Что такое классификатор?

Классификатор — это модель машинного обучения, которая используется для различения различных объектов на основе определенных признаков.

Принцип наивного байесовского классификатора:

Наивный байесовский классификатор — это вероятностная модель машинного обучения, которая используется для задачи классификации. Суть классификатора основана на теореме Байеса.

Теорема Байеса:

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

Используя теорему Байеса, мы можем найти вероятность того, что произойдет A, если произошло B. Здесь B — свидетельство, а A — гипотеза. Здесь делается предположение, что предикторы/функции независимы. То есть наличие одного конкретного признака не влияет на другой. Поэтому его называют наивным.

Пример:

Давайте возьмем пример, чтобы получить некоторую лучшую интуицию. Рассмотрим проблему игры в гольф. Набор данных представлен ниже.

	OUTLOOK	TEMPERATURE	HUMIDITY	WINDY	PLAY GOLF
0	Rainy	Hot	High	False	No
1	Rainy	Hot	High	True	No
2	Overcast	Hot	High	False	Yes
3	Sunny	Mild	High	False	Yes
4	Sunny	Cool	Normal	False	Yes
5	Sunny	Cool	Normal	True	No
6	Overcast	Cool	Normal	True	Yes
7	Rainy	Mild	High	False	No
8	Rainy	Cool	Normal	False	Yes

Мы классифицируем, подходит ли день для игры в гольф, учитывая особенности дня. Столбцы представляют эти функции, а строки представляют отдельные записи. Если мы возьмем первую строку набора данных, мы увидим, что для игры в гольф не подходит, если прогноз дождливый, температура жаркая, влажность высокая и не ветрено. Здесь мы делаем два предположения, одно из которых, как указано выше, мы считаем, что эти предикторы независимы. То есть, если температура жаркая, это не обязательно означает, что влажность высокая. Еще одно предположение, сделанное здесь, состоит в том, что все предикторы одинаково влияют на результат. То есть ветреный день не имеет большего значения при принятии решения играть в гольф или нет.

Согласно этому примеру, теорему Байеса можно переписать так:

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

Переменная y является переменной класса (играть в гольф), которая представляет, подходит ли игра в гольф или нет с учетом условий. Переменная X представляет параметры/функции.

X дается как,

$$X = (x_1, x_2, x_3, ..., x_n)$$

Здесь x_1,x_2....x_n представляют функции, т.е. они могут быть сопоставлены с внешним видом, температурой, влажностью и ветром. Подставив вместо X и расширив по цепному правилу, мы получим:

$$P(y|x_1, ..., x_n) = \frac{P(x_1|y)P(x_2|y)...P(x_n|y)P(y)}{P(x_1)P(x_2)...P(x_n)}$$

Теперь вы можете получить значения для каждого, просмотрев набор данных и подставив их в уравнение. Для всех записей в наборе данных знаменатель не меняется, он остается статичным. Следовательно, знаменатель можно убрать и ввести пропорциональность.

$$P(y|x_1, ..., x_n) \propto P(y) \prod_{i=1}^{n} P(x_i|y)$$

В нашем случае переменная класса (y) имеет только два результата: да или нет. Могут быть случаи, когда классификация может быть многовариантной. Следовательно, нам нужно найти класс y с максимальной вероятностью.

$$y = argmax_y P(y) \prod_{i=1}^{n} P(x_i|y)$$

Используя приведенную выше функцию, мы можем получить класс, учитывая предикторы.

Типы наивного байесовского классификатора:

Полиномиальный наивный байесовский метод:

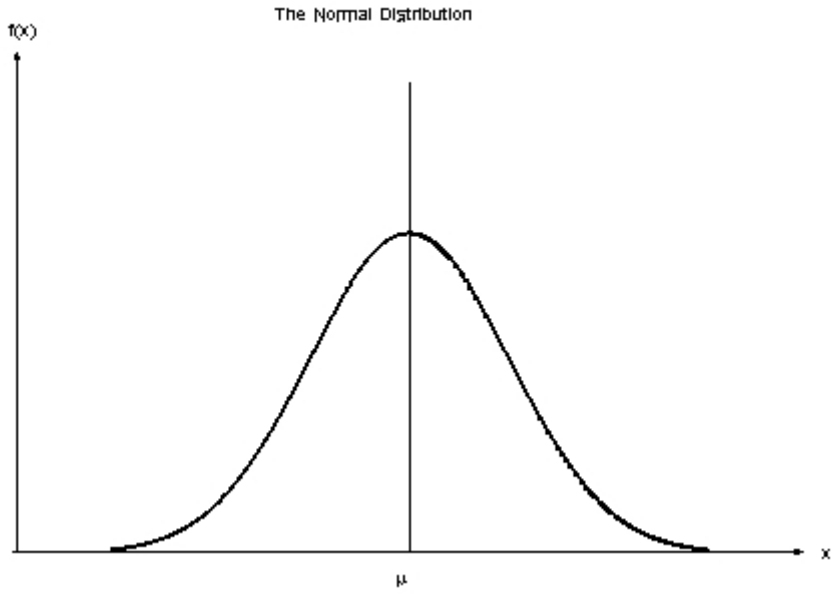
Это в основном используется для проблемы классификации документов, т. Е. Относится ли документ к категории спорта, политики, технологий и т. Д. Функции / предикторы, используемые классификатором, представляют собой частоту слов, присутствующих в документе.

Наивный Байес Бернулли:

Это похоже на полиномиальный наивный байес, но предикторы являются булевыми переменными. Параметры, которые мы используем для предсказания переменной класса, принимают только значения да или нет, например, встречается ли слово в тексте или нет.

Гауссовский наивный байесовский метод:

Когда предикторы принимают непрерывное значение и не являются дискретными, мы предполагаем, что эти значения взяты из гауссовского распределения.



Поскольку способ представления значений в наборе данных меняется, формула условной вероятности изменяется на

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

Вывод:

Алгоритмы наивного Байеса в основном используются в анализе настроений, фильтрации спама, системах рекомендаций и т. Д. Они быстры и просты в реализации, но их самый большой недостаток заключается в том, что предикторы должны быть независимыми. В большинстве реальных случаев предикторы зависимы, что снижает производительность классификатора.