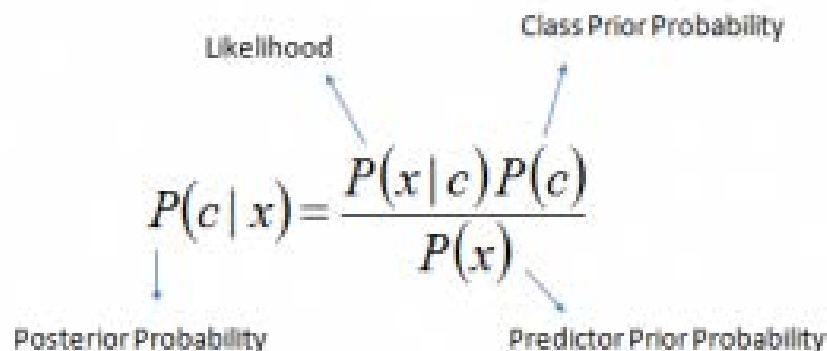


# Наивный Байесовский классификатор

# Теоретические основы

- Наивный байесовский алгоритм – это алгоритм классификации, основанный на теореме Байеса с допущением о независимости признаков.
- Теорема Байеса позволяет рассчитать апостериорную вероятность  $P(c|x)$  на основе  $P(c)$ ,  $P(x)$  и  $P(x|c)$ .

# Формулы



The diagram shows the formula for Posterior Probability,  $P(c|x) = \frac{P(x|c)P(c)}{P(x)}$ . Arrows point from descriptive labels to the components of the formula: 'Likelihood' points to  $P(x|c)$ , 'Class Prior Probability' points to  $P(c)$ , 'Posterior Probability' points to  $P(c|x)$ , and 'Predictor Prior Probability' points to  $P(x)$ .

$$P(c|x) = \frac{P(x|c)P(c)}{P(x)}$$

Likelihood

Class Prior Probability

Posterior Probability

Predictor Prior Probability

$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \dots \times P(x_n|c) \times P(c)$$

- $P(c|x)$  – апостериорная вероятность данного класса  $c$  (т.е. данного значения целевой переменной) при данном значении признака  $x$ .

# Пример

- Рассмотрим обучающий набор данных, содержащий один признак «Погодные условия» (weather) и целевую переменную «Игра» (play), которая обозначает возможность проведения матча. На основе погодных условий мы должны определить, состоится ли матч. Чтобы сделать это, необходимо выполнить следующие шаги.
- Шаг 1. Преобразуем набор данных в частотную таблицу (frequency table).

# Пример(2)

Weather	Play
Sunny	No
Overcast	Yes
Rainy	Yes
Sunny	Yes
Sunny	Yes
Overcast	Yes
Rainy	No
Rainy	No
Sunny	Yes
Rainy	Yes
Sunny	No
Overcast	Yes
Overcast	Yes
Rainy	No

Frequency Table		
Weather	No	Yes
Overcast		4
Rainy	3	2
Sunny	2	3
Grand Total	5	9

Likelihood table				
Weather	No	Yes		
Overcast		4	=4/14	0.29
Rainy	3	2	=5/14	0.36
Sunny	2	3	=5/14	0.36
All	5	9		
	=5/14	=9/14		
	0.36	0.64		

## Пример(3)

- Шаг 3. С помощью теоремы Байеса рассчитаем апостериорную вероятность для каждого класса при данных погодных условиях. Класс с наибольшей апостериорной вероятностью будет результатом прогноза.
- Задача. Состоится ли матч при солнечной погоде (sunny)?

$$P(\text{Yes} \mid \text{Sunny}) = P(\text{Sunny} \mid \text{Yes}) * P(\text{Yes}) / P(\text{Sunny})$$

# Положительные стороны

- Классификация, в том числе многоклассовая, выполняется легко и быстро.
- Когда допущение о независимости выполняется, НБА превосходит другие алгоритмы, такие как логистическая регрессия (logistic regression), и при этом требует меньший объем обучающих данных.
- НБА лучше работает с категориальными признаками, чем с непрерывными. Для непрерывных признаков предполагается нормальное распределение, что является

# Отрицательные стороны

- Если в тестовом наборе данных присутствует некоторое значение категориального признака, которое не встречалось в обучающем наборе данных, тогда модель присвоит нулевую вероятность этому значению и не сможет сделать прогноз. Это явление известно под названием «нулевая частота» (zero frequency). Данную проблему можно решить с помощью сглаживания. Одним из самых простых методов является сглаживание по Лапласу (Laplace smoothing).



# Приложения НБА

- Классификация в режиме реального времени. НБА очень быстро обучается, поэтому его можно использовать для обработки данных в режиме реального времени.
- Многоклассовая классификация. НБА обеспечивает возможность многоклассовой классификации. Это позволяет прогнозировать вероятности для множества значений целевой переменной.
- Классификация текстов, фильтрация спама, анализ тональности текста. При