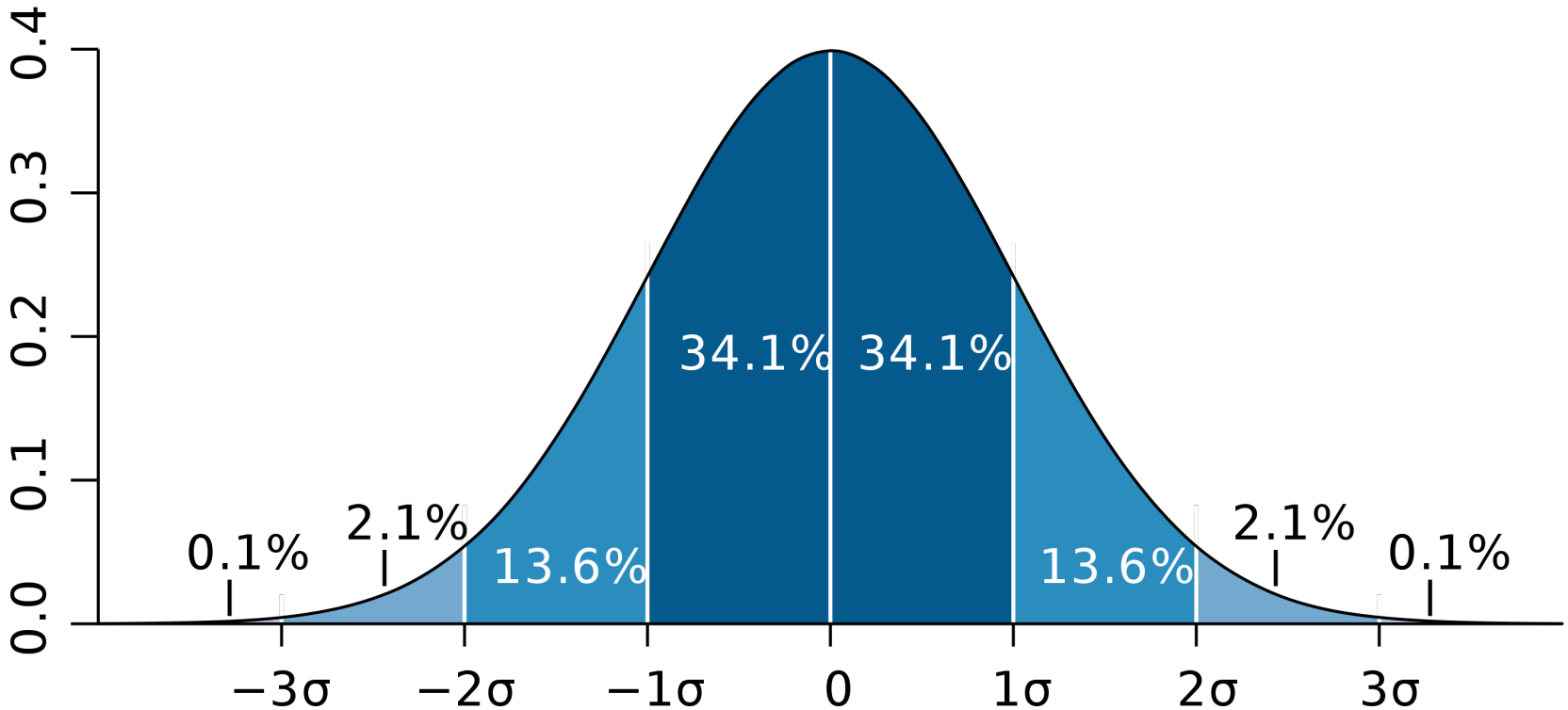
Standaryzacja

- Standaryzacja (ang. *Z-score standardization*) określa o różnicę między wartością, a wartością średnią w stosunku do odchylenia standardowego.

$$\overline{X} = \frac{X - \mu(X)}{\sigma(X)}$$

- Średnie wartości po standaryzacji przyjmą wartość zero, więc znak standaryzowanej zmiennej mówi czy jest ona większa, czy mniejsza od średniej.
- Wartości będą mieścić się w większości w przedziale $[-4, 4]$.

Rozkład danych po standaryzacji



Rysunek 3: Rozkład danych po standaryzacji [Wikipedia, 2019]

Ustandaryzowane dane

Tabela 6: Ustandaryzowane dane atletów wykonujących test Coopera

Wiek	Wzrost	Waga	Test Coopera
0.83	0.52	1.15	-0.12
-0.93	-1.47	-1.10	1.41
-0.80	0.70	-0.51	-0.35
0.90	0.25	0.46	-0.94

- Standaryzacja zależna jest od wartości średniej rozkładu, nie jest tak podatna na zaburzenia obserwacjami odstającymi jak normalizacja min-max.

Wartość odstająca

- Do przeciętnych i słabych wyników testu Coopera dodajemy obserwację wyniku sportowca

Tabela 7: Wyniki testów Coopera

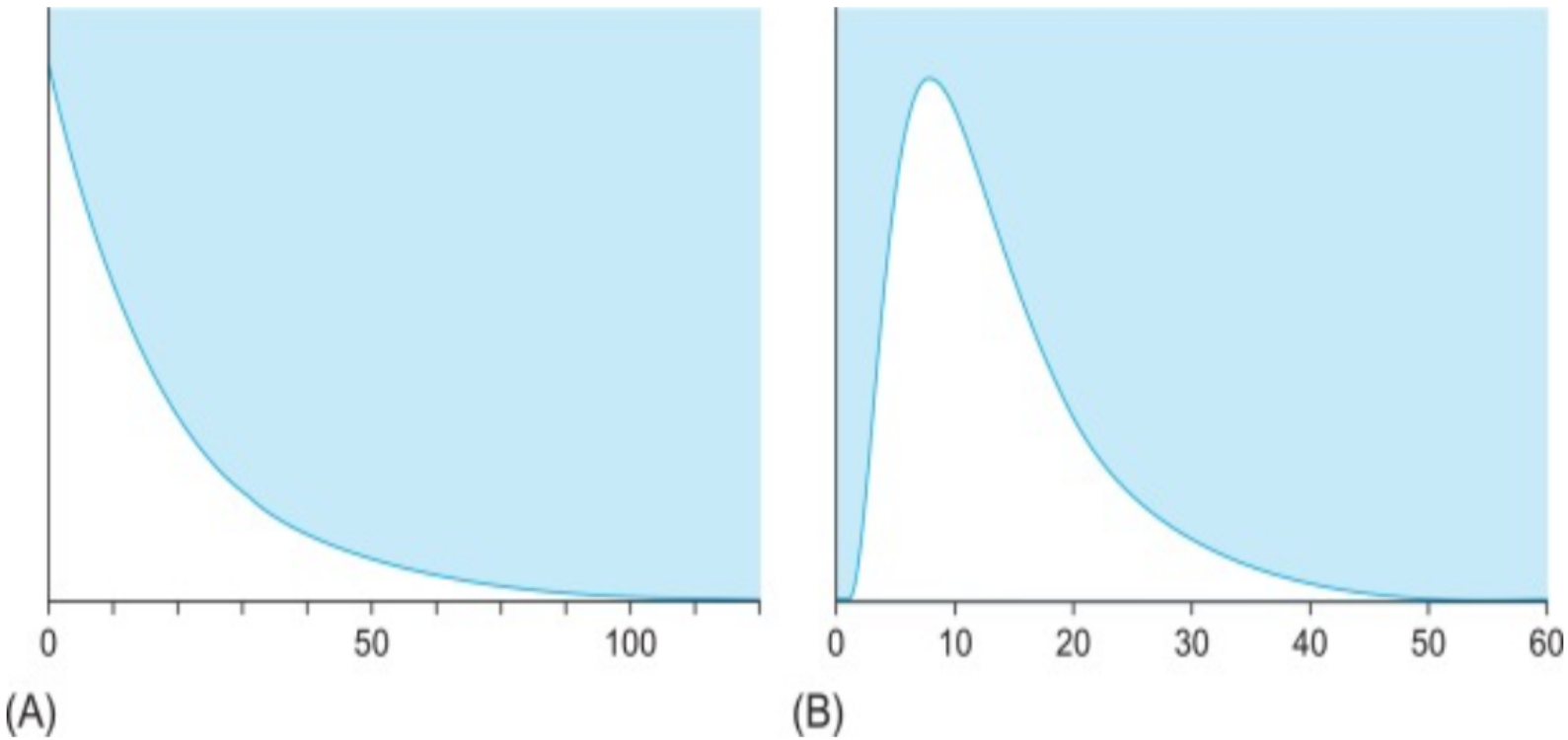
Dane	1.6	1.4	1.6	1.7	1.4	1.0	1.6	1.7
Standaryzacja	0.43	-0.43	0.43	0.86	-0.43	-2.15	0.43	0.86
	1.6	1.4	1.6	1.7	1.4	1.0	1.6	**3.7**
Uczący	-0.18	-0.43	-0.18	-0.06	-0.43	-0.92	-0.18	2.38
Testowy	0.43	-0.43	0.43	0.86	-0.43	-2.15	0.43	8.37

- Dodanie obserwacji do zbioru uczącego spowodowało, że wszystkie pozostałe obserwacje postrzegane są jako wyniki poniżej średniej.
- Dodanie obserwacji do zbioru uczącego informuje nas, że jest to obserwacja odstająca statystycznie występująca rzadziej niż raz na sto pomiarów.

Rozkład skośny

- Rozkład skośny jest niesymetryczną dystrybucją przypominającą rozkład normalny.
- W odróżnieniu do rozkładu normalnego, rozkład skośny charakteryzuje się ogonem, powstającym na skutek wolniejszego spadku jego wartości.
- Rozróżniamy rozkład prawostronnie skośny (z ogonem po prawej stronie) i lewostronnie skośny (z ogonem po lewej stronie). Który występuje częściej?

Rozkład prawostronnie skośny



Rysunek 4: Rozkład prawostronnie skośny [Siegel, 2016]

- Częściej występuje rozkład prawostronnie skośny, pojawiający się przy danych, które nie przyjmują ujemnych wartości.

Współczynnik skośności

- Istnieje kilka miar szacowania skośności, tzw. współczynników skośności
  - $A_d = \frac{\mu - d}{s}$
  - $A_m = 3\frac{\mu - m}{s}$
  - $A_Q = \frac{Q_1 + Q_3 - 2m}{Q_3 - Q_1}$
- gdzie
  - μ - średnia arytmetyczna
  - m - mediana
  - d - dominanta (moda)
  - Q_1, Q_3 - pierwszy i trzeci kwartyl
- Współczynnik skośności przyjmuje wartość zero dla rozkładu symetrycznego, ujemne dla lewostronnej skośności, dodatnie dla prawostronnej skośności.
- Różne miary mogą oceniać ten sam rozkład w różny sposób.

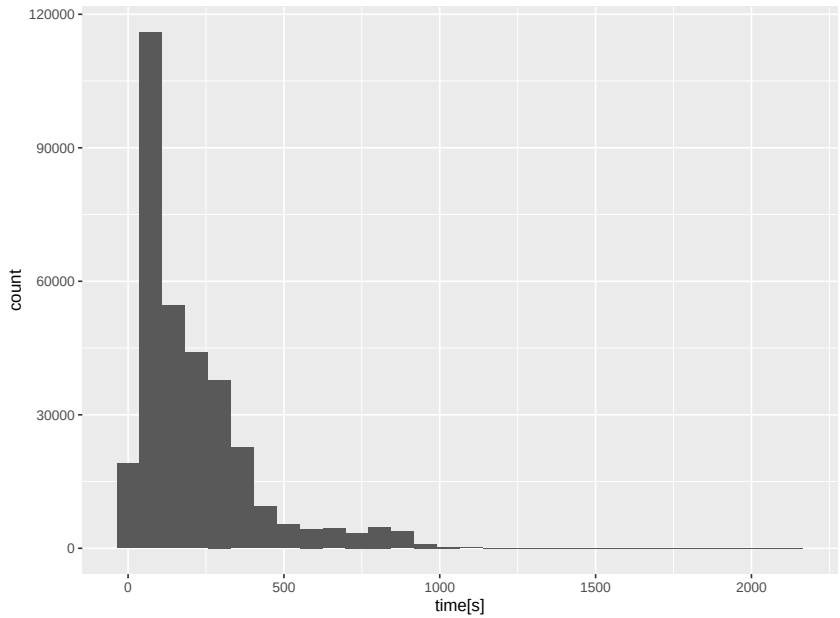
Problemy ze skośnością

- Większość metod statystycznych wymaga, aby rozkład wartości był rozkładem zbliżonym do normalnego.
- Użycie w ich przypadku rozkładu skośnego może prowadzić do błędnych wyników.
- W najlepszym wypadku mamy do czynienia z niewydajnością. Część informacji ze zbioru danych pozostaje niewykorzystana
  - Przy rozkładzie normalnym zakładamy symetryczność danych względem wartości oczekiwanej.
  - W wypadku rozkładu skośnego udającego rozkład normalny tracimy istotną informację o niesymetrycznym rozkładzie danych.

Usunięcie skośności

- W przypadku danych dodatnich możemy pozbyć się skośności poprzez zlogarytmowanie zmiennej.
- Logarytmowanie przekształca niesymetryczny rozkład w symetryczny, gdyż rozrzedza małe wartości i grupuje duże.
- Można równie dobrze stosować w celu usunięcia skośności ln jak i $\log_{10}$.

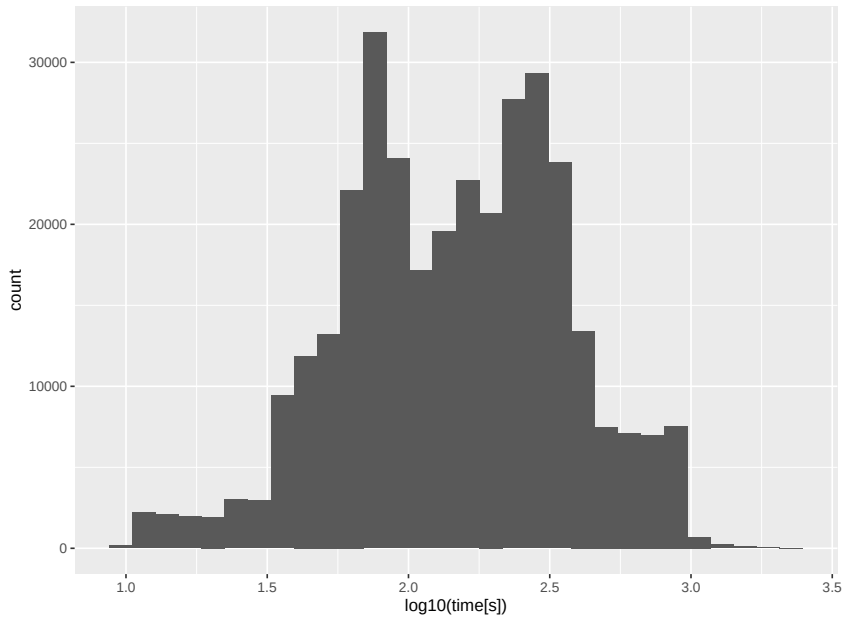
Czas przejazdu odcinka autostrady



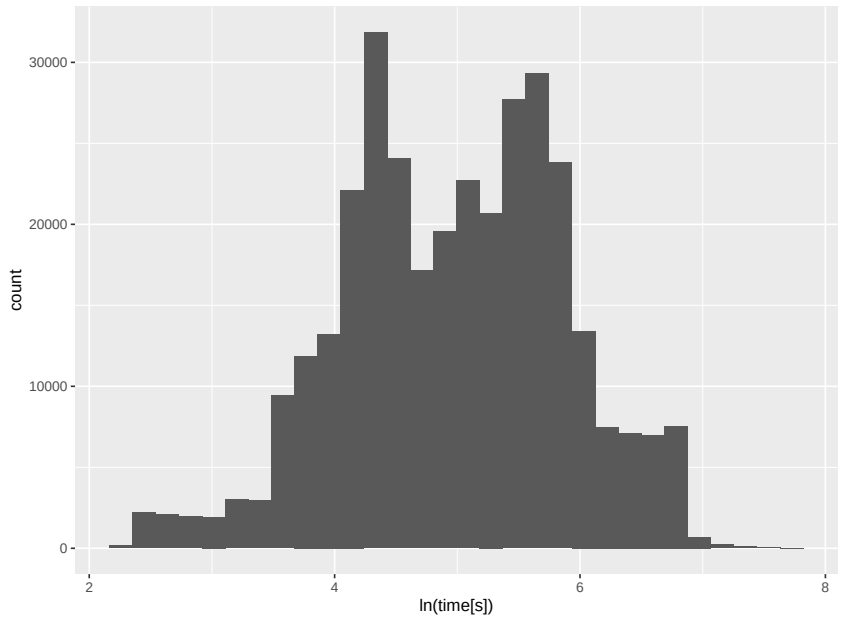
Rysunek 5: Rozkład czasu przejazdu przez odcinek autostrady

- Histogram przedstawia rozkład średniego czasu przejazdu odcinka brytyjskiej autostrady.

Logarytmizacja



Rysunek 6: Rozkład czasu po logarytmizacji $\log_{10}$



Rysunek 7: Rozkład czasu po logarytmizacji ln

Bibliografia I

[Liu and Dong, 2018] Liu, H. and Dong, G. (2018).

Feature Engineering for Machine Learning and Data Analytics.

CRC Press.

[Metageek, 2019] Metageek (2019).

Understanding wifi signal strength.

[Siegel, 2016] Siegel, A. F. (2016).

Practical Business Statistics.

[Wikipedia, 2019] Wikipedia (2019).

Rozkład normalny.

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.