Przypomnienie elementów z rachunku prawdopodobieństwa. Naiwny klasyfikator Bayesa. Aktualizacja rozkładów wg reguły Bayesa.

Przemysław Klęsk
pklesk@wi.zut.edu.pl

Literatura

1. D. Hand, H. Mannila, P. Smyth, *Eksploracja danych*. WNT, Warszawa, 2005.
2. J. Koronacki, J. Ćwik, *Statystyczne systemy uczące się*. WNT, Warszawa, 2005.
3. P. Cichosz, *Systemy uczące się*. WNT, 2007.

Przypomnienie elementów z rachunku prawdopodobieństwa

Prawdopodobieństwo warunkowe

Niech A i B są pewnymi podzbiorami przestrzeni zdarzeń Ω. $A, B \subset \Omega$. Prawdopodobieństwo zdarzenia A pod warunkiem, że zaszło zdarzenie B oblicza się następująco:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}, \quad P(B) \neq 0. \tag{1}$$

Inny sposób:

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{\#(A \cap B)}{\#\Omega}}{\frac{\#B}{\#\Omega}}$$
$$= \frac{\#(A \cap B)}{\#B}. \tag{2}$$

Przykład

Tabelka z danymi:

nr	płeć	palanie	sport
1	M	tak	nie
2	M	tak	tak
3	K	nie	nie
4	M	nie	tak
5	K	nie	tak
6	M	nie	tak
7	K	tak	tak
8	M	nie	tak
9	M	nie	nie
10	K	nie	nie

Jakie jest prawdopodobieństwo (w danych), że wylosowany człowiek jest palący, pod warunkiem, że uprawia sport?

$$P(\text{palenie} = \text{tak} | \text{sport} = \text{tak}) = \frac{2}{6}.$$

Jakie jest prawdopodobieństwo (w danych), że wylosowany człowiek uprawia sport pod warunkiem, że jest to niepalący mężczyzna?

$$P(\text{sport} = \text{tak} | \text{płeć} = \text{M}, \text{palenie} = \text{nie}) = \frac{3}{4}.$$

Jakie jest prawdopodobieństwo (w danych), że wylosowany człowiek jest kobietą i uprawia sport pod warunkiem, że nie pali?

$$P(\text{płeć} = \text{K}, \text{sport} = \text{tak} | \text{palenie} = \text{nie}) = \frac{1}{7}.$$

Przypomnienie elementów z rachunku prawdopodobieństwa

Manipulowanie wzorem prawdopodobieństwa warunkowego

Przenoszenie zdarzenia pod warunek:

$$P(A,B|C) = \frac{P(A,B,C)}{P(C)} = \frac{P(A,B,C)\cdot P(B,C)}{P(C)\cdot P(B,C)} = P(A|B,C)P(B|C). \tag{3}$$

„Wyłuskiwanie" zdarzenie spod warunku:

$$P(A|B,C) = \frac{P(A,B,C)}{P(B,C)} = \frac{P(A,B,C)\cdot P(C)}{P(B,C)\cdot P(C)} = \frac{P(A,B|C)}{P(B|C)}. \tag{4}$$

Zrobić przykład z przeniesieniem zdarzeń $A_1,\ldots,A_{n-1}$ pod warunek:
$P(A_1A_2A_3\cdots A_n|B) = P(A_2A_3\cdots A_n|A_1B)P(A_1|B) = \ldots$.

Przypomnienie elementów z rachunku prawdopodobieństwa

Niezależność zdarzeń

- Mówimy, że dwa zdarzenia A i B są niezależne (piszemy $A \perp B$), wtedy i tylko wtedy, gdy:

$$P(A \cap B) = P(A) \cdot P(B). \tag{5}$$

- W odpowiednio dużej populacji Ω, jeżeli $A \perp B$, to należy oczekiwać, że A z taką samą częstością pojawia się w Ω jak i w B, oraz odwrotnie że B z taką samą częstością pojawia się w Ω jak i w A.
- Czyli jeżeli $A \perp B$, to $P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{P(A)P(B)}{P(B)} = P(A)$.
- Jeżeli zdarzenia nie są niezależne to ich występowanie razem ma inne prawdopodobieństwo (częstość), niż iloczyn prawdopodobieństw. Możemy domniemywać korelacji — czyli istnienia pewnej przyczyny, która te zdarzenia wiąże.

Przypomnienie elementów z rachunku prawdopodobieństwa

Prawdopodobieństwo całkowite

Dla każdego rozbicia przestrzeni zdarzeń Ω na rozłączne podzbiory $B_1, B_2, \ldots, B_n$ (każdy o dodatniej mierze prawdopodobieństwa), tj:

$$\bigcup_{i=1}^{n} B_i = \Omega,$$
$$\forall i \neq j \quad B_i \cap B_j = \emptyset,$$
$$\forall i \quad P(B_i) > 0.$$

prawdopodobieństwo dowolnego zdarzenia A możemy obliczać wg wzoru:

$$\begin{aligned} P(A) &= P(A|B_1)P(B_1) + P(A|B_2)P(B_2) + \cdots + P(A|B_n)P(B_n) \\ &= \sum_{i=1}^{n} P(A|B_i)P(B_i). \end{aligned} \tag{6}$$

Zrobić rysunek.

Przykład

- Trzy fabryki produkują żarówki. Prawdopodobieństwa, że wyprodukowana żarówka będzie świeciła dłużej niż 5 lat, wynoszą dla tych fabryk odpowiednio: 0.9, 0.8, 0.7. Prawdopodobieństwa napotkania na rynku żarówek z danej fabryki, wynoszą odpowiednio: 0.3, 0.5, 0.2. Jakie jest prawdopodobieństwo, że losowo zakupiona żarówka będzie świeciła dłużej niż 5 lat?
- Jeżeli wiemy, że pewna losowo zakupiona żarówka świeciła dłużej niż 5 lat, to jakie jest prawdopodobieństwo, że pochodzi ona z drugiej fabryki?

Przypomnienie elementów z rachunku prawdopodobieństwa

Prawdopodobieństwo całkowite (odwrócenie)

$$
\begin{aligned}
P(B_i|A) &= \frac{P(B_i \cap A)}{P(A)} \\
&= \frac{P(B_i \cap A)}{P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n)} && (7) \\
&= \frac{P(B_i \cap A)\frac{P(B_i)}{P(B_i)}}{P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n)} && (8) \\
&= \frac{P(A|B_i)P(B_i)}{P(A|B_1)P(B_1) + \cdots + P(A|B_n)P(B_n)}. && (9)
\end{aligned}
$$

Innymi słowy, patrzymy na udział i-tego składnika w całej sumie obliczanej wg prawdopodobieństwa całkowitego.

Przypomnienie elementów z rachunku prawdopodobieństwa

Schemat Bernoulli'ego

Prawdopodobieństwo uzyskania k sukcesów w serii n prób, gdzie prawdopodobieństwo sukcesu w pojedynczej próbie jest równe p, a porażki $1 - p$:

$$P(p, k, n) = \binom{n}{k} p^k (1-p)^{n-k}. \tag{10}$$

Przypomnienie elementów z rachunku prawdopodobieństwa

Elementy dla zmiennych/przestrzeni ciągłych

- Operujemy na funkcji gęstości rozkładu prawdopodobieństwa $p(x)$ i dystrybuancie $F(x) = P(X \leqslant x)$.
- $\int_{-\infty}^{\infty} p(x)dx = 1.$
- $P(a \leqslant X \leqslant b) = \int_{a}^{b} p(x)dx.$
- $P(a \leqslant X \leqslant b) = F(b) - F(a).$
- Różniczka dystrybuanty: $dF(x) = p(x)dx.$
- Całka z funkcji gęstości: $\int p(x)dx = F(x) + C.$
- Dla zbiorów (wielowymiarowych) mierzalnych w sensie całki Lebesgue'a:

$$P(A) = \frac{\int_A 1 d\lambda}{\int_\Omega 1 d\lambda}. \tag{11}$$

(Przykłady: igła spuszczana na stół, spotkanie Romeo i Julii.)

Przypomnienie elementów z rachunku prawdopodobieństwa

Wartość oczekiwana

- Wartość oczekiwaną (lub średnią) pewnej zmiennej losowej X o skończonej liczbie możliwych wartości $\{x_1, \ldots, x_n\}$ obliczamy wg wzoru:

$$E(X) = x_1 \cdot P(X = x_1) + \cdots + x_n \cdot P(X = x_n)$$
$$= \sum_{i=1}^{n} x_i P(X = x_i). \tag{12}$$

- W przypadku ciągłym — gdy nieskończony zbiór możliwości — mamy:

$$E(X) = \int_{x \in X} x p(x) dx. \tag{13}$$

Przypomnienie elementów z rachunku prawdopodobieństwa

Wariancja — średni kwadrat odchylenia

- W przypadku dyskretnym:

$$D^2(X) = \sum_{i=1}^{n} (x_i - E(X))^2 P(X = x_i) \tag{14}$$

$$= E(X^2) - E^2(X). \tag{15}$$

- W przypadku ciągłym:

$$D^2(X) = \int_{x \in X} (x - E(X))^2 p(x) dx \tag{16}$$

$$= E(X^2) - E^2(X). \tag{17}$$

- Odchylenie standardowe to pierwiastek z wariancji.

Naiwny klasyfikator Bayesa

- Realizuje zadanie klasyfikacji dla dowolnej liczby klas, tj. odwzorowanie:

$$X \to \{1, 2, \ldots, m\},$$

gdzie w ogólności $X = X_1 \times X_2 \times \ldots \times X_n \subseteq \mathbb{R}^n$.

- Jest to klasyfikator probabilistyczny z dołożonym tzw. *naiwnym założeniem*, że zmienne wejściowe są niezależne: $\forall i \neq j \quad X_i \perp X_j$.
- Dzięki temu założeniu klasyfikator radzi sobie świetnie z dowolną liczbą zmiennych (mogą być setki czy nawet tysiące) — nie cierpi na *przekleństwo wymiarowości*. Wraz ze wzrostem liczby zmiennych złożoność skaluje się liniowo a nie wykładniczo.
- Ograniczenie: w praktyce wymaga zmiennych dyskretnych (kategoryczne, wyliczeniowe). Zmienne ciągłe muszą zostać zdyskretyzowane na przedziały.

Naiwny klasyfikator Bayesa

- Dany jest pewien zbiór danych z wyróżnioną zmienną decyzyjną (wyjściową):

X_1	X_2	$\cdots$	X_n	Y
3	1	$\cdots$	2	1
2	5	$\cdots$	4	2
1	4	$\cdots$	2	2
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$

- Na podstawie tabelki znamy rozkłady prawdopodobieństwa **wektorów** wejściowych w poszczególnych klasach tj. $X = x|Y = y$ (utożsamiając je z rozkładami częstości).
- Dla pewnego nowego $X = \underbrace{(x_1, x_2, \ldots, x_n)}_{x}$ chcemy wyznaczyć numer klasy y^*, która jest najbardziej prawdopodobna:

$$y^* = \arg\max_y P(Y = y|X = x). \tag{18}$$

Naiwny klasyfikator Bayesa

$$P(Y = y|X = x) = \frac{P(X = x|Y = y)P(Y = y)}{P(X)}$$
$$= \frac{P(X = x|Y = y)P(Y = y)}{P(X = x|Y = 1)P(Y = 1) + \cdots + P(X = x|Y = m)P(Y = m)}. \quad (19)$$

Mianownik stały i niezależny od y, dla którego badamy $P(Y = y|X = x)$, a zatem:

$$y^* = \arg\max_y P(Y = y|X = x) = \arg\max_y P(X = x|Y = y)P(Y = y). \quad (20)$$

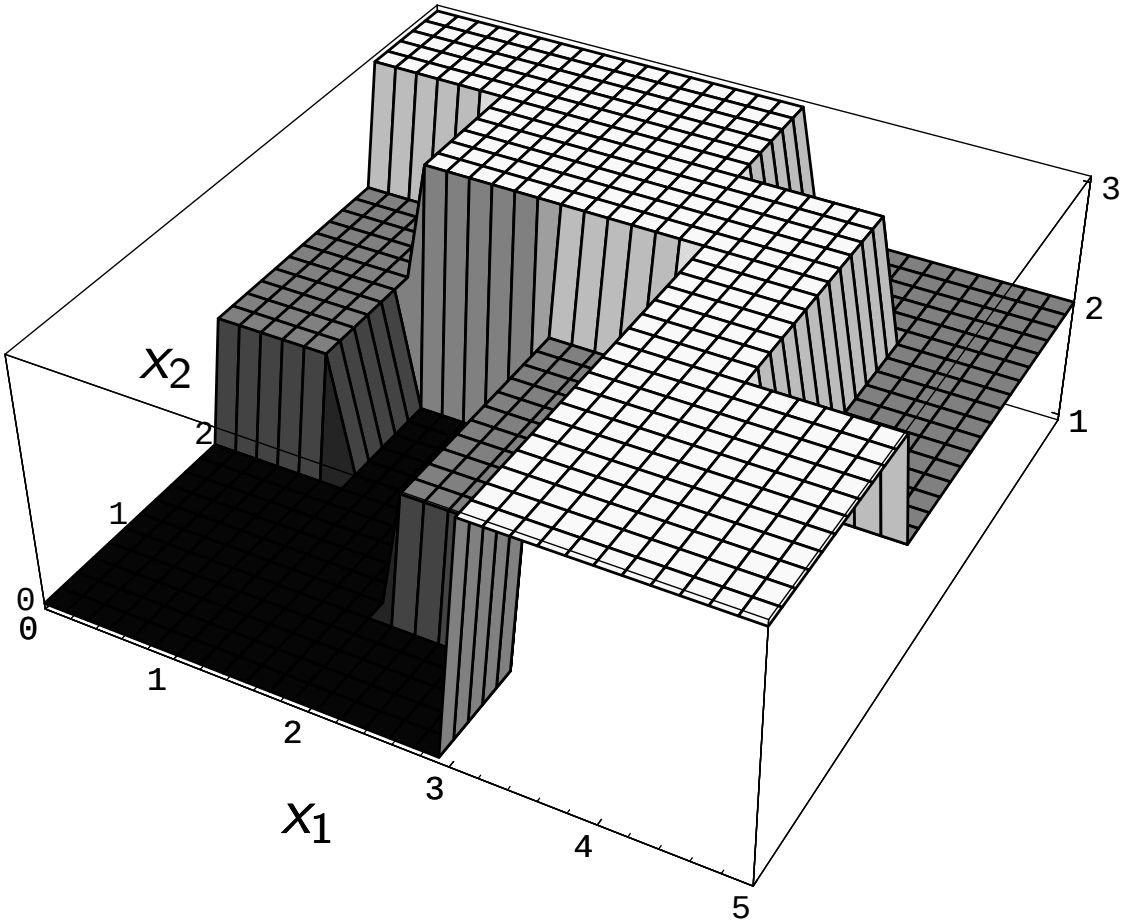
Wprowadzamy naiwne założenie o niezależności zmiennych, gdy w ogólności $x = (x_1, x_2, \ldots, x_n)$, mówiące że:

$$P(X = x|Y = y) = P(X_1 = x_1 \cap X_2 = x_2 \cap \cdots \cap X_n = x_n|Y = y)$$
$$= P(X_1 = x_1|Y = y)P(X_2 = x_2|Y = y) \cdots P(X_n = x_n|Y = y)$$
$$= \prod_{i=1}^{n} P(X_i = x_i|Y = y).$$

A zatem

$$y^* = \arg\max_y \prod_{i=1}^{n} P(X_i = x_i|Y = y)P(Y = y). \quad (21)$$

Przykładowy wykres działania klasyfikatora



$$f(x_1, x_2, \underbrace{x_3, \ldots, x_n}_{\text{const.}})$$

Rysunek: Odpowiedź klasyfikatora jako funkcja x_1 i x_2 przy ustalonych pozostałych zmiennych, np.:
$x = (x_1, x_2, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1, 1)$.

Naiwny klasyfikator Bayesa

- Zysk (z naiwności) — Niech dla uproszczenia każda ze zmiennych ma stałą lub średnią liczbę q wartości (moc dziedzin). Aby zapamiętać wszystkie możliwości dla $P(X_i = x_i | Y = y)$ potrzeba $O(n \cdot q \cdot m)$ pamięci. Bez naiwności byłoby to $O(q^n \cdot m)$, co w wielu przypadkach byłoby niemożliwe, więc klasyfikacja każdorazowo wymagałaby jednego przejścia po zbiorze danych, aby dla zadanego x wyznaczyć $P(X = x | Y = y)$.
- Trudność (bez naiwności) — Dla „małego" problemu, nawet jeżeli byłaby możliwość zapamiętania $O(q^n \cdot m)$ prawdopodobieństw łącznych $P(X = x | Y = y)$, to niektóre z nich mogłyby być błędnie zerowe (lub bardzo małe) z uwagi na braki realizacji wszystkich kombinacji wektora $x = (x_1, \ldots, x_n)$ w bazie danych, która jest do dyspozycji. Trudność ta znika przy naiwności, bo patrzymy na rozkłady pojedynczych zmiennych w klasach (a nie łączne).

Naiwny klasyfikator Bayesa

- „Epsilonowanie" — Gdy mało danych istnieje groźba, że $\prod_{i=1}^{n} P(X_i = x_i | Y = y) P(Y = y)$ stanie się (nieprawdziwie) zerem, gdy jeden z czynników jest zerem na podstawie tabelki z braku pewnych realizacji $X_i = x_i | Y = y$. Wyniki klasyfikacji mogą być z tego powodu czasami błędne. Dopuszczalna sztuczka: „wygłacić" wszystkie rozkłady warunkowe dodając do każdego punktu rozkładu pewną małą liczbę ϵ i normalizując do sumy równej jedności (dzieląc przez sumę).

Naiwny klasyfikator Bayesa

- Dla zachowania dobrych praktyk modelowania, klasyfikator powinien być budowany z rozbiciem zbioru danych na część uczącą i testującą lub z wykorzystaniem ogólniejszej techniki *kroswalidacji*.
- Po upewnieniu się, że na danym problemie klasyfikator poprawnie uogólnia — błędy na zbiorach testujących nie odbiegają znacząco od błędów na zbiorach uczących — można ostatecznie zbudować klasyfikator na całym zbiorze danych.

Reguła Bayesa

- Pozwala aktualizować nasz dotychczasowy model pewnego zjawiska na podstawie napływających danych.

$$P(\text{model}|\text{dane}) = \frac{P(\text{dane}|\text{model})P(\text{model})}{P(\text{dane})}. \quad (22)$$

- $$P(\text{model}|\text{dane}) = \frac{P(\text{dane}|\text{model})P(\text{model})}{P(\text{dane})} = \frac{P(\text{dane}|\text{model})P(\text{model})}{P(\text{dane}|\text{model}_1)P(\text{model}_1) + \ldots + P(\text{dane}|\text{model}_n)P(\text{model}_n)}. \quad (23)$$

- Pojęcia: a priori, a posteriori, likelihood.
- Wersja dla funkcji gęstości. Niech M oznacza nieskończony zbiór modeli, a m^* wyróżniony model, który chcemy zbadać:

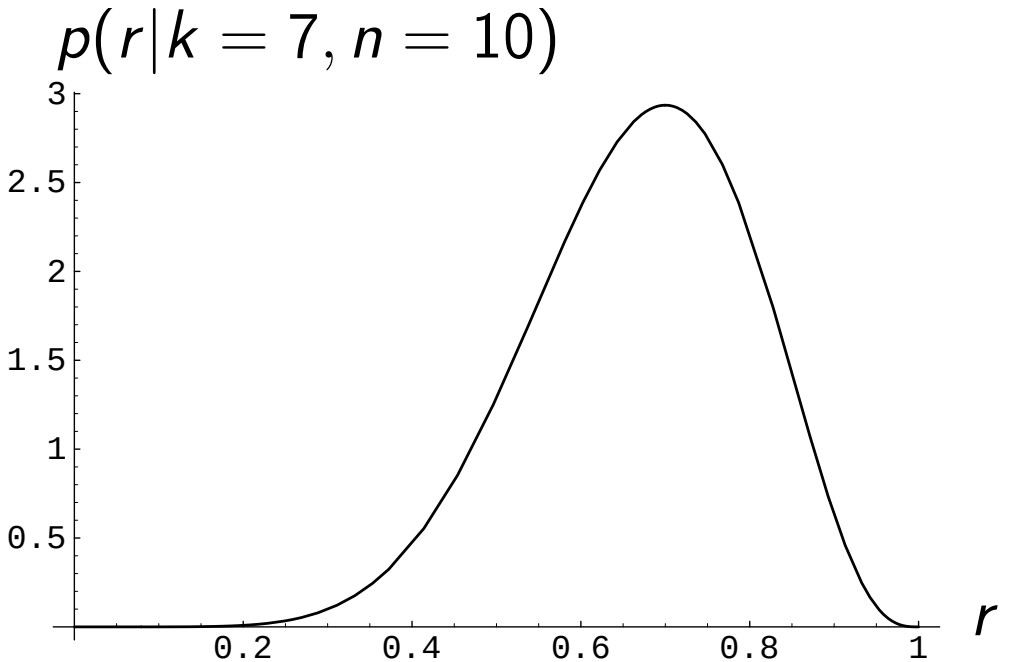
$$p(m^*|\text{dane}) = \frac{P(\text{dane}|m^*)p(m^*)}{\int_{m\in M} P(\text{dane}|m)p(m)dm}. \quad (24)$$

Reguła Bayesa

- W „jednorękich bandytach" mamy trzy różne rodzaje tarcz losujących symbol „7" z jednym prawdopodobieństw $r \in \{\frac{1}{2}, \frac{1}{4}, \frac{1}{8}\}$. Z jaką tarczą mamy do czynienia, jeżeli w próbie o liczności $n = 5$ mamy $k = 2$ sukcesy.
- Referendum: jaki rozkład na odsetek $r \in [0, 1]$ głosujących „tak" w całej populacji, jeżeli próba: $k = 7$ w $n = 10$.
- Jakie jest prawdopodobieństwo, że jutro wzejdzie słońce?
- Składanie reguły Bayesa z kilku prób (ze zmianą a priori), np.: $n_1 = 3$, $k_1 = 2$ i $n_2 = 2$ i $k_2 = 2$ lub $n = 5$ i $k = 4$.

Referendum

$$p(r|k=7,n=10) = \frac{r^7(1-r)^3}{\int_0^1 s^7(1-s)^3 ds} = \frac{r^7(1-r)^3}{\frac{1}{1320}}. \quad (25)$$



$$E(r) = \int_0^1 rp(r)dr = 1320\left(\frac{r^8}{8} - \frac{r^9}{3} + \frac{3r^{10}}{10} - \frac{r^{11}}{11}\right)\Big|_0^1 = \frac{2}{3}. \quad (26)$$

Czułość i specyficzność (klasyfikatora)

Czułość (ang. *sensitivity*)

$$\text{czułość} = \frac{\text{liczba prawdziwych pozytywnych}}{\text{liczba prawdziwych pozytywnych} + \text{liczba fałszywych negatywnych}} \tag{27}$$

Specyficzność lub swoistość (ang. *specifity*)

$$\text{specyficzność} = \frac{\text{liczba prawdziwych negatywnych}}{\text{liczba prawdziwych negatywnych} + \text{liczba fałszywych pozytywnych}} \tag{28}$$

Czułość i specyficzność a reguła Bayesa — przykład

Duża firma zamierza zrobić badania antynarkotykowe swoim pracownikom. Niech D i N oznaczają odpowiednio narkomana i nienarkomana, a $+$ i $-$ zdarzenia, że test wyszedł pozytywny lub negatywny. O pewnym teście narkotykowym wiadomo, że ma czułość $P(+|D) = 0.99$ i specyficzność $P(-|N) = 0.99$. Wiadomo także, że w całym dorosłym społeczeństwie mamy 0.005 narkomanów. Pytamy, ile wynosi $P(D|+)$? Czyli jakie jest prawdopodobieństwo, że ktoś jest faktycznie narkomanem, jeżeli go posądzimy o to na podstawie pozytywnego testu?

Czułość i specyficzność a reguła Bayesa — przykład

$$P(D|+) = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|N)P(N)}$$
$$= \frac{0.99 \cdot 0.005}{0.99 \cdot 0.005 + 0.01 \cdot 0.995} \approx 33\%.$$

Tylko około 33% spośród tych, dla których test dał wynik pozytywny, jest faktycznie narkomanami. Dlaczego? Wnioski?

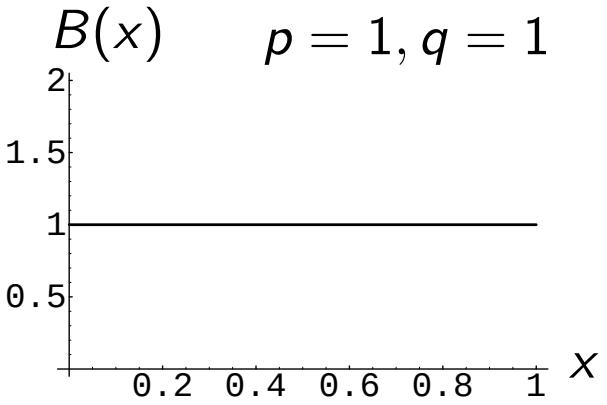
Rozkłady beta i uśrednianie ocen ekspertów

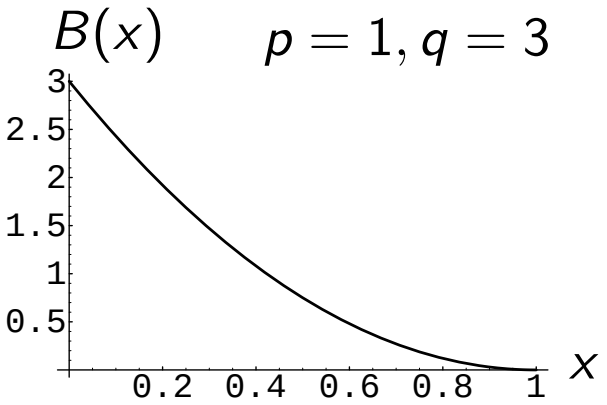
Parametryczna rodzina rozkładów beta $\{B(x,p,q)\}_{(p,q)\in\mathbb{N}^2}$

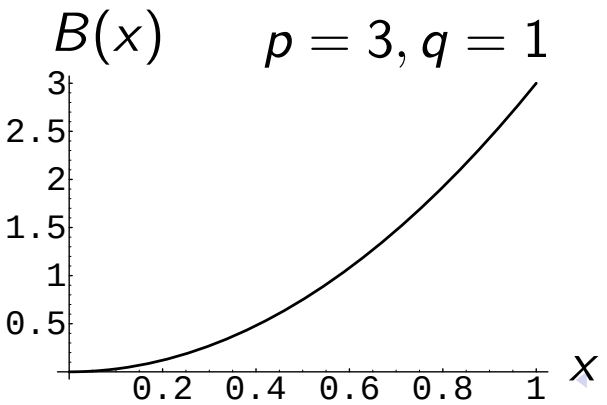
$$B(x,p,q) = \frac{x^{p-1}(1-x)^{q-1}}{\int_0^1 s^{p-1}(1-s)^{q-1}ds}, \tag{29}$$

gdzie p i q są odpowiednikami liczby sukcesów i porażek w rozkładzie dwumianowym (schemat Bernoulli'ego), tyle że powiększone o 1. Tzn. jeżeli np. $p = 1$ i $q = 1$, to odpowiada to tak naprawdę 0 sukcesów i 0 porażek i otrzymujemy rozkład jednostajny — często używany jako rozkład a priori.

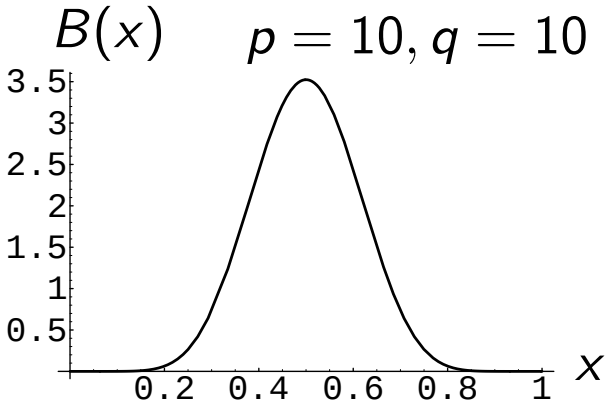
Rozkłady beta i uśrednianie ocen ekspertów

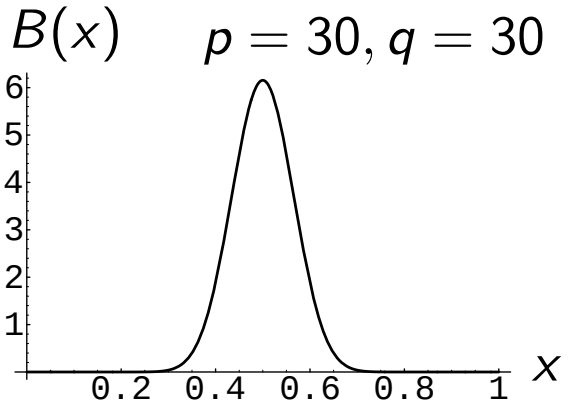


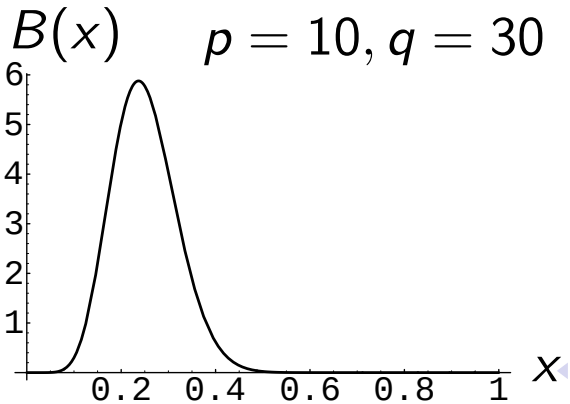




Rozkłady beta i uśrednianie ocen ekspertów







Rozkłady beta i uśrednianie ocen ekspertów

Dwóch ekspertów różnie wycenia tę samą nieruchomość. Jeden na 4 mln zł, a drugi na 10 mln zł. Wiemy, że za pierwszym ekspertem „stoi" baza danych 6 przypadków (doświadczenie zawodowe — tyle podobnych nieruchomości wycenił), a za drugim baza danych 2 przypadków?

1. Czy możemy wartość oczekiwaną oszacować jako: $\frac{6}{6+2}4 + \frac{2}{6+2}10 = 5.5$?
2. Jak możemy znaleźć funkcję gęstości rozkładu prawdopodobieństwa $p(r)$ na prawdziwą cenę nieruchomości $r \in [4, 10]$?
3. Czy wartość oczekiwana $E(r)$ policzona na podstawie funkcji gęstości rozkładu będzie dokładnie równa oszacowaniu z punktu 1?

Rozkłady beta i uśrednianie ocen ekspertów

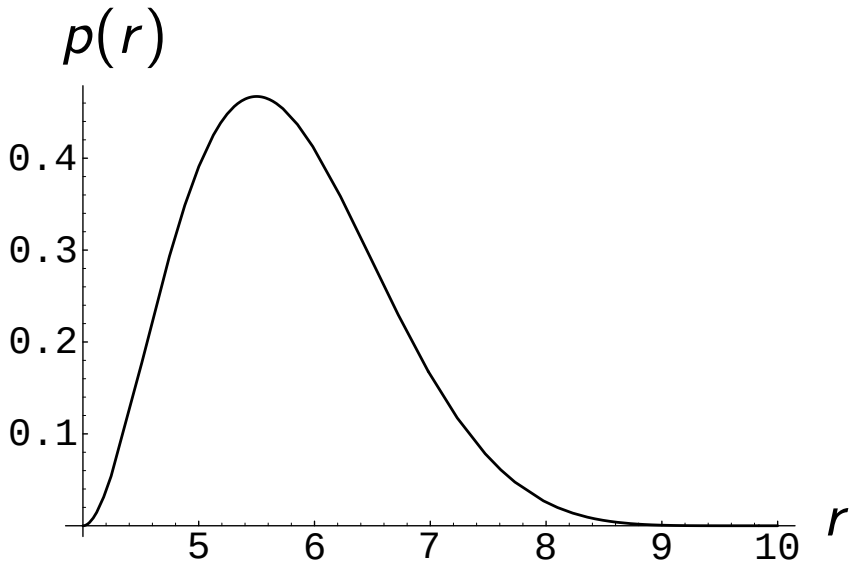
Zastosujemy odpowiedni rozkład beta, przy czym zmienną $x \in [0,1]$ zastępujemy zmienną $r \in [4,10]$ wg następującego przekształcenia:

$$x = \frac{r - r_{\min}}{r_{\max} - r_{\min}} = \frac{r-4}{10-4}. \tag{30}$$

Otrzymujemy:

$$p(r) = B(r, 2+1, 6+1) = \frac{(\frac{r-4}{10-4})^2(1-\frac{r-4}{10-4})^6}{\int_4^{10} (\frac{s-4}{10-4})^2(1-\frac{s-4}{10-4})^6 ds}$$
$$= \frac{7}{6}(1+\frac{4-r}{6})^6(-4+r)^2.$$

Rozkłady beta i uśrednianie ocen ekspertów



$$E(r) = \int_4^{10} rp(r)dr = \frac{7}{6}\left(\frac{125000r^2}{729} - \frac{275000r^3}{2187} + \frac{128125r^4}{2916} - \frac{6625r^5}{729} + \frac{20875r^6}{17496} - \frac{515r^7}{5103} + \frac{499r^8}{93312} - \frac{17r^9}{104976} + \frac{r^{10}}{466650}\right)\Big|_4^{10} = \frac{29}{5} = 5.8. \tag{31}$$