

Наивный байесовский классификатор

Теория и практическое применение

Перцев Дмитрий, Никита Швачко

2025

Что такое наивный Байес?

- Вероятностный классификатор на основе формулы Байеса
- Наивное предположение: признаки независимы при заданном классе
- Упрощает вычисления — вместо многомерной плотности использует одномерные
- Эффективен, но предположение не всегда верно на практике

Формула Байеса

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}$$

где:

- $P(A|B)$ — апостериорная вероятность
- $P(B|A)$ — правдоподобие
- $P(A), P(B)$ — априорные вероятности

Формула Байеса в машинном обучении

$$P(y_k|X) = \frac{P(y_k)P(X|y_k)}{P(X)}$$

Для нескольких признаков:

$$P(y_k|X_1, X_2, \dots, X_n) = \frac{P(y_k) \prod_{i=1}^n P(X_i|y_k)}{P(X_1, X_2, \dots, X_n)}$$

- $P(y_k|X)$ — апостериорная вероятность принадлежности образца к классу y_k с учётом его признаков X
- $P(X|y_k)$ — правдоподобие, то есть вероятность признаков X при заданном классе y_k ;
- $P(y_k)$ — априорная вероятность принадлежности случайно выбранного наблюдения к классу y_k ;
- $P(X)$ — априорная вероятность признаков X .

Правило классификации

На практике используют упрощенную формулу:

$$y_k \propto \arg \max_{y_k} P(y_k) \prod_{i=1}^n P(X_i|y_k)$$

- Знаменатель опускается (не зависит от класса)
- Выбирается класс с максимальной апостериорной вероятностью
- Параметры оцениваются методом максимального правдоподобия

Разновидности наивного Байеса

- **GaussianNB** — для непрерывных признаков с нормальным распределением
- **MultinomialNB** — для дискретных признаков (частоты слов)
- **ComplementNB** — для несбалансированных данных
- **BernoulliNB** — для бинарных признаков
- **CategoricalNB** — для категориальных данных

GaussianNB

Для непрерывных признаков с нормальным распределением

Формула вероятности

$$P(x_i|y) = \frac{1}{\sqrt{2\pi\sigma_y^2}} \exp\left(-\frac{(x_i - \mu_y)^2}{2\sigma_y^2}\right)$$

- μ_y — среднее признака в классе y
- σ_y — стандартное отклонение в классе y
- Параметры оцениваются по обучающим данным

Применение

- Медицинская диагностика
- Финансовый анализ
- Анализ сенсорных данных

MultinomialNB

Для дискретных признаков (мультиномиальное распределение)

Формула вероятности

$$P(x_i|y) = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

- N_{yi} — количество раз, когда признак i встречается в классе y
- N_y — общее количество всех признаков в классе y
- n — количество различных признаков
- α — сглаживающий параметр (обычно $\alpha = 1$)

Применение

- Классификация текстов
- Анализ тональности
- Категоризация документов

BernoulliNB

Для бинарных признаков (распределение Бернулли)

Формула вероятности

$$P(x_i|y) = P(x_i = 1|y)x_i + (1 - P(x_i = 1|y))(1 - x_i)$$

- $P(x_i = 1|y)$ — вероятность того, что признак i равен 1 в классе y
- x_i — значение признака (0 или 1)
- Подходит для признаков наличия/отсутствия

Применение

- Спам-фильтрация (наличие слов)
- Медицинская диагностика (наличие симптомов)
- Анализ кликов (click-through rate)

CategoricalNB

Для категориальных признаков

Формула вероятности

$$P(x_i = t \mid y = c; \alpha) = \frac{N_{tic} + \alpha}{N_c + \alpha n_i}$$

- N_{tic} — количество раз, когда признак x_i принимает значение t в классе c
- N_c — общее количество признаков в классе c
- n_i — количество доступных значений признака i
- α — сглаживающий параметр

Применение

- Демографические данные
- Опросы пользователей
- Маркетинговые исследования

ComplementNB

Для несбалансированных наборов данных

Формула весов

$$\hat{\theta}_{ci} = \frac{\alpha_i + \sum_{j:y_j \neq c} d_{ij}}{\alpha + \sum_{j:y_j \neq c} \sum_k d_{kj}}$$

$$w_{ci} = \log \hat{\theta}_{ci}, \quad w_{ci} = \frac{w_{ci}}{\sum_j |w_{ci}|}$$

- Оценивает вероятность признака при **дополнении** класса
- Учитывает отсутствие признаков в других классах
- Менее чувствителен к смещению выборки

Применение

- Обнаружение мошенничества
- Медицинская диагностика редких заболеваний

Сравнение типов наивного Байеса

Тип	Тип признаков	Основное применение
GaussianNB	Непрерывные, нормальное распределение	Научные данные, измерения
MultinomialNB	Дискретные, частоты	Текстовая классификация
BernoulliNB	Бинарные	Наличие/отсутствие признаков
CategoricalNB	Категориальные	Демографические данные
ComplementNB	Дискретные, несбалансированные	Редкие события, аномалии

Принцип работы GaussianNB

- 1 Расчет априорных вероятностей классов
- 2 Расчет средних и стандартных отклонений признаков по классам
- 3 Расчет вероятностной плотности тестовых признаков
- 4 Вычисление апостериорных вероятностей
- 5 Выбор класса с максимальной вероятностью

Пример: фильтрация спама

Подробный расчет с использованием MultinomialNB

Обучающая выборка

Message	Class
Hi, how are you?	Not spam
Congratulations, you won a prize!	Spam
Buy the product now and get a discount!	Spam
Let's walk this evening	Not spam

Классифицируемое сообщение

"Hi, you won a discount and you can get the prize this evening"

Сглаживание Лапласа

Добавляем $+1$ к каждой частоте для избежания нулевых вероятностей

Расчет частот и вероятностей

Word	Frequency		Smoothed Frequency		Probability	
	Not Spam	Spam	Not Spam	Spam	Not Spam	Spam
hi	1	0	2	1	$2/28 = 0.0714$	$1/33 = 0.03$
how	1	0	2	1	0.0714	0.03
are	1	0	2	1	0.0714	0.03
you	1	1	2	2	0.0714	0.06
congratulations	0	1	1	2	0.0357	0.06
won	0	1	1	2	0.0357	0.06
a	0	2	1	3	0.0357	0.09
prize	0	1	1	2	0.0357	0.06
buy	0	1	1	2	0.0357	0.06
the	0	1	1	2	0.0357	0.06
product	0	1	1	2	0.0357	0.06
now	0	1	1	2	0.0357	0.06
and	0	1	1	2	0.0357	0.06
get	0	1	1	2	0.0357	0.06
discount	0	1	1	2	0.0357	0.06
let's	1	0	2	1	0.0714	0.03
walk	1	0	2	1	0.0714	0.03
this	1	0	2	1	0.0714	0.03
evening	1	0	2	1	0.0714	0.03

Формулы расчета

Математическое обоснование

Основная формула MultinomialNB

$$P(x_i|y) = \frac{N_{yi} + \alpha}{N_y + \alpha n}$$

Априорные вероятности классов

$$P(Spam) = \frac{2}{4} = 0.5, \quad P(Not\ Spam) = \frac{2}{4} = 0.5$$

Формула классификации для сообщения

$$P(C|M) = P(C) \cdot \prod_{i=1}^n P(w_i|C), \quad w_i \in M$$

Формулы расчета

Для нашего сообщения

$$\begin{aligned} P(C|M) = & P(C) \cdot P('hi'|C) \cdot P('you'|C) \cdot P('won'|C) \cdot P('a'|C) \cdot P('discount'|C) \\ & \cdot P('and'|C) \cdot P('you'|C) \cdot P('can'|C) \cdot P('get'|C) \cdot P('the'|C) \cdot P('prize'|C) \\ & \cdot P('this'|C) \cdot P('evening'|C) \end{aligned}$$

Подробный расчет для Spam

$$\begin{aligned} P(\text{Spam}|M) = & 0.5 \times P('hi'|\text{Spam}) \times P('you'|\text{Spam}) \times P('won'|\text{Spam}) \times P('a'|\text{Spam}) \\ & \times P('discount'|\text{Spam}) \times P('and'|\text{Spam}) \times P('you'|\text{Spam}) \times P('can'|\text{Spam}) \\ & \times P('get'|\text{Spam}) \times P('the'|\text{Spam}) \times P('prize'|\text{Spam}) \times P('this'|\text{Spam}) \\ & \times P('evening'|\text{Spam}) \end{aligned}$$

Численный расчет

$$\begin{aligned} P(\text{Spam}|M) = & 0.5 \times 0.03 \times 0.06 \times 0.06 \times 0.09 \times 0.06 \times 0.06 \times 0.06 \times 0.03 \\ & \times 0.06 \times 0.06 \times 0.06 \times 0.03 \times 0.03 \\ = & 0.5 \times (0.03^4) \times (0.06^9) \times 0.09 \\ = & 0.5 \times 0.00000081 \times 0.000000010077696 \times 0.09 \\ \approx & 6.12 \times 10^{-18} \end{aligned}$$

Подробный расчет для Not Spam

$$\begin{aligned} P(\text{Not Spam}|M) = & 0.5 \times P(\text{'hi'}|\text{Not Spam}) \times P(\text{'you'}|\text{Not Spam}) \times P(\text{'won'}|\text{Not Spam}) \\ & \times P(\text{'a'}|\text{Not Spam}) \times P(\text{'discount'}|\text{Not Spam}) \times P(\text{'and'}|\text{Not Spam}) \\ & \times P(\text{'you'}|\text{Not Spam}) \times P(\text{'can'}|\text{Not Spam}) \times P(\text{'get'}|\text{Not Spam}) \\ & \times P(\text{'the'}|\text{Not Spam}) \times P(\text{'prize'}|\text{Not Spam}) \times P(\text{'this'}|\text{Not Spam}) \\ & \times P(\text{'evening'}|\text{Not Spam}) \end{aligned}$$

$$\begin{aligned} P(\text{Not Spam}|M) = & 0.5 \times 0.0714 \times 0.0714 \times 0.0357 \times 0.0357 \times 0.0357 \times 0.0357 \\ & \times 0.0714 \times 0.0357 \times 0.0357 \times 0.0357 \times 0.0357 \times 0.0357 \\ & \times 0.0357 \times 0.0714 \end{aligned}$$

Численный расчет

$$\begin{aligned} P(\text{Not Spam}|M) &= 0.5 \times (0.0714^4) \times (0.0357^{10}) \times 0.0714 \\ &= 0.5 \times 0.000026 \times 3.3 \times 10^{-15} \times 0.0714 \\ &\approx 2.45 \times 10^{-18} \end{aligned}$$

Результат классификации

Сравнение вероятностей

Итоговые вероятности

- $P(\text{Spam}|M) \approx 6.12 \times 10^{-18}$
- $P(\text{Not Spam}|M) \approx 2.45 \times 10^{-18}$

$$\hat{y} = \arg \max_{y \in \{\text{Spam}, \text{Not Spam}\}} P(y|M)$$

Результат

Поскольку $P(\text{Spam}|M) > P(\text{Not Spam}|M)$, сообщение классифицируется как **SPAM**

В реальных системах используют логарифмы вероятностей чтобы избежать численных проблем с очень малыми числами:

$$\log P(C|M) = \log P(C) + \sum_{i=1}^n \log P(w_i|C)$$

Преимущества и недостатки

Преимущества:

- Простота реализации и интерпретации
- Высокая скорость работы
- Хорошо работает с небольшими наборами данных
- Устойчив к шуму в данных
- Разные варианты для разных типов данных

Недостатки:

- Наивное предположение о независимости признаков
- Может уступать в точности более сложным алгоритмам
- Проблема нулевой вероятности (решается сглаживанием)
- Чувствительность к выбросам (для GaussianNB)

Когда использовать наивный Байес?

Рекомендуется использовать когда:

- Мало обучающих данных
- Требуется быстрая классификация
- Признаки условно независимы
- Нужен базовый алгоритм для сравнения
- Простота реализации важнее максимальной точности

Не рекомендуется когда:

- Признаки сильно коррелированы
- Требуется высокая точность
- Достаточно данных для более сложных моделей

Заключение

- Наивный Байес — мощный и эффективный классификатор
- Разные варианты для разных типов данных и задач
- Широко применяется в текстовой классификации и не только
- Прост в понимании и реализации
- Отличная отправная точка для решения задач классификации
- Несмотря на "наивность" часто показывает хорошие результаты