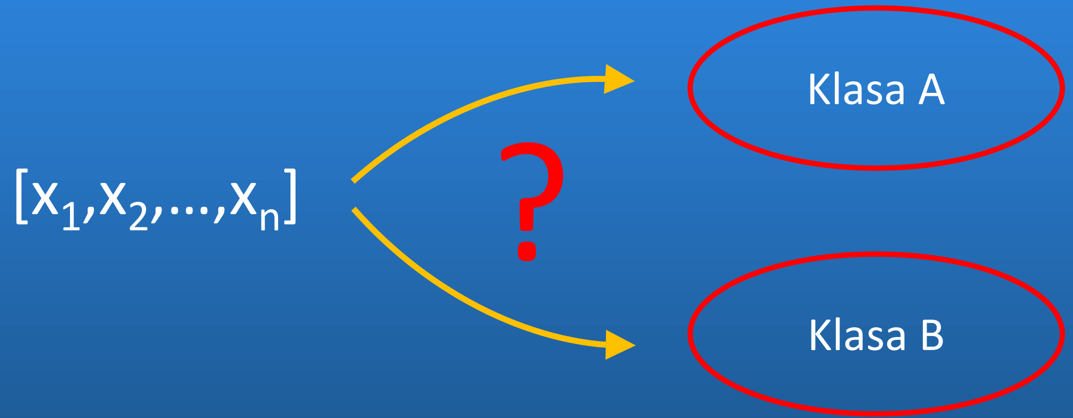
Klasyfikacja statystyczna

Klasyfikacja statystyczna - rodzaj algorytmu statystycznego, który przydziela obserwacje statystyczne do klas, bazując na atrybutach (cechach) tych obserwacji.

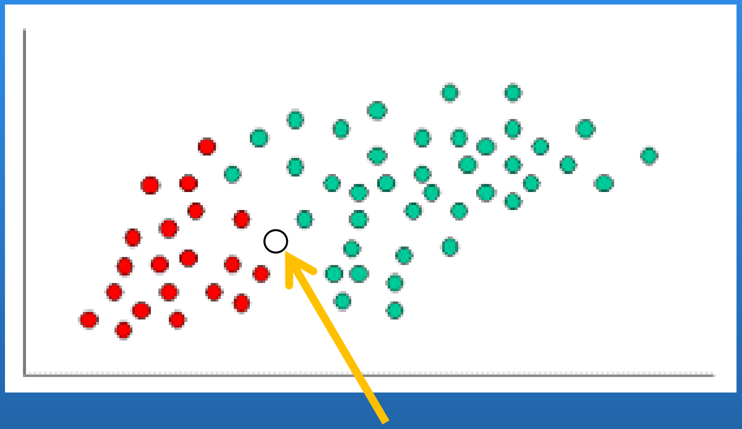


Rozpatrzmy dwa przykłady.

Klasyfikacja statystyczna

Przykład 1

Rozważmy następujący wykres rozrzutu:



Nowy obiekt chcemy zakwalifikować do jednej z dwóch kategorii: zielone lub czerwone.

Klasyfikacja statystyczna

Przykład 2

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Rozpatrzmy samochód: Red, Domestic, SUV

Chcemy go zakwalifikować do jednej z dwóch kategorii:

Stolen_Yes, Stolen_No

Klasyfikatory kNN

Klasyfikator kNN – klasyfikator k-najbliższych sasiadów (ang. k-nearest neighbor classifier)

Klasyfikator ten należy do grupy algorytmów opartych o analizę przypadku.

Idea klasyfikacji polega na wyszukiwaniu tych zgromadzonych przypadków, które mogą być zastosowane do klasyfikacji nowych sytuacji.

Klasyfikacja nowych przypadków jest realizowana na bieżąco, tzn. wtedy gdy pojawia się potrzeba klasyfikacji nowego przypadku.

Klasyfikatory kNN

Załóżmy, że dowolny przykład ze zbioru treningowego jest n-wymiarowym wektorem, reprezentującym punkt w przestrzeni n-wymiarowej nazywanej przestrzenią wzorców.

Klasyfikatory 1NN

Klasyfikacja nowego przypadku X:

- Poszukujemy punktu w przestrzeni wzorców, który jest „najbliższy” X
- Przypadek X klasyfikujemy jako należący do klasy, do której należy ten „najbliższy” punkt

Klasyfikatory kNN

Wada klasyfikatora 1NN

Duża czułość na punkty osobliwe i szum w danych treningowych

Rozwiązaniem jest rozważenie k najbliższych sąsiadów!

Otrzymujemy w ten sposób klasyfikator kNN.

Klasyfikatory kNN

Klasyfikatory kNN

Klasyfikacja nowego przypadku X:

- Poszukujemy w przestrzeni wzorców k najbliższych sąsiadów X.
- Przypadek X klasyfikujemy jako należący do klasy, która dominuje w zbiorze k najbliższych sąsiadów

Klasyfikatory kNN

Jak wybrać k?

- k < sqrt(n), gdzie n jest liczbą wszystkich przypadków.
- Mała wartość k – mała „stabilność”, k jest czułe na „szum”.
- Duża wartość k – mniejsza precyzja, bierzemy pod uwagę przypadki, które nie sąsiadują z kwalifikowanym.

Klasyfikatory kNN

Problemy

Problemy związane z klasyfikatorem kNN:

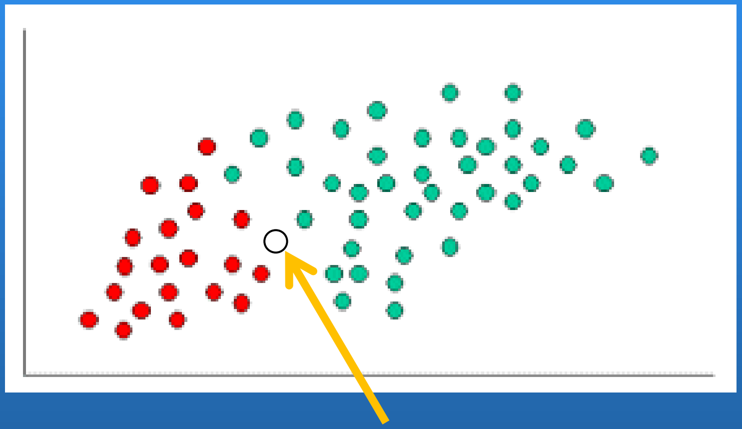
- Jak zdefiniować punkt „najbliższy”?

Rozwiązanie:

- W przypadku atrybutów liczbowych stosujemy euklidesową miarę odległości
- Stosujemy też inne miary odległości: blokową (Manhattan), Mińkowskiego itd.

Naiwny klasyfikator Bayesa

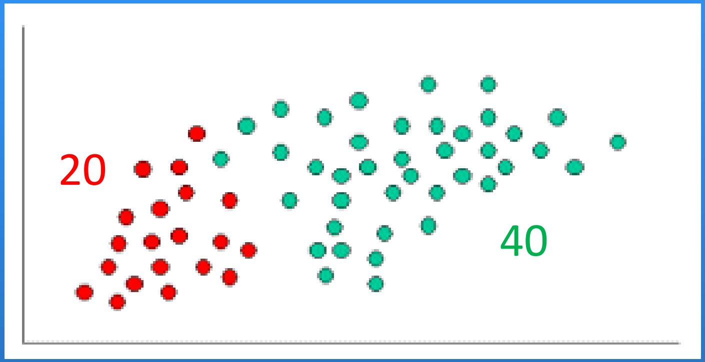
Mamy dany następujący wykres rozrzutu:



Nowy obiekt chcemy zakwalifikować do jednej z dwóch kategorii: zielone lub czerwone.

Naiwny klasyfikator Bayesa

Załóżmy na początek, że nowego obiektu jeszcze nie mamy:



Co możemy powiedzieć o takiej próbie?

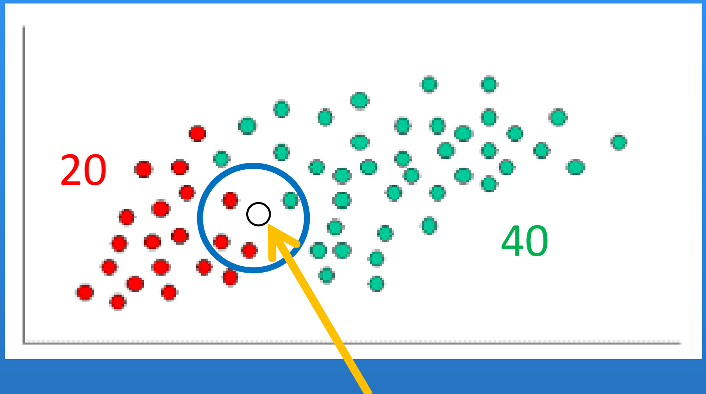
Prawdopodobieństwo, że nowy$\in$C=P(nowy$\in$C)=20/60=1/3

Prawdopodobieństwo, że nowy$\in$Z=P(nowy$\in$Z)=40/60=2/3

Jest to tak zwane prawdopodobieństwo a priori. Oparte jest ono na posiadanych już danych.

Naiwny klasyfikator Bayesa

Spróbujmy teraz zakwalifikować nasz obiekt:



Obliczmy teraz szanse w oparciu o sąsiedztwo punktu:

Szansa, że nowy będzie zielony = Sz(nowy$\in$Z) = 1/40

Szansa, że nowy będzie czerowny = Sz(nowy$\in$C) = 3/20

Druga szansa jest większa bo w pobliżu nowego obiektu więcej jest obiektów czerwonych.

Naiwny klasyfikator Bayesa

Argument maksimum (arg max lub argmax) to zbiór argumentów funkcji dla jakich osiąga ona maksimum:

$$\underset{x \in T}{\arg\max}\, f(x) = \left\{ x \in T \colon f(x) = \max_{t \in T} f(t) \right\}.$$

Przykłady

$$\underset{x \in (-1;1)}{\arg\max}\, x = \emptyset$$

$$\underset{x \in \langle -2\pi;\pi \rangle}{\arg\max}\, \sin(x) = \{-\tfrac{3\pi}{2}, \tfrac{\pi}{2}\}$$

$$\underset{x \in \mathbb{R}}{\arg\max}\, \cos(x) = \{\ldots, -4\pi, -2\pi, 0, 2\pi, 4\pi, \ldots\}$$

Naiwny klasyfikator Bayesa

Załóżmy, że chcemy dokonać klasyfikacji do jednej z klas $V=\{v_1,v_2,\ldots,v_m\}$.

Zgodnie z regułami klasyfikacji Bayesowskiej otrzymujemy:

$$V_{nb} = \operatorname{argmax}_{v_j \in V} P(v_j) \prod P(a_i|v_j)$$

gdzie:

$P(v_j)$ - prawdopodobieństwo, że $v = v_j$

$$P(a_i|v_j) = \frac{n_c}{n}$$

$n =$ the number of training examples for which $v = v_j$

$n_c =$ number of examples for which $v = v_j$ and $a = a_i$

Naiwny Klasyfikator Bayesa

Przykład (ponownie)

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Rozpatrzmy samochód: Red, SUV, Domestic

Chcemy go zakwalifikować do jednej z dwóch kategorii:

Stolen - Yes, Stolen - No

Naiwny Klasyfikator Bayesa

Przykład

Musimy policzyć:

$$v = \text{argmax}_{\,b \in \{Yes, No\}} \Pr(b) \prod_i \Pr(a_i \mid b)$$

No to policzmy:

iloczyn

$$v = \text{argmax}_{\,b \in \{Yes, No\}} \Pr(b) \cdot$$

$$\Pr(Red \mid b) \cdot$$

$$\Pr(Domestic \mid b) \cdot$$

$$\Pr(SUV \mid b) \cdot$$

Naiwny Klasyfikator Bayesa

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Przykład

Liczymy:

Pr(Yes) = 1/2 Pr(No) = 1/2

Pr(Red|Yes) = 3/5 Pr(Red|No) = 2/5

Pr(Domestic|Yes) = 2/5 Pr(Domestic|No) = 3/5

Pr(SUV|Yes) = 1/5 Pr(SUV|No) = 3/5

Wówczas:

$1/2 \cdot 3/5 \cdot 2/5 \cdot 1/5 = 6/250$ $1/2 \cdot 2/5 \cdot 3/5 \cdot 3/5 = 18/250$

Otrzymujemy odpowiedź No.

Naiwny Klasyfikator Bayesa

Uwaga

W powyższych wyliczeniach założyliśmy, że:

$$\Pr(A \mid B) = \frac{n_{A \wedge B}}{n_B}$$

gdzie:

- $n_{A \wedge B}$ to ilość przypadków $A \wedge B$
- n_B to ilość przypadków B

Problem może się pojawić wtedy gdy $n_{A \wedge B} = 0$.

Naiwny Klasyfikator Bayesa

Uwaga

ID	wiek	dochód	student	status	kupi_komputer
1	<=30	wysoki	nie	kawaler	nie
2	<=30	wysoki	nie	żonaty	nie
3	31..40	wysoki	nie	kawaler	tak
4	>40	średni	nie	kawaler	tak
5	>40	niski	tak	kawaler	tak
6	>40	niski	tak	żonaty	nie
7	31..40	niski	tak	żonaty	tak
8	<=30	średni	nie	kawaler	nie
9	<=30	niski	tak	kawaler	tak
10	>40	średni	tak	kawaler	tak
11	<=30	średni	tak	żonaty	tak
12	31..40	średni	nie	żonaty	tak
13	31..40	wysoki	tak	kawaler	tak
14	>40	średni	nie	żonaty	nie

P(wiek='31..40'|kupi_komputer='nie') = 0

Naiwny Klasyfikator Bayesa

<u>Uwaga</u>

W celu uniknięcia sytuacji w której licznik będzie równy 0 wprowadzamy dwie dodatkowe liczby p i m.

Przy czym:

- p – nasza estymacja Pr(A|B), zwykle przyjmujemy 1/(ilość wartości związanych z klasyfikacją)
- m – pewna stała

Zatem otrzymujemy:

$$\Pr(A \mid B) = \frac{n_{A \wedge B} + pm}{n_{B} + m}$$

Naiwny Klasyfikator Bayesa

Przykład (ponownie)

Example No.	Color	Type	Origin	Stolen?
1	Red	Sports	Domestic	Yes
2	Red	Sports	Domestic	No
3	Red	Sports	Domestic	Yes
4	Yellow	Sports	Domestic	No
5	Yellow	Sports	Imported	Yes
6	Yellow	SUV	Imported	No
7	Yellow	SUV	Imported	Yes
8	Yellow	SUV	Domestic	No
9	Red	SUV	Imported	No
10	Red	Sports	Imported	Yes

Rozpatrzmy samochód: Red, SUV, Domestic

Chcemy go zakwalifikować do jednej z dwóch kategorii:

Stolen - Yes, Stolen - No

Naiwny Klasyfikator Bayesa

Przykład (jeszcze raz!)

Liczymy:

$$\Pr(\text{Yes}) = 1/2 \qquad \Pr(\text{No}) = 1/2$$

$$\Pr(\text{Red}\,|\,\text{Yes}) = \frac{3+3\cdot 0.5}{5+3} \qquad \Pr(\text{Red}\,|\,\text{No}) = \frac{2+3\cdot 0.5}{5+3}$$

$$\Pr(\text{Domestic}\,|\,\text{Yes}) = \frac{2+3\cdot 0.5}{5+3} \qquad \Pr(\text{Domestic}\,|\,\text{No}) = \frac{3+3\cdot 0.5}{5+3}$$

$$\Pr(\text{SUV}\,|\,\text{Yes}) = \frac{1+3\cdot 0.5}{5+3} \qquad \Pr(\text{SUV}\,|\,\text{No}) = \frac{3+3\cdot 0.5}{5+3}$$

Wówczas:

0,037   0,069

Otrzymujemy także odpowiedź No.