Naive Bayes Classifier

1. Wet van Bayes

Van alle patiënten in de hematologie afdeling is 1% geïnfecteerd met het Hiv-virus, 97% daarvan testen positief. Dit is 4% van alle niet geïnfecteerde patiënten. Bereken de kans dat een patiënt die positief wordt getest effectief het Hiv-virus heeft.

C_1 = "HIV+ patiënten"
C_2 = "HIV- patiënten "
X = "test resultaat"

$P(C_1) = 0.01$ $P(X = + \mid C_1) = 0.97$
$P(C_2) = 0.99$ $P(X = + \mid C_2) = 0.04$

$$P(C_1 \mid X = +) = \frac{P(C_1) \cdot P(X = + \mid C_1)}{P(C_1) \cdot P(X = + \mid C_1) + P(C_2) \cdot P(X = + \mid C_2)}$$

Terminologie

- $P(C_1)$, $P(C_2)$: eerdere waarschijnlijkheid
- $P(C_1 \mid X = +)$, $P(C_2 \mid X = +)$: latere waarschijnlijkheid

In het algemeen

- n klassen: C1, C2, …, Cn
- k voorspellende variabelen $X_1, X_2, \ldots, X_k$

Vraag

Als $X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k$, wat is dan de meest waarschijnlijke klasse?

$P(C_i \mid X_1 = x_1, X_2 = x_2, \ldots, X_k = x_k)$?

= $P(C_i \mid x_1, x_2, \ldots, x_k)$?

Wet van Bayes:

Wet van Bayes:

$$P(C_i \mid x_1, x_2, \ldots, x_k) = \frac{P(C_i) \cdot P(x_1, x_2, \ldots, x_k \mid C_i)}{P(x_1, x_2, \ldots x_k)}$$

Voorbeeld

- Titanic dataset
- 4 variabelen, 2201 observaties
- X_1: klasse
  ('0' = personeel, '1'= duurste klasse, '2' = middle klasse, '3' = goedkoopste klasse)
- X_2: leeftijd
  ('1' = volwassen, '0' = kind)
- X_3: geslacht
  ('1' = man, '0' = vrouw)
- C: gered of niet
  ('1' = gered, '0' = niet gered)

Voorbeeld vraag: Zal een meisje in de laagste klasse gered worden of niet?

P(gered | goedkoopste, kind, vrouw)= ?

P(niet gered | goedkoopste, kind, vrouw)= ?

$$P(\text{gered} \mid \text{goedkoopste, kind, vrouw}) = \frac{P(\text{gered}) \bullet P(\text{goedkoopste, kind, vrouw} \mid \text{gered})}{P(\text{goedkoopste, kind, vrouw})}$$

$$P(\text{niet gered} \mid \text{goedkoopste, kind, vrouw}) = \frac{P(\text{niet gered}) \bullet P(\text{goedkoopste, kind, vrouw} \mid \text{niet gered})}{P(\text{goedkoopste, kind, vrouw})}$$

Opmerkingen:

- noemer P(goedkoopste, kind,vrouw) geen reden om te vergelijken.

- Om alle vragen te beantwoorden, moeten de volgende dingen gekend zijn

  P(gered) en P(niet gered)
  P(x1, x2, x3 | gered) en
  P(x1, x2, x3 | niet gered)
  Voor alle combinaties van x_1, x_2, x_3
  → 2•4•2•2 = 32 mogelijkheden
  Geschatte waarschijnlijkheden via frequenties in training set

Probleem:

- Grote hoeveelheid van mogelijkheidheden x1, x2, x3

- Bepaalde combinaties kunnen niet voorkomen in een klasse C_i in de training set

→ schatting: $P(x_1, x_2, x_3 \mid C_i) = 0$.
→ $P(C_i \mid x_1, x_2, x_3) = 0$

1. Oplossing: Naive Bayes

- Assumptie: X_1, X_2, X_3 niet onderling afhankelijk in elke klasse C_i

→ $P(x_1; x_2; x_3 \mid C_i)$
$= P(x_1 \mid C_i) \bullet P(x_2 \mid C_i) \bullet P(x_3 \mid C_i)$

- P(x1 | Ci), P(x2 | Ci), P(x3 | Ci) Schatting via frequenties in training set

- Minder snel nul; minder kansen: 2•(4+2+2) = 16