Opracowane zagadnienia ze Wstępu do Statystycznej Analizy Danych

Mariusz Strzelecki (szczeles@mat.uni.torun.pl)
tu się ładnie wpiszcie :D

25 maja 2009

1 Typy zmiennych w analizie danych. Rozkład częstości. Graficzne przedstawienie danych

2 Zmienna losowa, rozkład, dystrybuanta

W ujęciu intuicyjnym zmienna losowa to taka zmienna, która w wyniku doświadczenia przyjmuje wartość liczbową zależną od przypadku, a więc nie dającą się ustalić przed przeprowadzeniem doświadczenia.

Niech $(\Omega, \mathcal{F}, P)$ będzie dowolną przestrzenią probablistyczną.

Zmienną losową nazywamy dowolną funkcję $X\colon \Omega \to \mathbb{R}$, która jest mierzalna i spełnia warunek: $\{\omega\colon X(\omega) \in B\} \in \mathcal{F}$ dla każdego dowolnego, rozsądnego $B \subset \mathbb{R}$.

Rozważmy dwie σ-algebry, największą: 2^Ω i najmniejszą: $\{\emptyset, \Omega\}$. W tych obu przypadkach: jakie musi być $X\colon \Omega \to \mathbb{R}$, aby $X^{-1}(B) \in \mathcal{F}$?

Dla $\mathcal{F} = 2^\Omega$: X - dowolna funkcja

Dla $\mathcal{F} = \{\emptyset, \Omega\}$ X - funkcja stała. Gdyby tak nie było, czyli $X\colon \Omega \to \{x_1, x_2\}$ to $\{\omega\colon X(\omega) \in \{x_1\}\} \neq \emptyset, \neq \Omega$ oraz $\{\omega\colon X(\omega) \in \{x_2\}\} \neq \emptyset, \neq \Omega$ (a razem dają Ω).

Definicja. *Zmienną losową nazywamy dowolną funkcję $X\colon \Omega \to \mathbb{R}$ spełniającą waruenk $X^{-1}(B) \in \mathcal{F} \forall_{B \in \mathbb{B}(\mathbb{R})}$, gdzie $\mathbb{B}(\mathbb{R})$ to σ-algebra zbiorów borelowskich.*

Funkcje z tej definicji nazywamy mierzalnymi $(\Omega, \mathcal{F}) \to (\mathbb{R}, \mathbb{B}(\mathbb{R}))$ (bo jest to odwzorowanie przenoszące badania prawdopodobieństwa z niewygodnej przestrzeni probabilistycznej do dobrze znanej przestrzeni euklidesowej)

Pytanie: co to jest zbiór borelowski? Najprościej rzecz umując jest to taki rozsądny pozdbiór $\mathbb{R}$, na przykład $\{x_0\}$, $[a, b]$, $(x0, \infty)$, $\mathbb{N}$. W praktyce trudno spotkać zbiór nieborelowski, ale istnieją, a nawet jest ich więcej niż borelowskich.

Fakt. *Niech $X\colon \Omega \to \mathbb{R}$ - zmienna losowa. Rozważmy $Y = g(X)\colon \mathbb{R} \to \mathbb{R}$. Y jest zmienną losową $\iff$ g jest funkcją mierzalną $(\mathbb{R}, \mathbb{B}(\mathbb{R})) \to (\mathbb{R}, \mathbb{B}(\mathbb{R}))$*

I to tyle o zmiennych losowych, co zrobić, żeby je ogarnąć? Może przykład!

Rzucamy raz kostką i badamy wyniki. Nasza $\Omega = (\omega_6, ..., \omega_6)$, gdzie ω_i oznacza "wyrzucono i".

I teraz: zmienna losowa $X\colon \Omega \to \mathbb{R}$ ma się następująco: $X(\omega_i) = i$. A co ze zbiorami borelowskimi? Wydaje

mi się, że X^{-1} można zapisać "taki iloczyn zdarzeń elementarnych ω_j, kiedy $j \in B$". Czyli weźmy śmieszny $B_0 = \{4\} \cup (1.2, 2.4]$ i wtenczas $X^{-1}(B_0) = \{\omega_2, \omega_4\}$. Jeśli źle mówię, poprawcie mnie.

Dalej: tego nie było na wykładzie, ale pojawia sie wszędzie: zmienna losowa X jest *typu skokowego*, jeśli przyjmuje przeliczalną liczbę wartości. Wtedy wartości zmiennej losowej (*punkty skokowe*) oznaczamy przez $x_1, x_2, ...$, a prawdopodobieństwa, z jakimi sa one realizowane (*skoki*) przez $p_1, p_2, ...$ Zmienna losowa jest *typu ciągłego* jeśli jej możliwe wartości tworzą przedział $\in \mathbb{R}$. Dla zmiennej losowej typu ciągłego możliwe jest określenie prawdopodobieństwa, że przyjmuje ona wartość należącą do dowolnego zbioru jej wartości. Sposób rozdysponowania całej "masy"prawdopodobieństwa (równej 1) pomiędzy wartości, jakie przyjmuje dana zmienna losowa, określamy mianem jej *rozkładu prawdopodobieństwa.*

Wracamy do wykładu!

Definicja. *Rozkładem zmiennej losowej X nazywamy miarę probabilistyczną (prawdopodobieństwo) postaci: $P_X(B) = p(X^{-1}(B)), \forall_{B \in \mathbb{B}(\mathbb{R})}$. Zauważmy, że $p\colon \mathcal{F} \to [0,1]$, zaś $P_X\colon \mathbb{B}(\mathbb{R}) \to [0,1]$.*

Uwaga. *Tak określony rozkład to rzeczywiście prawdopodobieństwo, ponieważ*
$P_X(\mathbb{R}) = p(X^{-1}(\mathbb{R})) = p(X^{-1}(\bigcup_{n=1}^{\infty}(-n, n))) = p(\Omega) = 1.$

Uwaga. $P_X(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P_X(A_i)$, *gdy A_i - rozłączne.*

Dowód. $P_X(\bigcup_{i=1}^{\infty} A_i) = p(X^{-1}(\bigcup_{i=1}^{\infty} A_i)) = p(\bigcup_{i=1}^{\infty} X^{-1}(A_i)) = \sum_{i=1}^{\infty} p(X^{-1}(A_i)) = \sum_{i=1}^{\infty} P_X(A_i).$ □

Definicja. *Dystrybuantą zmiennej losowej X nazywamy funkcję $F_X\colon \mathbb{R} \to [0,1]$ określoną:*
$F_X(a) = P_X((-\infty, a]) = P(X \leqslant a)$. *Dla rozkładu dyskretnego (kiedy zmienne sa skokowe) możemy zapisać* $F_X(a) = \sum_{k:x_k \leqslant a} P_X(\{x_k\}) = \sum_{k:\ x_k \leqslant a} p_k$.

Charakterystyczne własności dystrybuanty:

1. Funkcja niemalejąca
2. Funkcja prawdostonnie ciągła
3. $\lim_{x \to -\infty} F_X(x) = 0, \lim_{x \to +\infty} F_X(x) = 1$.

3 Rozkłady dyskretne i absolutnie ciągłe. Przykłady

Definicja. *Rozkład zmiennej losowej X nazywamy dyskretnym, jeśli $\exists_{S \subset \mathbb{R}} \sharp S \leqslant \sharp\mathbb{N}, P_X(S) = 1$.*

Uwaga. $B \in \mathbb{B}(\mathbb{R}), P_X(B) = P_X(B \cap S) = \sum_{k:x_k \in B} P(\{x_k\}) = \sum_{k:\ x_k \in B} p_k$
A $p_k = P_X(\{x_k\}) = P(X = x_k)$.
Wniosek z tego prosty: rozkład możemy zadać przez pary liczb (x_k, p_k) dla $k = 1, 2, ...$

Przykład. *(1) Jednopunktowy, $S = x_1$. Wtenczas $P(X = x_1) = 1$.*

(2) Dwupunktowy, $S = x_1, x_2$. Wtenczas $P(X = x_1) = p, P(X = x_2 = 1 - p)$.
Jeśli $x_1 = 1$ i $x_2 = 0$ to rozkład nazywamy zerojedynkowym.

(3) Dwumianowy, $S = 0, 1, 2, ..., n$. $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, k = 0, 1, ..., n$.
n i p to parametry rozkładu, skojarzenie z prawdopodobieństwem k-sukcesów w niezależnych próbach Bernouliego jak najbardziej na miejscu.

(4) Poissona, $S = 0, 1, ...$ $P(X = k) = \frac{\lambda^k}{k!} e^{-\lambda}$.

$\lambda > 0$ - *parametr rozkładu.*