Eksploracja danych

naiwny klasyfikator bayesowski

Piotr Lipiński

Naiwny klasyfikator bayesowski

- Klasyfikacja danych reprezentowanych przez d-wymiarowe wektory

  $$\mathbf{x} = (x_1, x_2, \ldots, x_d)$$

  z użyciem K klas $C_1, C_2, \ldots, C_K$

- Klasyfikator będzie definiowany przez warunkowe rozkłady prawdopodobieństwa (dla każdej klasy C_k):

  $$P(C_k \mid \mathbf{x}) = P(C_k \mid x_1, x_2, \ldots, x_d)$$

- Mając takie rozkłady, dla każdego wektora danych **x** będzie można wyznaczyć najbardziej prawdopodobną klasę (która będzie wynikiem klasyfikacji wektora danych **x**).

Naiwny klasyfikator bayesowski

- Korzystając z twierdzenia Bayesa mamy:

$$P(C_k \mid \mathbf{x}) = P(C_k)\, P(\mathbf{x} \mid C_k) / P(\mathbf{x})$$

- Mianownik tego ułamka nie jest istotny, bo nie zależy od klasy. Licznik natomiast można przedstawić jako rozkład łączny:

$$P(C_k, x_1, x_2, \ldots, x_d)$$

- Rozkład łączny można rozpisać jako:

$P(C_k, x_1, x_2, \ldots, x_d) =$

$P(C_k)\, P(x_1, x_2, \ldots, x_d \mid C_k) =$

$P(C_k)\, P(x_1 \mid C_k)\, P(x_2, \ldots, x_d \mid C_k, x_1) =$

$P(C_k)\, P(x_1 \mid C_k)\, P(x_2 \mid C_k, x_1)\, P(x_2, \ldots, x_d \mid C_k, x_1, x_2) = \ldots =$

$P(C_k)\, P(x_1 \mid C_k)\, P(x_2 \mid C_k, x_1) \ldots P(x_d \mid C_k, x_1, x_2, \ldots, x_{d-1})$

Naiwny klasyfikator bayesowski

- Jeżeli założymy (naiwnie) warunkową niezależność zmiennych losowych $x_1, x_2, \ldots, x_d$, to:

$P(x_2 \mid C_k, x_1) = P(x_2 \mid C_k)$

$P(x_3 \mid C_k, x_1, x_2) = P(x_3 \mid C_k)$

$\ldots$

$P(x_d \mid C_k, x_1, x_2, \ldots, x_{d-1}) = P(x_d \mid C_k)$

- Zatem rozkład łączny upraszcza się do:

$P(C_k, x_1, x_2, \ldots, x_d) = P(C_k)\, P(x_1 \mid C_k)\, P(x_2 \mid C_k) \ldots P(x_d \mid C_k)$

Naiwny klasyfikator bayesowski

- Konstrukcja klasyfikatora wymaga więc wyznaczenia:
  - $P(C_k)$
    - można to estymować na podstawie zbioru danych uczących (częstość występowania poszczególnych klas w zbiorze danych uczących)
  - $P(x_i \mid C_k)$
    - dla prostych (dyskretnych) przypadków można to estymować na podstawie zbioru danych uczących
    - dla bardziej złożonych przypadków należy przyjąć jakiś model tego rozkładu prawdopodobieństwa (popularne są rozkłady gaussowskie, dwumianowe, Bernoulliego, itp.)

Przykład

- Celem jest klasyfikacja owoców na podstawie ich cech. Skupiamy się na rozpoznawaniu banana, pomarańczy i innego owocu (trzy klasy) na podstawie trzech cech binarnych: długi, słodki, żółty.
- Dane uczące zawierają informacje o 1000 owoców opisanych przez trzy cechy (długi, słodki, żółty).
- Tabela poniżej zawiera podsumowanie zbioru danych uczących.

typ	długi	niedługi	słodki	niesłodki	żółty	nieżółty	razem
banan	400	100	350	150	450	50	500
pomarańcza	0	300	150	150	300	0	300
inny	100	100	150	50	50	150	200
razem	500	500	650	350	800	200	1000

Przykład

- Prawdopodobieństwa bezwarunkowe (prior probabilities)
  - P(banan) = 500 / 1000 = 0.50
  - P(pomarańcza) = 300 / 100 = 0.30
  - P(inny) = 200 / 1000 = 0.20
- Prawdopodobieństwa zdarzeń (evidence probabilities)
  - P(długi) = 500 / 1000 = 0.50
  - P(słodki) = 650 / 1000 = 0.65
  - P(żółty) = 800 / 1000 = 0.80

typ	długi	niedługi	słodki	niesłodki	żółty	nieżółty	razem
banan	400	100	350	150	450	50	500
pomarańcza	0	300	150	150	300	0	300
inny	100	100	150	50	50	150	200
razem	500	500	650	350	800	200	1000

Przykład

- Prawdopodobieństwa warunkowe
  - P(długi | banan) = 400 / 500 = 0.80
  - P(długi | pomarańcza) = 0 / 300 = 0
  - P(długi | inny) = 100 / 200 = 0.50
  - P(słodki | banan) = 350 / 500 = 0.70
  - ....
  - P(żółty | banan) = 450 / 500 = 0.90
  - ....
  - (należy wyliczyć wszystkie prawdopodobieństwa warunkowe $P(x_i \mid C_k)$, dla cech x_i = długi, słodki, żółty i klas C_k = banan, pomarańcza, inny)
- Uczenie naiwnego klasyfikatora bayesowskiego to właśnie wyliczenie tych prawdopodobieństw.

typ	długi	niedługi	słodki	niesłodki	żółty	nieżółty	razem
banan	400	100	350	150	450	50	500
pomarańcza	0	300	150	150	300	0	300
inny	100	100	150	50	50	150	200
razem	500	500	650	350	800	200	1000

Przykład

- Nauczonego naiwnego klasyfikara bayesowskiego można użyć do rozpoznawania nieznanych owoców (opisanych przez rozważane trzy cechy).
- Zastanówmy się czym może być owoc długi, słodki i żółty. Są trzy możliwości - może to być banan, pomarańcza lub inny owoc. Będziemy rozpatrywać więc

  P(banan | długi, słodki, żółty)

  P(pomarańcza | długi, słodki, żółty)

  P(inny | długi, słodki, żółty)
- P(banan | długi, słodki, żółty) =

$$\frac{P(\text{banan})\,P(\text{długi} \mid \text{banan})\,P(\text{słodki} \mid \text{banan})\,P(\text{żółty} \mid \text{banan})}{P(\text{długi})\,P(\text{słodki})\,P(\text{żółty})} =$$

$$\frac{0.50 * 0.80 * 0.70 * 0.90}{0.50 * 0.65 * 0.80} =$$

0.252000 / 0.260000 = 0.969231

Przykład

- P(pomarańcza | długi, słodki, żółty) = 0

  bo P(długi | pomarańcza) = 0
- P(inny | długi, słodki, żółty) =

$$\frac{0.20 * 0.50 * 0.75 * 0.25}{0.50 * 0.65 * 0.80} =$$

0.018750 / 0.260000 = 0.072115
- Ostatecznie więc owoc długi, słodki i żółty jest najprawdopodobniej bananem.
- Naturalnie, jak łatwo zauważyć, liczenie mianownika było niepotrzebne, bo do porównywania prawdopodobieństw warunkowych wystarcza sam licznik.
- Na czym polegała „naiwność” klasyfikatora bayesowskiego w tym przykładzie?