### Наївний байєсів класифікатор

Наївний баєсів класифікатор (Naive Bayes) - це сімейство алгоритмів класифікації, які діють за правилом: Кожен параметр (ознака) класифікованих даних розглядається незалежно від інших параметрів класу. Два параметри називаються незалежними, якщо значення одного параметра не впливає на інший.

Наприклад, є набір даних про пацієнта: пульс, рівень холестерину, вага, ріст і поштовий індекс. Всі параметри будуть незалежними, якщо їх значення не впливають один на одного. Для цього набору даних розумно припустити, що ріст пацієнта і поштовий індекс є незалежними, оскільки жодним чином не пов'язані між собою.

Зазвичай, параметри набору даних не є повністю незалежними. Припущення, що всі параметри набору даних є незалежними - це досить наївне припущення і зазвичай так не буває. В наведеному прикладі є 3 співвідношення, які можна вважати залежними:

* Якщо ріст збільшився, ймовірно, збільшилася вага.
* Якщо збільшилася вага, ймовірно, збільшився рівень холестерину.
* Якщо збільшився рівень холестерину, ймовірно, збільшився пульс.

Теорема Байєса дозволяє передбачити клас на підставі набору ознак, використовуючи ймовірність. Спрощене рівняння для класифікації виглядає так:

Рівняння знаходить ймовірність класу А, на підставі ознак 1 і 2. Тобто, якщо присутні ознаки 1 і 2, то, ймовірно, це дані класу А.

**Приклад реалізації наївного байєсівського класифікатора**

Існує навчальний набір даних про 1000 фруктів.

* Фрукт може бути бананом, апельсином або чимось іншим (це класи).
* Фрукт може бути довгим, солодким або жовтим (це ознаки).

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Клас** | **Довгий** | **Солодкий** | **Жовтий** | **Всього** |
| Банан | 400 | 350 | 450 | **500** |
| Апельсин | 0 | 150 | 300 | **300** |
| Інше | 100 | 150 | 50 | **200** |
| **Всього** | **500** | **650** | **800** | **1000** |

В цьому навчальному наборі даних зазначене

* Серед 500 бананів - 400 довгі, 350 солодкі і 450 жовті.
* Серед 300 апельсинів немає довгих, 150 солодких і 300 жовтих.
* Серед решти 200 фруктів - 100 довгих, 150 солодких і 50 жовтих.

Якщо надано лише дані про довжину, солодкість і колір фрукта (не знаючи його класу), то можна обчислити ймовірність того, що фрукт виявиться бананом, апельсином або чимось іншим. Припустимо, що невідомий фрукт довгий, солодкий і жовтий.

Для обчислення ймовірності потрібно виконати 4 кроки:

**Крок 1:** Щоб обчислити ймовірність того, що невідомий фрукт є бананом, слід перевірити, чи цей фрукт схожий на банан. Обчислюється ймовірність класу «Банан» на підставі ознак «довгий», «солодкий», «жовтий»:

P (Банан | Довгий, Солодкий, Жовтий)

**Крок 2:** Почнемо з чисельника і підставимо все значення в рівняння:

* P (Довгий | Банан) = 400/500 = 0.8
* P (Солодкий | Банан) = 350/500 = 0.7
* P (Жовтий | Банан) = 450/500 = 0.9
* P (Банан) = 500/1000 = 0.5

Перемноживши значення (відповідно до рівняння), отримується:

* P (Довгий | Банан) х P (Солодкий | Банан) х P (Жовтий | Банан) х P (Банан) =  
  0.8 x 0.7 x 0.9 x 0.5 = 0.252

Аналогічно виконуються обчислення для інших класів.

* P (Довгий | Апельсин) х P (Солодкий | Апельсин) х P (Жовтий | Апельсин) х P (Апельсин)=0 x 0.5 x 1 x 0.3 = 0
* P (Довгий | Інше) х P (Солодкий | Інше) х P (Жовтий | Інше) х P (Інше) =  
  0.5 x 0.75 x 0.25 x 0.2 = 0.018

**Крок 3:** Обчислюється знаменник, він буде однаковим для всіх наступних обчислень.

* Р (Довгий) = 500/1000 = 0.5
* Р (Солодкий) = 650/1000 = 0.65
* Р (Жовтий) = 800/1000 = 0.8
* Р (Довгий) х Р (Солодкий) х Р (Жовтий) = 0.5 х 0.65 х 0.8 = 0.26

**Крок 4:** Виконуються обчислення ймовірності

* P (Банан | Довгий, Солодкий, Жовтий) = 0.252 / 0.26 = 0.97
* P (Апельсин | Довгий, Солодкий, Жовтий) = 0 / 0.26 = 0
* P (Інше | Довгий, Солодкий, Жовтий) = 0.018 / 0.26 = 0.072

Оскільки 0,97 більше ніж 0,072, то наївний байєсів алгоритм класифікує довгий, солодкий і жовтий фрукт як банан.

Цей метод вимагає навчання, оскільки алгоритм використовує розмічений набір даних для побудови таблиці. Якщо частотні таблиці вже обчислені, класифікація невідомого фрукта містить лише обчислення ймовірностей для всіх класів, а потім вибір найбільшої ймовірності.