KAD opracowany

Disclaimer:
Notatki na podstawie wykładów, studenckiego opracowania z lat poprzednich i strony GUS. Mogą pojawić się błędy, za które nie odpowiadam. Brakuje także średnich i wariancji dla szeregów rozdzielczych. Aczkolwiek ja osobiście dostałem 5 - Ajvar :D

(To z !!! na pewno było już na kolosie w 2025 lub 2026)
(To z !! jest względnie prawdopodobne, że się pojawi)

Szereg statystyczny - ciąg wielkości statystycznych zaobserwowanych wg określonego kryterium.

- nieuporządkowany - wg kolejności badania
- uporządkowany - wg wartości rosnących/malejących

Szereg rozdzielczy - szereg statystyczny opisujący informacje dotyczące badanej cechy danej populacji/próby, w którym wartości tej cechy są uporządkowane i pogrupowane wg określonych kryteriów

- punktowy - grupuje się wartości zmiennej każdej wartości cechy i podaje liczebność/częstość (używamy jak jest mało unikalnych wartości)
- przedziałowy - wymaga pogrupowania wartości cechy w przedziały klasowe. Liczebności i częstości podaje się dla tak utworzonych przedziałów (używamy jak jest dużo unikalnych wartości)

Cecha statystyczna - właściwość, którą badamy, mierzymy u jednostek zbiorowości?
Wartość cechy - konkretna liczba/wartość, którą cecha przyjmuje u danej jednostki

---

Średnia

Arytmetyczna zwykła

Suma wartości zmiennej wszystkich jednostek badanej zbiorowości podzielona przez liczbę tych jednostek.

- Jest wrażliwa na wpływ wartości skrajnych, , bo każdy składnik sumy jest tak samo ważny
- jest reprezentatywna dla rozkładów o niewielkiej asymetrii

$$\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$$

x_i - wariant cechy mierzalnej
n - liczebność badanej zbiorowości

Arytmetyczna ważona

Średnia elementów o różnych przypisanych wagach. Przez to elementy o większej wadze mają większy wpływ na średnią.

- Wagi to współczynniki, które przyporządkowuje się wartościom cechy, aby nadać im pożądane znaczenie

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

w_i - waga cechy
x_i - wartość cechy
n - liczebność cechy

Harmoniczna

Odwrotność średniej arytmetycznej z odwrotności poszczególnych wartości danej cechy.

- Cechy często podawane są w jednostkach względnych np. prędkości w km/h

$$H = \frac{N}{\sum_{i=1}^{N} \frac{1}{x_i}}$$

N - liczebność cechy
x_i - wartość cechy

Geometryczna

Pierwiastek stopnia n z iloczynu n wartości zmiennej.

- stosuje się przy badaniu średniego tempa zmian zjawisk
- gdy wykres układa się na kształt logarytmiczny

$$\bar{x}_G = \left(\prod_{i=1}^{n} x_i \right)^{1/n}$$

x_i - wartość cechy
n - liczebność cechy

Mediana !!!

Wartość środkowa dzieląca zbiorowość w postaci uporządkowanego szeregu statystycznego na dwie równe części.

Dla nieparzystej liczby wyrazów:

$$M = x_{\frac{n+1}{2}}$$

x_i - wartość z uporządkowanego szeregu
n - liczba wyrazów w szeregu

Dla parzystej liczby wyrazów:

$$M = \frac{x_{\frac{n}{2}} + x_{\frac{n+1}{2}}}{2}$$

Dla szeregu rozdzielczego:
Należy znaleźć przedział medianowy, czyli taki, w którym znajduje się element środkowy

$$M = x_0 + \frac{l_M}{f_0}(\frac{n}{2} - f_1)$$

n - liczba wyrazów w szeregu
x_0 - dolna granica przedziału, w którym leży mediana
l_M - rozpiętość przedziału klasy mediany
f_0 - liczebność w klasie mediany
f_1 - łączna liczebność w przedziałach poprzedzających klasę mediany

---

Dominanta (moda, wartość modalna)

Wartość najczęściej występująca w zbiorze danych. Jest miarą tendencji centralnej.

- może być więcej niż jedna dominanta

Wariancja

Średnia arytmetyczna z kwadratów odchyleń poszczególnych wartości cechy od średniej arytmetycznej. **Pozwala oszacować przeciętne zróżnicowanie badanej zbiorowości ze względu na wybraną cechę**

$$Var(x) = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n}$$

$\bar{x}$ - średnia arytmetyczna
x_i - poszczególne wartości cechy
n - liczebność cechy

- Gdy znamy część populacji(próbę), a odchylenie chcemy policzyć dla całej populacji, w mianowniku piszemy $n-1$

Odchylenie

Przeciętne

Średnia arytmetyczna bezwzględnych wartości odchyleń zmiennej od wartości jej średniej arytmetycznej. Stosuje się, gdy:

- rozkład badanej cechy jest symetryczny lub zbliżony do symetrycznego
- istnieją skrajne wartości, które po podniesieniu do kwadratu znacząco obciążyłyby wartość odchylenia standardowego

$$d = \frac{|x_1 - \bar{x}| + |x_2 - \bar{x}| + \cdots + |x_n - \bar{x}|}{n}$$

x_i - wartości cechy
$\bar{x}$ - średnia arytmetyczna
n - liczebność cechy

Standardowe!!!

Pierwiastek z wariancji. Określa o ile wszystkie jednostki badanej zbiorowości różnią się średnio od średniej arytmetycznej.

- Korzystając z tego można określić obszar, w jakim badana cecha przyjmuje wartości typowe. Przyjmuje się, że jest to przedział o długości dwóch odchyleń standardowych ($-\sigma < xtypowe < \sigma$)

$$\sigma = \sqrt{Var(x)}$$

Asymetria

Miara naruszenia symetryczności danej cechy.
$A_s > 0$ - Występuje asymetria **prawostronna**,
$A_s = 0$ - Rozkład jest **symetryczny**,
$A_s < 0$ - asymetria **lewostronna**

$$A_s = \frac{\bar{x} - D}{\sigma}$$

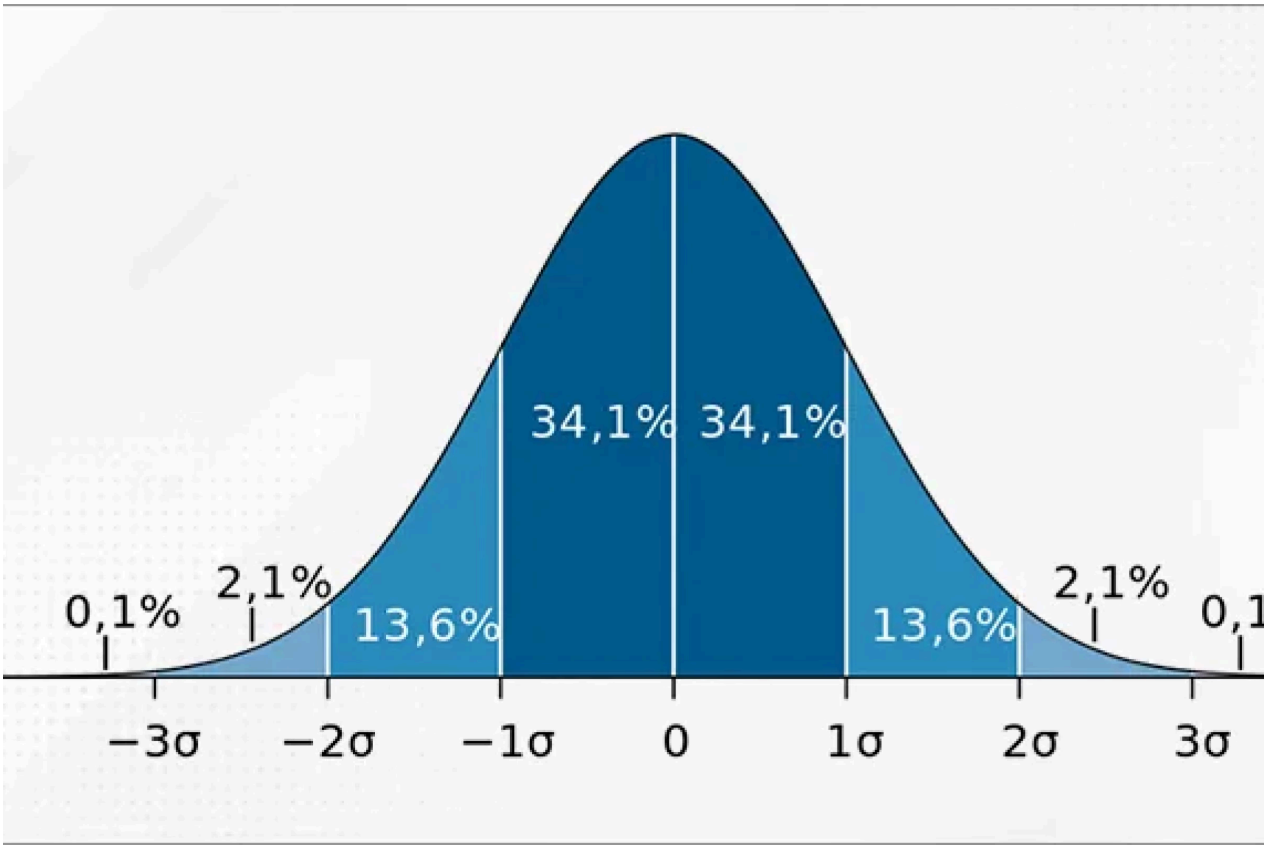
A_s - współczynnik asymetri
$\bar{x}$ - średnia arytmetyczna

D - dominanta
σ - odchylenie standardowe

Reguła trzech sigm B)

Mówi, że w rozkładzie normalnym:

- 68,2% wartości cechy leży w odległości jednej sigmy(odchylenia standardowego σ) od średniej,
- 95,4% w odległości dwóch sigm
- 99,7% w odległości trzech sigm



Kowariancja!!!

Miara wspólnego zróżnicowania dwóch zbiorów danych od średniej.

- Dodatnia mówi o zmianach w tym samym kierunku
- Ujemna o zmianach w kierunku przeciwnym.
- Wartości bliskie 0 sugerują brak prostej, liniowej zależności
  Służy do obliczania **korelacji liniowej Pearsona**

$$Cov(x,y) = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n}$$

x, y - zmienne, między którymi liczymy kowariancję
$\bar{x}, \bar{y}$ - średnie arytmetyczne
x_i, y_i - poszczególne wartości cech
n - liczba obserwacji

Współczynnik korelacji liniowej Pearsona!!!

Jest to miara liniowej zależności między zmiennymi x i y. Stosuje się go do badania korelacji cech ilościowych. To normalizacja kowariancji, dzięki której współczynnik ten nie zależy od jednostek i mieści się w zakresie od -1 do 1, gdzie:

- 1 - idealna, rosnąca zależność liniowa
- 0 - brak korelacji liniowej
- -1 - idealna, malejąca zależność liniowa

$$r(x, y) = \frac{Cov(x, y)}{\sigma(x)\sigma(y)}$$

$Cov(x, y)$ - kowariancja między zmiennymi
$\sigma(x), \sigma(y)$ - odchylenia standardowe zmienych

Regresja!!!

Liniowa

Stosowana do opisu liniowej zależności zmiennej y od zmiennej x. Celem modelu regresji linowej jest znalezienie najlepszego dopasowania prostoliniowego modelu, który pozwala prognozować wartość y na podstawie wartości x.

$$y = f(x) = \beta_0 + \beta_1 x + \epsilon$$

β_0, β_1 - parametry modelu (współczynniki regresji)
ϵ - składnik losowy (różnica między wartością przewidywaną, a rzeczywistą)

Wielokrotna

Model regresji służący do opisu wpływu n zmiennych x_i na jedną zmienną y

$$y = f(x) = \beta_0 + \sum_{i=1}^{n} \beta_i x_i + \epsilon$$

β_0 - wyraz wolny
β_i - współczyniki regresji odzwiercielające wpływ i-tej zmiennej na zmienną y
ϵ - składnik losowy - różnica między wartością przewidywaną, a rzeczywistą

Częstość względna!!!

Wynik dzielenia liczebności danej klasy przez liczbę wszystkich jednostek, czyli **proporcja wystąpień danej klasy w całym zbiorze danych.** Jej wyniki mieszczą się w przedziale od 0 do 1

$$f_i = \frac{n_i}{N}$$

f_i - częstosć względna klasy o indeksie i
n_i liczebność klasy o indeksie i
N - łączna liczba elementów

Wzór Bayesa!!!

Służy do określania prawdopodobieństwa przynależności obiektu do j-tej klasy na podstawie obserwacji x.

$$p_j(x) = \frac{p_j f_j(x)}{f(x)}$$

$p_j(x)$ - **prawdopodobieństwo warunkowe po zaobserwowaniu x, że badany obiekt należy do klasy j**
p_j - **prawdopodobieństwo wystąpienia klasy j na podstawie ogólnych** danych
$f_j(x)$ - **funkcja gęstości prawdopodobieństwa dla klasy j** (jak bardzo typowa jest obserwacja x dla klasy j) Np funkcja temperatur diagnozowanych przy grypie
$f(x)$ - **funkcja gęstości dla obserwacji x we wszystkich klasach**

Czy dobór jednostki ma znaczenie?

Tak. Jeżeli jedna cecha przyjmuje wartości ze znacznie szerszego przedziału niż druga, to będzie mieć większy wpływ na wynik klasyfikacji. Z tego powodu zakresy wartości cech powinny zostać ujednolicone.

Normalizacja

Z-score

Rodzaj normalizacji, w wyniku którego zmienna uzyskuje wartość oczekiwaną 0 i odchylenie standardowe 1.

- Stosujemy ją w przypadkach, gdy mamy do czynienia z elementami odstającymi o bardzo dużej lub małej wartości

$$x^{\#} = \frac{x - \bar{x}}{\sigma}$$

$x^{\#}$ - wartość znormalizowana
$\bar{x}$ - wartość średnia
σ - odchylenie standardowe

Min - max

Sprowadzenie wartości wszystkich atrybutów do przedziału $[0, 1]$. Dzięki temu odległości nie są zdominowane przez żaden atrybut.

$$x^{\#} = \frac{x - x_{\min}}{x_{\max} - x_{\min}}$$

$x^{\#}$ - wartość znormalizowana
x - wartość bieżąca
$xmin, xmax$ - minimalna i maksymalna wartość w całym zbiorze danych

---

Błędy

Bezwzględny

Różnica między wartością obliczoną, a prawdziwą

$$\delta_i = x_i - x_0$$

δ_i - błąd bezwzględny ze znakiem
x_i - wartość obliczona
x_0 - wartość prawdziwa

Względny

Iloraz błędu bezwzględnego i wartości prawdziwej

$$\frac{x_i - x_0}{x_0}$$

Względny procentowy

Błąd względny wyrażony w procentach (mnożymy razy 100%)

Wskaźnik

Stosunek liczebności/wielkości dwóch zbiorowości pozostających ze sobą w logicznym związku. np. gęstość zaludnienia (osoby/1000m^2)

$$W = \frac{W_1}{W_2}$$

To jest ułamkowy, mogą być jeszcze procentowe i promilowe, czyli ułamkowy $\times 100\%$ i $\times 100 promili$

Indeks

Miara dynamiki - zmian w reprezentatywnej grupie pojedynczych punktów danych.

Prosty

Ułamkowy

$$I_u = \frac{x_t}{x_0}$$

x_t - okres badawczy
x_0 - okres porównawczy

Procentowy

$I_p =$ ułamkowy $\times 100\%$

Promilowy

$I_{pm} =$ ułamkowy $\times 1000 promili$

Złożony!!

Służą do pomiaru dynamiki procesów złożonych z kilku składników. Wyznaczane są dla grupy badanej zbiorowości. Najczęściej przytaczanym przykładem użycia jest łączna ocena zmian ilościowych i zmian cen.

$$w = l \times c$$

w - realny wydatek
c - cena jednostkowa artykułu
l - liczba artykułów

Związek między indeksami

$$I_w = \frac{w_t}{w_0}, I_l = \frac{l_t}{l_0}, I_c = \frac{c_t}{c_0}$$

Jest identyczny jak między kategoriami

$$I_w = I_l \times I_c$$

Wartość agregatu W jest sumą wydatków na każdy artykuł, czyli sumą iloczynów ilości i cen poszczególnych artykułów:

$$W = \sum_{j=1}^{k} w_j = \sum_{j=1}^{k} l_j c_j$$

Indeks wydatku (wartości)!!!

Wskaźnik używany do porównania wartości całkowitej grupy produktów lub usług w różnych momentach czasu. Pozwala mierzyć zmiany wartości w czasie uwzględniając zarówno zmiany cen, jak i ilości. Indeks mierzy o ile zmieniła się łączna wartość grupy produktów w okresie t w porównaniu do okresu bazowego 0

$$I_w = \frac{\sum_{j=1}^{k} w_j^t}{\sum_{j=1}^{k} w_j^0} = \frac{\sum_{j=1}^{k} l_j^t \times c_j^t}{\sum_{j=1}^{k} l_j^0 \times c_j^0}$$

Indeks wg Laspeyresa!!

ilości
Indeks informujący, jak zmieniłaby się łączna wartość całego agregatu produktów w okresie badanym w stosunku do okresu bazowego, gdyby ceny wszystkich produktów w obydwu porównywanych okresach były takie same.

$$I_l^L = \frac{\sum_{j=1}^{k} l_j^t \times c_j^0}{\sum_{j=1}^{k} l_j^0 \times c_j^0}$$

cen
Indeks informujący, jak zmieniłaby się łączna wartość całego agregatu produktów w okresie badanym w stosunku do okresu bazowego, gdyby ilości tych produktów w obydwu porównywanych okresach były takie same.

$$I_c^L = \frac{\sum_{j=1}^{k} l_j^0 \times c_j^t}{\sum_{j=1}^{k} l_j^0 \times c_j^0}$$

Indeks wg Paasche'a!!

ilości
Indeks informujący, jak zmieniłaby się łączna wartość całego agregatu produktów w okresie badanym w stosunku do okresu bazowego, gdyby ceny wszystkich produktów w obydwu porównywanych okresach były takie same.

$$I_l^P = \frac{\sum_{j=1}^{k} l_j^t \times c_j^t}{\sum_{j=1}^{k} l_j^0 \times c_j^t}$$

cen
Indeks informujący, jak zmieniłaby się łączna wartość całego agregatu produktów w okresie

badanym w stosunku do okresu bazowego, gdyby ilości w obydwu porównywanych okresach były takie same.

$$I_c^P = \frac{\sum_{j=1}^{k} l_j^t \times c_j^t}{\sum_{j=1}^{k} l_j^t \times c_j^0}$$

Klasyfikacja i grupowanie

klasyfikacja!!! - ostatni etap procesu rozpoznawania, który polega na zakwalifikowaniu danego wektora opisującego obiekt do jednej ze wcześniej zdefiniowanych kategorii (klas). W przeciwieństwie do grupowania, podczas klasyfikacji zakładamy, że obiekt ma jednoznacznie określoną klasę.

grupowanie!! - wyodrębnianie zbiorów jednostek o identycznych lub podobnych właściwościach. Ze względu na kierunek budowania pewnych poziomów grupowania wyróżniamy grupowanie:

- aglomeracyjne (od szczegółu do ogółu)
- deglomeracyjne (od ogółu do szczegółu)

zliczanie - określenie liczby jednostek posiadających określony wariant cechy, który to wariant wyznacza w zbiorowości pewną klasę

k średnich!!

Przykład algorytmu grupowania płaskiego. Polega na podziale zbioru danych klastry, gdzie każdy klaster reprezentowany jest przez średnią, czyli centroid.

1. Początkowo położenie centroidów ustalane jest poprzez wybranie k losowych wektorów cech ze zbioru danych.
2. Następnie przypisuje się każdemu wektorowi cech najbliżej leżący centroid wyznaczając w ten sposób klastry
3. Dla każdego klastra aktualizuje się położenie centroidów poprzez obliczenie średniej położenia należących do niego punktów i przypisanie jej i-temu centroidowi.
4. Punkty 2 i 3 wykonuje się aż do uzyskania warunku stopu, czyli póki położenie centroidów ulega zmianie.

kNN!!

Metoda klasyfikacji polegająca na zakwalifikowaniu danego wektora cech ze zbioru testowego do jednej z klas na podstawie danych ze zbioru treningowego.

1. Obliczenie odległości obiektu klasyfikowanego od każdego punktu ze zbioru treningowego

2. Wybór k obiektów o najmniejszych odległościach od tego obiektu
3. Zbadanie z jakich klas pochodzą obiekty wybrane w kroku 2
4. wybór najliczniej reprezentowanej klasy
5. rozstrzygnięcie ewentualnych remisów (na przykład, kiedy istnieją dwie najliczniej reprezentowane klasy). Można to zrobić na przykład losowo lub przez wybór nieparzystego k

Ocena jakości klasyfikacji

sensitivity (czułość) - true positive rate - stosunek do sumy w wierszu - jak dużo prawdziwych obiektów wykryłeś TP / TP + FN

precision (precyzja) - true positives w stosunku do true positives + false positives - stosunek do sumy w kolumnie TP / TP + FP

Dwa powyższe parametry są wzajemnie przeciwstawne

specifitity (swoistość) - true negative rate TN / TN + FP

```
   P  N <- wynik
P TP FN
N FP TN
```

Odległość euklidesowa

$$d(p_1, p_2) = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2}$$

Iloczyn skalarny

$$d(x, y) = 1 - \frac{\|x\|\|y\| \cos(x, y)}{\|x\|\|y\|}$$
