Написание простейшего классификатора

Иногда перед нами стоит следующая задача: на основании имеющегося набора признаков сделать какое-либо предположение относительно класса, к которому относится рассматриваемый объект. Например, на основании ряда характеристических оценок опухоли отнести ее к классу доброкачественных или злокачественных.

В данной задаче мы будем пользоваться набором данных об опухолях груди, который был предоставлен доктором Вильямом Уолбергом (Dr. William H. Wolberg). Им были собраны данные о 699 доброкачественных и злокачественных опухолях. На основании биопсии каждая из них была оценена по 9 признакам:

1. Толщина скопления клеток
2. Оценка однородности размеров клеток
3. Оценка однородности форм клеток
4. Межклеточная мембранная адгезия
5. оценка размера отдельной клетки эпителия
6. Единичные "голые" ядра
7. Рыхлый (деконденсированный) хроматин
8. Нормальные ядрышки
9. Динамика митоза

Каждый признак имеет оценку от 1 до 10. После этого дается классификация опухоли – злокачественная или доброкачественная.

Используемый набор данных включает в себя 699 строк, каждая из которых состоит из 11 полей:

Код пациента – 9 признаков – класс опухоли

Класс опухоли 4 – злокачественная

Класс опухоли 2 - доброкачественная

Нам нужно на основании уже имеющихся данных об опухолях отнести новый образец к одному из двух классов. Для этого необходимо написать классификатор и обучить его. Мы будем использовать самый простой вариант.

Необходим какой-то способ разбиения опухолей на 2 класса: доброкачественная и злокачественная. Мы поступим следующим образом:

- По каждому из 9 параметров найдем среднее значение в каждом из классов
- Найдем среднее между средними

Полученные 9 значений и будут границами разделения на классы. См. рисунок 1.

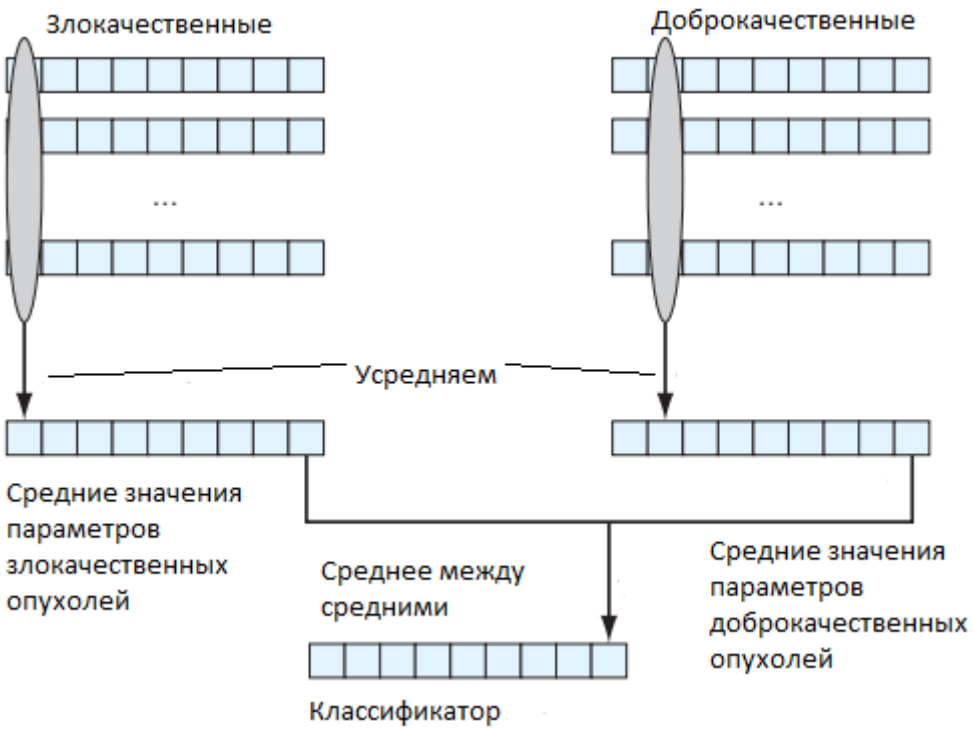

Рисунок 1 – принцип построения классификатора

Чтобы классифицировать неизвестную опухоль, нужно сравнить полученные граничные значения с параметрами опухолями и сделать простое голосование. Если параметр опухоли больше, чем граничное значение, то по этому параметру опухоль классифицируется как злокачественная, если меньше или равен, то, соответственно, как доброкачественная. Такой шаг выполняется по каждому из параметров. Например, если классификатор выглядел так:

1 2 3 4 5 6 7 8 9

Поступившие параметры опухоли такие 1 1 1 3 4 5 8 4 6, то

Параметр 1 доброкачественная (1 = 1)

Параметр 2 доброкачественная (1<2)

Параметр 3 доброкачественная (1<3)

Параметр 4 доброкачественная (3<4)

Параметр 5 доброкачественная (4<5)

Параметр 6 доброкачественная (5<6)

Параметр 7 злокачественная (8>7)

Параметр 8 доброкачественная (4<8)

Параметр 9 доброкачественная (6<9)

Итого 8 голосов за доброкачественную, 1 за злокачественную, поэтому опухоль надо классифицировать как доброкачественную.