Klasyfikacja metodą Bayesa

Tadeusz Pankowski
www.put.poznan.pl/~tadeusz.pankowski

Klasyfikacja metodą Bayesa – prawdopodobieństwo warunkowe i bezwarunkowe

1. Klasyfikacja Bayesowska jest klasyfikacją statystyczną. Pozwala przewidzieć prawdopodobieństwo przynależności obiektu do klasy. Opiera się na *twierdzeniu Bayesa*.
2. Rozważamy hipotezę, że obiekt o właściwości X należy do klasy C.
3. Z punktu widzenia zadania klasyfikacji chcemy obliczyć prawdopodobieństwo P(C|X) tego, że obiekt o właściwości X należy do klasy C.

Klasyfikacja metodą Bayesa – prawdopodobieństwo warunkowe i bezwarunkowe (c.d.)

- P(C|X) jest tzw. *prawdopodobieństwem warunkowym (a posteriori)* zdarzenia C (należenie do klasy C) pod warunkiem zajścia zdarzenia X (posiadanie właściwości X).
- Przykład:
  Rozważamy populację owoców. Przypuśćmy, że rozważaną właściwością X obiektu *o* jest:

  X : „*o* jest owocem okrągłym i czerwonym”,

  a C jest klasą *jabłek*.

  Wówczas P(C|X) wyraża nasze przekonanie, że owoc jest jabłkiem, jeśli zaobserwowaliśmy, że jest on okrągły i czerwony.

Klasyfikacja metodą Bayesa – prawdopodobieństwo warunkowe i bezwarunkowe (c.d.)

- Przeciwieństwem prawdopodobieństwa *warunkowego* jest prawdopodobieństwo *bezwarunkowe*. W naszym przypadku P(C) wyraża prawdopodobieństwo, że obserwowany owoc jest jabłkiem. Prawdopodobieństwo to nazywamy prawdopodobieństwem *bezwarunkowym (a priori*).
- Prawdopodobieństwo warunkowe, P(C|X), uwzględnia więcej informacji (wiedza X), podczas gdy prawdopodobieństwo bezwarunkowe P(C) jest niezależne od X.

Klasyfikacja metodą Bayesa – prawdopodobieństwo warunkowe i bezwarunkowe (c.d.)

- Podobnie, P(X|C) jest prawdopodobieństwem warunkowym zajścia zdarzenia X pod warunkiem zajścia zdarzenia C. Jest więc na przykład prawdopodobieństwem, że owoc jest okrągły i czerwony, jeśli wiadomo, że jest jabłkiem.
- P(X) jest prawdopodobieństwem bezwarunkowym. Oznacza więc na przykład prawdopodobieństwem tego, że obserwowany (wybrany losowo) owoc jest czerwony i okrągły.

Klasyfikacja metodą Bayesa – twierdzenie Bayesa

- Twierdzenia Bayesa pokazuje, w jaki sposób obliczyć prawdopodobieństwo warunkowe P(C|X), jeśli znane są prawdopodobieństwa: warunkowe P(X|C) oraz bezwarunkowe P(C) i P(X).
- Prawdopodobieństwa: P(X|C), P(C) oraz P(X) mogą być bezpośrednio wyliczone z danych zgromadzonych w treningowym zbiorze danych (w bazie danych).
- Twierdzenie Bayesa:

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)}$$

Twierdzenie Bayesa

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)}$$

Dowód (z prawdopodobieństwa warunkowego):

$$P(C \cap X) = P(X)P(C \mid X) = P(C)P(X \mid C)$$

$$P(C \mid X) = \frac{P(X \mid C)P(C)}{P(X)}$$

Klasyfikacja Bayesa (naiwna)

- Każdy obiekt traktowany jest jako wektor X (krotka) wartości atrybutów A_1, ..., A_n:
  $X = (x_1, x_2, ..., x_n)$.
- Niech C_1, ..., C_m będą klasami, do których może należeć X, P(C|X) niech oznacza prawdopodobieństwo przynależności X (ściślej: obiektów o właściwości X) do klasy C.
- W klasyfikacji Bayesa przypisujemy X do tej klasy, do której prawdopodobieństwo warunkowe przynależności X jest największe.
- X jest więc przypisany do C_i, jeśli
  $P(C_i|X) \geq P(C_k|X)$, dla każdego k, $1 \leq k \leq m$, $k \neq i$.

Klasyfikacja Bayesa

1. W klasyfikacji Bayesa maksymalizujemy: $P(C_i \mid X) = \frac{P(X \mid C_i)P(C_i)}{P(X)}$
2. Ponieważ P(X) jest stałe, więc wystarczy maksymalizować iloczyn
   $P(X|C_i)P(C_i)$.
3. Ponadto przyjmujemy:
   $P(C_i) = s_i / s$,
   gdzie s oznacza liczbę obiektów w zbiorze treningowym, a s_i oznacza liczbę obiektów w klasie C_i.
4. Dla $X = (x_1, x_2, ..., x_n)$, wartość $P(X|C_i)$ obliczamy jako iloczyn:
   $P(X|C_i) = P(x_1|C_i)*P(x_2|C_i)* ... *P(x_n|C_i)$,
   przy czym:
   $P(x_k|C_i) = s_{ik} / s_{i,}$
   gdzie s_{ik} oznacza liczbę obiektów klasy C_i, dla których wartość atrybutu A_k jest równa x_k, a s_i oznacza liczbę wszystkich obiektów klasy C_i w zadanym zbiorze treningowym.

Przykład klasyfikacji metodą Bayesa

Algorytm: Klasyfikacja obiektów metodą Bayesa.

- *Wejście*:
  - Rozważana populacja obiektów (klientów) opisana jest za pomocą czterech atrybutów: *Wiek*, *Dochód*, *Studia*, *OcenaKred.*
  - Interesuje nas przynależność obiektów do jednej z dwóch klas: *klienci kupujący komputery* (o etykiecie TAK) i *klienci nie kupujący komputerów* (o etykiecie NIE).
  - Z bazy danych wybrano zbiór treningowy (rozważany przez nas przy budowie drzew decyzyjnych – następny slajd).
  - Obiekt *X* o nieznanej przynależności klasowej ma postać:
    X = (Wiek = „<=30”, Dochód = „średni”, Student = „tak”, OcenaKred = „dobra”)
- *Wyjście:*
  - Określić przynależność obiektu *X* do klasy C_1 („tak”) lub C_2 („nie”) za pomocą klasyfikacji Bayesa.

Przykład klasyfikacji metodą Bayesa

Zbiór treningowy:

ID	Wiek	Dochód	Studia	OcenaKred	ZakupKomp
1	<=30	wysoki	nie	dobra	nie
2	<=30	wysoki	nie	znakomita	nie
3	31..40	wysoki	nie	dobra	tak
4	>40	średni	nie	dobra	tak
5	>40	niski	tak	dobra	tak
6	>40	niski	tak	znakomita	nie
7	31..40	niski	tak	znakomita	tak
8	<=30	średni	nie	dobra	nie
9	<=30	niski	tak	dobra	tak
10	>40	średni	tak	dobra	tak
11	<=30	średni	tak	znakomita	tak
12	31..40	średni	nie	znakomita	tak
13	31..40	wysoki	tak	dobra	tak
14	>40	średni	nie	znakomita	nie

Klasyfikowany obiekt:
X = (Wiek = „<=30”, Dochód = „średni”, Student = „tak”, OcenaKred = „dobra”)

Przykład klasyfikacji metodą Bayesa (c.d.)

1. Należy obliczyć, dla jakiej wartości i, i =1, 2, iloczyn
   $P(X|C_i)*P(C_i)$,
   osiąga maksimum.
2. $P(C_i)$ oznacza prawdopodobieństwo bezwarunkowe przynależności obiektu do klasy (inaczej: prawdopodobieństwo klasy) C_i, i = 1, 2.
   Ze zbioru treningowego obliczamy:
   $P(C_1) = 9/14 = 0.643$
   $P(C_2) = 5/14 = 0.357$
3. Prawdopodobieństwa warunkowe $P(X|C_i)$ są odpowiednio równe iloczynom prawdopodobieństw warunkowych:
   $P(X|C_1)$ = P(Wiek=„<=30”|C_1) * P(Dochód=„średni”|C_1)
   P(Studia=„tak”|C_1) * P(OcenaKred=„dobra”|C_1),
   $P(X|C_2)$ = P(Wiek=„<=30”|C_2) * P(Dochód=„średni”|C_2)
   P(Studia=„tak”|C_2) * P(OcenaKred=„dobra”|C_2),

Przykład klasyfikacji metodą Bayesa (c.d.)

ID	Wiek	Dochód	Studia	OcenaKred	ZakupKomp
1	<=30	wysoki	nie	dobra	nie
2	<=30	wysoki	nie	znakomita	nie
3	31..40	wysoki	nie	dobra	tak
4	>40	średni	nie	dobra	tak
5	>40	niski	tak	dobra	tak
6	>40	niski	tak	znakomita	nie
7	31..40	niski	tak	znakomita	tak
8	<=30	średni	nie	dobra	nie
9	<=30	niski	tak	dobra	tak
10	>40	średni	tak	dobra	tak
11	<=30	średni	tak	znakomita	tak
12	31..40	średni	nie	znakomita	tak
13	31..40	wysoki	tak	dobra	tak
14	>40	średni	nie	znakomita	nie

Ze zbioru treningowego obliczamy:

P(Wiek=„<=30”|C_1) = 2/9 = 0.222
P(Dochód=„średni”|C_1) = 4/9 = 0.444
P(Studia=„tak”|C_1) = 6/9 = 0.667
P(OcenaKred=„dobra”|C_1) = 6/9 = 0.667
P(Wiek=„<=30”|C_2) = 3/5 = 0.600
P(Dochód=„średni”|C_2) = 2/5 = 0,400
P(Studia=„tak”|C_2) = 1/5 = 0.200
P(OcenaKred=„dobra”|C_2) = 2/5 = 0.400

Stąd:

$P(X|C_1) = 0.222*0.444*0.667*0.667 = 0.044$

$P(X|C_1)P(C_1) = 0.044*0.643 = 0.028$

$P(X|C_2) = 0.600*0.400*0.200*0.400 = 0.019$

$P(X|C_2)P(C_2) = 0.019*0.357 = 0.007$

X – został zaklasyfikowany do C_1.

Inne zastosowania – przykład

Twierdzenia Bayesa można użyć do interpretacji rezultatów badania przy użyciu testów wykrywających narkotyki. Załóżmy, że przy badaniu narkomana test wypada pozytywnie w 99% przypadków, zaś przy badaniu osoby nie zażywającej narkotyków wypada negatywnie w 99% przypadków.

Pewna firma postanowiła przebadać swoich pracowników takim testem wiedząc, że 0,5% z nich to narkomani. Chcemy obliczyć prawdopodobieństwo, że osoba u której test wypadł pozytywnie rzeczywiście zażywa narkotyki. Oznaczmy następujące zdarzenia:

- D - dana osoba jest narkomanem
- N - dana osoba nie jest narkomanem
- + - u danej osoby test dał wynik pozytywny
- – - u danej osoby test dał wynik negatywny

Inne zastosowania – przykład

$P(D)$ = 0,005, gdyż 0,5% pracowników to narkomani
$P(N) = 1 - P(D)$ = 0,995
$P(+ \mid D)$ = 0,99, skuteczność testu przy badaniu narkomana
$P(- \mid N)$ = 0,99, gdyż taką skuteczność ma test przy badaniu osoby nie będacej narkomanem
$P(+ \mid N) = 1 - P(- \mid N)$ = 0,01
Mając te dane chcemy obliczyć prawdopodobieństwo, że osoba u której test wypadł pozytywnie, rzeczywiście jest narkomanem. Tak więc:

$$P(D|+) = \frac{P(+|D)P(D)}{P(+)} = \frac{P(+|D)P(D)}{P(+|D)P(D) + P(+|N)P(N)} = \frac{0,99 \cdot 0,005}{0,99 \cdot 0,005 + 0,01 \cdot 0,995} = 0,3322$$

Inne zastosowania – przykład

Mimo potencjalnie wysokiej skuteczności testu, prawdopodobieństwo, że narkomanem jest badany pracownik u którego test dał wynik pozytywny, jest równe około 33%, więc jest nawet bardziej prawdopodobnym, ze taka osoba nie zażywa narkotyków.

Ten przykład pokazuje, dlaczego ważne jest, aby nie polegać na wynikach tylko pojedynczego testu.

Przykład 2

Wśród studentów jest 60% mężczyzn i 40% kobiet.
Połowa studentek chodzi w spodniach, a połowa w spódniczkach. Wszyscy studenci noszą spodnie.
Obserwator widzi osobę w spodniach. Jakie jest prawdopodobieństwo, że jest to studentka?

P(A) = 0.4 – prawdopodobieństwo, że osoba jest studentką,
P(A') = 0.6 – prawdopodobieństwo, że osoba jest studentem,
P(B|A) = 0.5 – prawdopodobieństwo, że osoba będąca studentką nosi spodnie,
P(C|A) = 0.5 – prawdopodobieństwo, że osoba będąca studentką nosi spódnicę,
P(B|A') = 1 – prawdopodobieństwo, że osoba będąca studentem nosi spodnie.

Obliczyć:
P(A|B) – prawdopodobieństwo, że osoba w spodniach jest studentką

Obliczamy:
P(B) – prawdopodobieństwo, że losowo wybrana osoba nosi spodnie
P(B) = P(B|A)P(A) + P(B|A')P(A') = 0.5×0.4 + 1×0.6 = 0.8

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} = \frac{0.5 \times 0.4}{0.8} = 0.25.$$