Klasyfikacja Bayesowska

Wykład wer. 2020

Podstawowe metody (klasyfikacyjne)

- metody symboliczne (drzewa i reguły decyzyjne),
- metody oparte na logice matematycznej (ILP),
- sztuczne sieci neuronowe,
- metody k-najbliższych sąsiadów,
- **klasyfikacja bayesowska (Naive Bayes),**
- analiza dyskryminacyjna (statystyczna),
- metody wektorów wspierających,
- regresja logistyczna,
- klasyfikatory genetyczne.
- ...

Klasyfikacja Bayesowska

- Prob. Tw. Bayesa: Pozwala na obliczanie prawdopodobieństw związanych z hipotezami; efektywne podejście do wielu problemów praktycznych.
- Przyrostowość: Każdy nowy przykład może zmienić oszacowanie prawdopodobieństw; Możliwość uwzględniania prawdopodobieństw apriori.
- Predykcja probabilistyczna: Predykcja wielu hipotez, "ważonych" za pomocą prawdopodobieństw.
- Standard: Pomimo, że w pewnych przypadkach metody bayesowskie są nieatrakcyjne obliczeniowo, to mogą dostarczać tzw. standardu optymalnych decyzji, z którym można porównywać inne metody.

Reguła Bayesa (przypomnienie)

- Probability of event *H* given evidence *E* :

$$\Pr[H \mid E] = \frac{\Pr[E \mid H]\Pr[H]}{\Pr[E]}$$

- *A priori* probability of *H* : $\Pr[H]$
  - Probability of event *before* evidence is seen
- *A posteriori* probability of *H* : $\Pr[H \mid E]$
  - Probability of event *after* evidence is seen

from Bayes "Essay towards solving a problem in the doctrine of chances" (1763)

Thomas Bayes

Born: 1702 in London, England
Died: 1761 in Tunbridge Wells, Kent, England

Twierdzenie Bayes'a

- Dla zbioru uczącego *D, prawdopodobieństwo posteriori hipotezy h, P(h|D)* wynika z twierdzenia Bayes'a:

$$P(h|D) = \frac{P(D|h)P(h)}{P(D)}$$

- MAP (maximum posteriori) hypothesis:

$$h_{MAP} \equiv \arg\max_{h \in H} P(h|D) = \arg\max_{h \in H} P(D|h)P(h).$$

- Praktyka: założenie znajomości prawdo-podobieństw *a priori* – trudności z dużą liczbą reprezentatywnych przykładów

Klasyfikator baysowski

- Przykłady uczące opisane zbiorem atrybutów warunkowych *a1*,*a2*,…,*am*; mogą być należeć do jednej z klas ze zbioru **K** = {*Kj* : *j* = 1,...,*r*}
- Nowy (klasyfikowany) obiekt opisany wartościami atrybutów warunkowych ⟨*a1* = *v1*, *a2* = *v2*,…, *am* = *vm*⟩
- Przydziel obiekt do najbardziej prawdopodobnej klasy, tzn. do klasy K_{MAP} która maksymalizuje:

$$K_{MAP} = \arg\max_{K_j \in \mathbf{K}} P(K_j \mid a_1 = v_1, \ldots, a_m = v_m)$$

$$K_{MAP} = \arg\max_{K_j \in \mathbf{K}} \frac{P(a_1 = v_1, \ldots, a_m = v_m \mid K_j) \cdot P(K_j)}{P(a_1 = v_1, \ldots, a_m = v_m)}$$

$$\arg\max_{K_j \in \mathbf{K}} P(a_1 = v_1, \ldots, a_m = v_m \mid K_j) \cdot P(K_j)$$

Naiwny klasyfikator bayesowski

- Założenie upraszczające: niezależność warunkowa atrybutów a_i ($i = 1,\ldots, n$):

$$P(K_j \mid a_1 = v_1, \cdots, a_m = v_m) = P(K_j) \prod_{i=1}^{m} (P(a_i = v_i \mid K_j)$$

- Znacznie obniża wymogi co do estymacji prawdopodobieństw *a priori* , niższe koszty obliczeniowe.

Zbiór uczący

- Przykład Quinlan'a (*Play a game*).

Outlook	Tempreature	Humidity	Windy	Class
sunny	hot	high	false	N
sunny	hot	high	true	N
overcast	hot	high	false	P
rain	mild	high	false	P
rain	cool	normal	false	P
rain	cool	normal	true	N
overcast	cool	normal	true	P
sunny	mild	high	false	N
sunny	cool	normal	false	P
rain	mild	normal	false	P
sunny	mild	normal	true	P
overcast	mild	high	true	P
overcast	hot	normal	false	P
rain	mild	high	true	N

Naiwny klasyfikator bayesowski (II)

- Dla danego przykładu, można obliczyć prawdopodobieństwa

Outlook	P	N
sunny	2/9	3/5
overcast	4/9	0
rain	3/9	2/5
Tempreature		
hot	2/9	2/5
mild	4/9	2/5
cool	3/9	1/5

Humidity	P	N
high	3/9	4/5
normal	6/9	1/5
Windy		
true	3/9	3/5
false	6/9	2/5

Analiza przykładu

Outlook			Temperature			Humidity			Windy			Play	
	Yes	*No*		*Yes*	*No*		*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	Hot	2	2	High	3	4	False	6	2	9	5
Overcast	4	0	Mild	4	2	Normal	6	1	True	3	3		
Rainy	3	2	Cool	3	1								
Sunny	2/9	3/5	Hot	2/9	2/5	High	3/9	4/5	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	Mild	4/9	2/5	Normal	6/9	1/5	True	3/9	3/5		
Rainy	3/9	2/5	Cool	3/9	1/5								

Outlook	Temp.	Humidity	Windy	Play
Sunny	Cool	High	True	?

Nowy przykład:

Likelihood of the two classes

For “yes” = 2/9 × 3/9 × 3/9 × 3/9 × 9/14 = 0.0053

For “no” = 3/5 × 1/5 × 4/5 × 3/5 × 5/14 = 0.0206

Conversion into a probability by normalization:

P(“yes”) = 0.0053 / (0.0053 + 0.0206) = 0.205

P(“no”) = 0.0206 / (0.0053 + 0.0206) = 0.795

The “zero-frequency problem”

- Co zrobimy gdy wartość atrybutu nie występi w kombinacji z pewną klasą?
  (np. “Humidity = high” for class “yes”)
  - Prawdopodob. zero! $\Pr[Humidity = High \mid yes] = 0$
  - *A posteriori* probability także zero! $\Pr[yes \mid E] = 0$
    (Niezależnie od innych wartości!)
- Rozwiązania: dodaj 1 do licznika dla każej value-class pary (*Laplace estimator)*
- Rezultat: probabilities will never be zero!
  (also: stabilizes probability estimates)
- Możliwe inne estymaty

*Modified probability estimates

- In some cases adding a constant different from 1 might be more appropriate
- Example: attribute *outlook* for class *yes*

$$\frac{2+\mu/3}{9+\mu} \qquad \frac{4+\mu/3}{9+\mu} \qquad \frac{3+\mu/3}{9+\mu}$$

Sunny ***Overcast*** ***Rainy***

- Weights don't need to be equal (but they must sum to 1)

$$\frac{2+\mu p_1}{9+\mu} \qquad \frac{4+\mu p_2}{9+\mu} \qquad \frac{3+\mu p_3}{9+\mu}$$

Missing values

- Uczenie: przykład – z niezdefiniowanymi wartościami nie uwzględniamy w oszacowaniu częstości
- Predykcja: atrybut pomijany w obliczeniach
- Przykład:

Outlook	Temp.	Humidity	Windy	Play
?	Cool	High	True	?

Likelihood of “yes” = 3/9 × 3/9 × 3/9 × 9/14 = 0.0238
Likelihood of “no” = 1/5 × 4/5 × 3/5 × 5/14 = 0.0343
P(“yes”) = 0.0238 / (0.0238 + 0.0343) = 41%
P(“no”) = 0.0343 / (0.0238 + 0.0343) = 59%

Atrybuty liczbowe

- Typowe założenie – przyjmij wybrany rozkład, np. normalny
- PDF – funkcja rozkładu normalnego zdefiniowana przez dwa parametry (wartość oczekiwana i odchylenie):
  - *Sample mean* μ

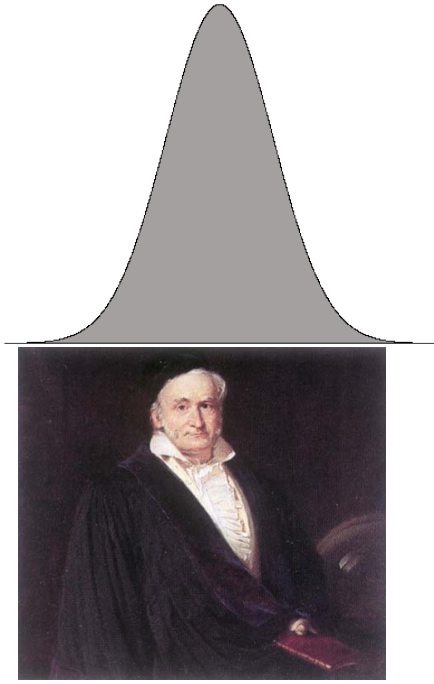
$$\mu = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - *Standard deviation* σ

$$\sigma = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \mu)^2$$

  - Then the density function *f(x) is*

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Karl Gauss, 1777-1855
great German mathematician

Przeliczenie przykładu

Outlook			Temperature		Humidity		Windy			Play	
	Yes	*No*	*Yes*	*No*	*Yes*	*No*		*Yes*	*No*	*Yes*	*No*
Sunny	2	3	64, 68,	65, 71,	65, 70,	70, 85,	False	6	2	9	5
Overcast	4	0	69, 70,	72, 80,	70, 75,	90, 91,	True	3	3		
Rainy	3	2	72, …	85, …	80, …	95, …					
Sunny	2/9	3/5	μ =73	μ =75	μ =79	μ =86	False	6/9	2/5	9/14	5/14
Overcast	4/9	0/5	σ =6.2	σ =7.9	σ =10.2	σ =9.7	True	3/9	3/5		
Rainy	3/9	2/5									

- Wykorzystaj funkcję gęstości:

$$f(temperature = 66 \mid yes) = \frac{1}{\sqrt{2\pi}\,6.2} e^{-\frac{(66-73)^2}{2*6.2^2}} = 0.0340$$

Klasyfikacja nowych faktów

- Nowe dane:

Outlook	Temp.	Humidity	Windy	Play
Sunny	66	90	true	?

Likelihood of “yes” = 2/9 × 0.0340 × 0.0221 × 3/9 × 9/14 = 0.000036
Likelihood of “no” = 3/5 × 0.0291 × 0.0380 × 3/5 × 5/14 = 0.000136
P(“yes”) = 0.000036 / (0.000036 + 0. 000136) = 20.9%
P(“no”) = 0.000136 / (0.000036 + 0. 000136) = 79.1%

Naïve Bayes: dyskusja

- Naïve Bayes works surprisingly well (even if independence assumption is clearly violated)
- Why? Because classification doesn't require accurate probability estimates *as long as maximum probability is assigned to correct class*
- However: adding too many redundant attributes will cause problems (e.g. identical attributes)
- Note also: many numeric attributes are not normally distributed (→ *kernel density estimators*)

Naïve Bayes rozszerzenia

- Improvements:
  - select best attributes (e.g. with greedy search)
  - often works as well or better with just a fraction of all attributes
- Bayesian Networks

Rozszerzymy informacje o k-NN kolejny wykład