Наивный байесовский классификатор основывается на применении формулы Байеса, что контексте машинного обучения формулируется следующим образом:

$$P(C_k|x) = \frac{P(x|C_k) \cdot P(C_k)}{P(x)}$$

- C_k - номер класса.
- x - экземпляр данных (вектор признаков классифицируемого объекта).
- $P(C_k|x)$ - вероятность принадлежности экземпляра x классу C_k (постериорная вероятность). По сути является ответом байессовского классификатора.
- $P(x|C_k)$ - вероятность наблюдения экземпляра x в предположении, что он пренадлежит классу C_k. Также называется значением правдоподобия (вычисляется с помощью функции правдоподобия).
- $P(C_k)$ - априорная вероятность появления класса C_k. Считается как частота появления класса C_k в датасете.
- $P(x)$ - безусловная вероятность появления экземпляра x вне зависимости от его класса. Может быть вычислена по следующей формуле:

$$P(x) = \sum_k P(x|C_k) \cdot P(C_k)$$

Во время классификации экземпляра x вы вычисляете вероятность $P(C_k|x)$ для всех классов, а после выбираете класс с наибольшей вероятностью.

Особенность наивного байесовского классификатора заключается в его "наивности": предполагается, что признаки независимы друг от друга. Благодаря этому предположению очень сильно упрощается вычисление значения правдоподобия:

$$P(x|C_k) = P(x_1|C_k) \cdot P(x_2|C_k) \cdot \ldots \cdot P(x_m|C_k)$$

Теперь вместо сложной многомерной функции правдоподобия $P(x|C_k)$, что моделирует взаимоотношения между всеми признаками x_i (и непонятно как вычисляется), у нас имеется произведение одномерных функций правдоподобия. Ниже будут приведены конкретные функции правдоподобия, что нужно использовать в данной лабораторной работе.

Классификатор "Гаусса"

Функция правдоподобия для классификатора Гаусса:

$$P(x_i|\mu_{i,C_k}, \sigma^2_{i,C_k}, C_k) = \frac{1}{\sqrt{2\pi\sigma^2_{i,C_k}}} \cdot \exp\left(-\frac{(x_i - \mu_{i,C_k})^2}{2\sigma^2_{i,C_k}}\right)$$

- x_i - значение i-го признака.
- μ_{i,C_k} - среднее значение i-го признака для класса C_k.
- σ^2_{i,C_k} - дисперсия i-го признака для класса C_k.

Мультиномиальный классификатор

Мультиномиальный классификатор используется для работы с целочисленными данными (когда признаки объектов представлены неотрицательными числами). Пример задачи: классификация документов. В рамках данной задачи в качестве признака x_i выступает количество вхождений i-го слова (есть заранее заданный список слов) в документ x.

> Пример документа x:
>
> "Мне нравится кофе. Кофе бодрит, а бодрости мне не достаёт никогда."
>
> Список слов (словарь):
>
> 1. мне
> 2. нравится
> 3. кофе
> 4. чай
>
> Пример соответствующего вектора признаков x:
>
> { 2, 1, 2, 0 }.

В терминах мультиномиального распределения слова называются атрибутами (неофициальная терминология) объекта x, а признак x_i представляет количество появлений i-го атрибута в объекте x.

Функция правдоподобия для мультиномиального классификатора:

$$P(x|C_k) = \frac{(\sum_{i=1}^{n} x_i)!}{\prod_{i=1}^{n} x_i!} \prod_{i=1}^{n} p_{ik}^{x_i}$$

- p_{ik} - вероятность появления i-го атрибута в объекте x в предположении, что он принадлежит классу C_k.

Рекомендуется сначала посчитать логарифм от этой функции для избежания численного переполнения, а после применить экспоненту к посчитанному значению.

В случае мультиномиального классификатора некоторые из параметров p_{ik} могут оказаться равными нулю. Это может привести к тому, что $P(x|C_k)$ будет всегда равняться нулю для некоторых классов. Для избежания данной ситуация необходимо "сглаживать" значения p_{ik} посредством внесения корректировок в процесс вычисления p_{ik}. Детали сглаживания можно найти по этой ссылке. (в качестве значения параметра α берите единицу)

Доп. ссылки:
Naive Bayes classifier - Wikipedia
Bayes' theorem - Wikipedia