УДК 51

МАТЕМАТИЧЕСКИЕ МЕТОДЫ, ИСПОЛЬЗУЕМЫЕ В МАШИННОМ ОБУЧЕНИИ

Кивишев К.А., студент гр. ПИб-192, II курс
Балахнин Е.Е., студент гр. ПИб-192, II курс
Гутова Е.В., ст. преподаватель кафедры математики
Кузбасский государственный технический университет
имени Т.Ф. Горбачева
г. Кемерово

На данный момент существуют различные технологии распознавания того, что происходит на изображении. Телефоны активируются только при распознавании лица пользователя или отпечатка пальца, камеры могут регистрировать скорость автомобиля, существуют бесконтактные магазины, в которых стоят камеры и они запоминают что взял покупатель и после того как покупатель выйдет из магазина, с него снимется сумма денег за взятые им с собой товары. Все эти возможности основаны на распознавании изображения, которое реализуется с помощью машинного обучения.

Машинное обучение основывается на том, что компьютеру выдается набор данных и с помощью этого набора данных происходит обучение. Обычно исследование набора данных начинается со следующего алгоритма: сбор и очистка данных; визуальных анализ данных, их распределение, статистики; анализ зависимости (корреляции) между переменными (признаками); отбор и определение признаков, которые будут использоваться для построения моделей; разделение на данные для обучения модели и тестовые; построение моделей на данных для обучения, оценка результата на тестовых данных; интерпретация полученной модели, визуализация результатов. Основные принципы подготовки данных в которых используется различные специализированные математические методы для достижения цели: классификация; регрессия; кластеризация.

Классификация — это отнесение объектов к определенному классу по набору признаков. Существуют несколько методов классификации данных: формула Байеса; дерево принятия решений. Метод классификации данных по формуле Байеса состоит в том, что высчитывается вероятность по всем гипотезам при выполнении события и после чего она сравнивается, чья вероятность окажется больше, такую гипотезу компьютер посчитает верной и выведет её пользователю.

Пример: Дано 100 фруктов, из которых 50 – бананы, 30 – апельсины, 20 – остальные. Все данные приведены в таблице ниже. Надо определить фрукт, который является длинным, сладким и желтым.

Фрукты	Длинный	Сладкий	Желтый	Всего фруктов
Банан	40	35	45	50
Апельсин	0	15	30	30
Другое	10	15	5	20

H_1 – фрукт является бананом, $P(H_1) = \frac{50}{100}$

H_2 – фрукт является апельсином, $P(H_2) = \frac{30}{100}$

H_3 – фрукт относится к другим, $P(H_3) = \frac{20}{100}$

Высчитывается вероятность при каждой гипотезе.

$$P_{H_1}(A) = P(A_1 \cdot A_2 \cdot A_3) = \frac{40}{50} \cdot \frac{35}{50} \cdot \frac{45}{50} = 0{,}504$$

$$P_{H_2}(A) = P(A_1 \cdot A_2 \cdot A_3) = \frac{0}{30} \cdot \frac{15}{30} \cdot \frac{30}{30} = 0$$

$$P_{H_3}(A) = P(A_1 \cdot A_2 \cdot A_3) = \frac{10}{20} \cdot \frac{15}{20} \cdot \frac{5}{20} = 0{,}09375$$

Далее высчитывается общая вероятность и после чего, используя все данные рассчитываются вероятности по формуле Байеса.

$$P(A) = P(H_1) \cdot P_{H_1}(A) + P(H_2) \cdot P_{H_2}(A) + P(H_3) \cdot P_{H_3}(A) =$$

$$= \frac{50 \cdot 0{,}504}{100} + \frac{30 \cdot 0}{100} + \frac{20 * 0{,}09375}{100} = 0{,}27075$$

$$P_A(H_1) = \frac{P(H_1) \cdot P_{H_1}(A)}{P(A)} = \frac{\frac{50}{100} \cdot 0{,}504}{0{,}27075} \approx 0{,}931$$

$$P_A(H_2) = \frac{P(H_2) \cdot P_{H_2}(A)}{P(A)} = \frac{\frac{30}{100} \cdot 0}{0{,}27075} = 0$$

$$P_A(H_3) = \frac{P(H_3) \cdot P_{H_3}(A)}{P(A)} = \frac{\frac{20}{100} \cdot 0{,}09375}{0{,}27075} \approx 0{,}014$$

Сравнивая полученные значения по формуле Байеса заметно, что значение $P_A(H_1)$ намного больше, в отличие от остальных и поэтому компьютер посчитает, что фрукт является бананом.

Метод дерева принятия решений состоит в том, что существует дерево, которое состоит из «веток» и «листьев». Ветки содержат в себе атрибуты, от которых зависит целевая функция, а листья содержат в себе целевую функцию. Процесс решения происходит до того момента, пока компьютер не дойдет до конца ветки. Компьютер не может понять просто так, что он дошёл до конца дерева, для этого он рассчитывает Энтропию ветви, которая

вычисляется по формуле: $S = -\sum_{i=1}^{N} p_i \log_2 p_i$, где p_i – вероятность нахождения системы в i – ом состоянии, N – количество возможных состояний

Пример: Дано 10 шариков, 7 зеленых и 3 красных и расставлены они в порядке, показанном на рисунке 1.

Рисунок 1. Расстановка шаров

Далее рассчитывается Энтропия

$$S_0 = -\frac{7}{10}\log_2\frac{7}{10} - \frac{3}{10}\log_2\frac{3}{10} = 0{,}8816$$

После чего шары разделяются на 2 части не меняя последовательности (рис. 2)

Рисунок 2. Разделение шаров на 2 группы

Далее рассчитывается Энтропия для каждой группы

$$S_1 = -\frac{3}{5}\log_2\frac{3}{5} - \frac{2}{5}\log_2\frac{2}{5} = 0{,}971$$

$$S_2 = -\frac{4}{5}\log_2\frac{4}{5} - \frac{1}{5}\log_2\frac{1}{5} = 0{,}722$$

Каждая группа шаров так же делится на 2 группы (рис. 3) и в итоге видно, что в каждой группе находится один или несколько шаров одного цвета.

Рисунок 3. Разделение шаров на группы

Высчитывается Энтропия каждой группы

$$S_3 = -\frac{3}{3}\log_2 3 = 0$$

$$S_4 = -\frac{2}{2}\log_2\frac{2}{2} = 0$$

$$S_5 = -\frac{4}{4}\log_2\frac{4}{4} = 0$$

$$S_6 = -\frac{1}{1}\log_2\frac{1}{1} = 0$$

Видно, что значение каждой Энтропии равняется нулевому, что сигнализирует об остановке процесса. Дерево решений показано на рисунке 4.

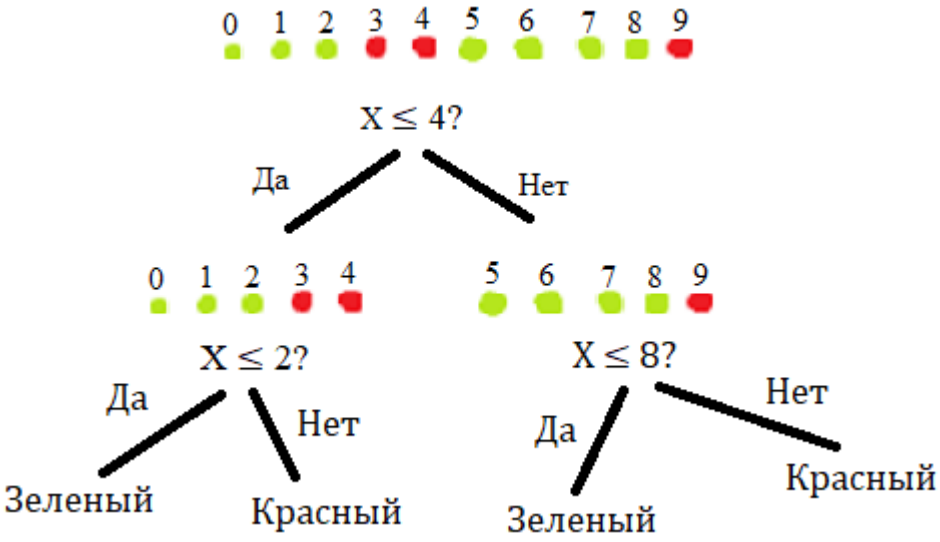

Рисунок 4. Дерево решений

Для компьютера такая модель выглядит следующим образом: ему дан набор порядковых номеров. Компьютер принимает на вход какой-то порядковый номер (x), рассчитывает Энтропию и после чего проверяет значение порядкового номера. Если значение меньше или равно 4, то компьютер идет по левой ветке, иначе по правой. Рассчитывает Энтропию и идет дальше проверять условие и при истинном условии он идет по истинной ветке до того момента, пока Энтропия не будет равна 0.

Подведём итоги, машинное обучение основано на анализе данных, которые относятся к определенному классу и компьютеру приходится находить признаки, которые определяют данные к классу. Данная процедура определения данных к какому-либо классу называется Классификацией и выполняется с помощью формулы Байеса, дерева принятия решений и многих других.