Uniwersytet Przyrodniczy we Wrocławiu

Data Mining
Wykład 5

Naiwny klasyfikator Bayes'a

Naiwny klasyfikator Bayesa

- Naiwny klasyfikator Bayesa jest klasyfikatorem statystycznym - oparty na twierdzeniu Bayesa
- Niech X oznacza przykład, którego klasa nie jest znana. Każdy przykład jest reprezentowany w postaci n-wymiarowego wektora, $X=(x_1, x_2, \ldots, x_n)$
- P(C|X) prawdopodobieństwo a-posteriori (prawdopodobieństwo obliczane na podstawie wyników doświadczenia, czyli częstości), że przykład X należy do klasy C

Reguła Bayesa

Przykład X klasyfikujemy jako pochodzący z tej klasy C_i, dla której wartość $P(C_i|X)$, i = 1, 2, ..., m, jest największa.

Naiwny klasyfikator Bayesa - Przykła

- Przykład: Dany zbiór przykładów opisujących wnioski kredytowe klientów banku:

P(Ryzyko=niskie | Wiek=38, Status=rozwodnik, Dochód=niski, Dzieci=2)

- oznacza prawdopodobieństwo a-posteriori, że klient, X=(38, rozwodnik, niski, 2), składający wniosek kredytowy jest klientem o niskim ryzyku kredytowym (klient wiarygodny)

Twierdzenie Bayesa

- W jaki sposób oszacować prawdopodobieństwo a-posteriori P(C|X)?

P(C|X) = (P(X|C) * P(C))/P(X),

- P(C) oznacza prawdopodobieństwo a-priori wystąpienia klasy C (tj. prawdopodobieństwo, że dowolny przykład należy do klasy C),
- P(X|C) oznacza prawdopodobieństwo a-posteriori, że X należy do klasy C,
- P(X) oznacza prawdopodobieństwo a-priori wystąpienia przykładu X

Naiwny klasyfikator Bayesa (1)

- Dany jest zbiór treningowy D składający się z n przykładów
- Załóżmy, że atrybut decyzyjny przyjmuje m różnych wartości definiując m różnych klas C_i , i = 1, ..., m
- Niech si oznacza liczbę przykładów z D należących do klasy C_i
- Klasyfikator Bayesa przypisuje nieznany przykład X do tej klasy C_i, dla której wartość $P(C_i|X)$ jest największa

Naiwny klasyfikator Bayesa (2)

- Prawdopodobieństwo P(X) jest stałe dla wszystkich klas - klasa C_i , dla której wartość $P(C_i | X)$ jest największa, to klasa C_i , dla której wartość $P(X | C_i) * P(C_i)$ jest największa
- Wartości $P(C_i)$ zastępujemy estymatorami s_i/n (względną częstością klasy C_i), lub zakładamy, że wszystkie klasy mają to samo prawdopodobieństwo $P(C_1) = P(C_2) = ... = P(C_m)$

Uniwersytet Przyrodniczy we Wrocławiu

Naiwny klasyfikator Bayesa (3)

- W jaki sposób obliczyć $P(X|C_i)$?
- Dla dużych zbiorów danych, o dużej liczbie deskryptorów, obliczenie $P(X|C_i)$ będzie bardzo kosztowne
- Wymaga ono oszacowania ogromnej liczby prawdopodobieństw i jest rzędu k^p, gdzie p oznacza zmienne, natomiast k oznacza liczbę wartości tych zmiennych np. dla p=30 -> 2^{30} czyli około 10^9
- Przyjmując założenie o niezależności atrybutów, możemy przyjąć, że wszystkie zmienne są warunkowo niezależne przy danych klasach. Wówczas możemy zastąpić prawdopodobieństwo warunkowe P(X|Ci) iloczynem prawdopodobieństw

$$P(X|C_i) = \prod_{j=1}^{n} P(x_j | C_i)$$

Uniwersytet Przyrodniczy we Wrocławiu

Naiwny klasyfikator Bayesa (4)

- Prawdopodobieństwa $P(x_1|C_i)$, $P(x_2|C_i)$, ..., $P(x_n|C_i)$ można estymować w oparciu o zbiór treningowy następująco:

jeżeli j-ty atrybut jest atrybutem kategorycznym, to $P(x_j | C_i)$ estymujemy względną częstością występowania przykładów z klasy C_i posiadających wartość x_j dla j-tego atrybutu, (s_{ij}/s_i)

jeżeli j-ty atrybut jest atrybutem ciągłym, to $P(x_j | C_i)$ estymujemy funkcją gęstości Gaussa

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

(zakładając rozkład normalny wartości atrybutów)

Uniwersytet Przyrodniczy we Wrocławiu

Przykład (1)

- Rozważmy Przykład:

  Chcemy dokonać predykcji klasy, do której należy nowy przypadek
  - C1 (kupi_ komputer ='tak')
  - C2 (kupi _ komputer ='nie')

- Nowy przypadek:
  - X = (wiek='<=30', dochód='średni', student = 'tak', status='kawaler')
  - Maksymalizujemy wartość $P(X/C_i)*P(C_i)$, dla i=1,2

Przykład (2)

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

Przykład (3)

P(kupi_komputer = 'tak') = P(C1) = 9/14 = 0.643

P(kupi_ komputer = 'nie') = P(C2) = 5/14= 0.357

P(wiek <= '30' | kupi_ komputer = 'tak') = 2/9 = 0.222
P(wiek <= '30' | kupi_ komputer = 'nie') = 3/5 = 0.6
P(dochód = 'średni' | kupi_ komputer = 'tak') = 4/9 = 0.444
P(dochód = 'średni' | kupi_ komputer = 'nie') = 2/5 = 0.4
P(student = 'tak' | kupi_ komputer = 'tak') = 6/9 = 0.667
P(student = 'tak' | kupi_ komputer = 'nie') = 1/5 = 0.2
P(status = 'kawaler' | kupi_ komputer = 'tak') = 6/9 = 0.667
P(status ='kawaler' | kupi_ komputer = 'nie') = 2/9 = 0.4

Przykład (4)

Korzystając z obliczonych prawdopodobieństw, otrzymujemy:

P(X| kupi_komputer='tak') = 0.222 * 0.444 * 0.667 * 0.667 = 0.044
P(X| kupi_komputer='nie') = 0.600 * 0.400 * 0.200 * 0.400 = 0.019

Stąd:

P(X| kupi_ komputer='tak') * P(kupi_ komputer='tak') = 0.044 * 0.643 **= 0.028**

P(X| kupi_ komputer='nie') * P(kupi_ komputer='nie') = 0.019 * 0.357 **= 0.007**

Naiwny klasyfikator Bayesa zaklasyfikuje nowy przypadek X do klasy:

kupi_komputer = 'tak'

Problem „częstości zero"

A co jeżeli dana wartość atrybutu nie występuje dla wszystkich klas?

Przykładowo: wiek='31..40' dla klasy „nie'

– Prawdopodobieństwo wynosi 0, tj. P(wiek='31..40'|kupi_komputer='nie') = 0

– A-posteriori prawdopodobieństwo również wynosi 0

Rozwiązanie:

dodać 1 do licznika wystąpień każdej pary <wartość atrybutu – klasa> (estymator Laplace'a)

Podsumowanie - Naiwny klasyfikator Bayesa

- Założenie o niezależności atrybutów znacznie redukuje koszt obliczeń
- Jeżeli założenie jest spełnione, naiwny klasyfikator Bayes'a jest optymalny, tzn. zapewnia najlepszą dokładność klasyfikacji w porównaniu z innymi klasyfikatorami
- Założenie rzadko spełnione w praktyce – jednakże naiwny klasyfikator Bayes'a jest zadziwiająco dokładny