УДК 519.6

КЛАССИФИКАЦИЯ ДАННЫХ КАК ЗАДАЧА МАШИННОГО ОБУЧЕНИЯ

Фрузорова В. А., Попок Р.В., Метельский А. В.
Белорусский национальный технический университет, Минск

Машинное обучение — класс методов искусственного интеллекта, одной из задач которых является не прямое выполнение задачи, а обучение решению множества сходных задач. В основе этих методов лежат такие разделы математики, как комбинаторика, дискретная математика, математический анализ, линейная алгебра, статистика, теория вероятности.

Другая цель машинного обучения — предсказать результат по входным данным. Чем более разнообразными они будут, тем сложнее программе выявить закономерности и дать точный результат. Примером может служить распознавание букв в тексте (однообразные данные) и распознавание изображения на фотографии (разнообразные данные).

Еще одной важной задачей машинного обучения является классификация объектов. Для классификации всегда нужен Учитель — заранее размеченные данные с признаками и категориями, по которым программа будет учиться разделять объекты на классы. Такой подход позволяет создавать различные классификаторы для различных целей. Например, новостные ленты в социальных сетях, предлагающие новости исходя из интересов пользователей, или классификация текстов по тематике и языкам, что необходимо для поисковых систем.

Не менее востребованная задача, которую решают с помощью машинного обучения – выбор наилучшего варианта из числа доступных. По такому принципу работают автомобили с автопилотом: программа выбирает наилучший вариант развития событий.

Раньше все спам-фильтры использовали наивный байесовский классификатор, основанный на теореме Байеса:

$$P(H_i / A) = \frac{P(H_i) \cdot P(A / H_i)}{P(A)}, \; i = \overline{1, n},$$

Рисунок 1. Формула Байеса

$\boldsymbol{P(H_i / A)}$ - вероятность события $\boldsymbol{H_i}$, если наступит событие $\boldsymbol{A}$

$\boldsymbol{P(H_i)}$ – априорная вероятность события$\boldsymbol{H_i}$

$\boldsymbol{P(A / H_i)}$ – вероятность события $\boldsymbol{A}$ при истинности $\boldsymbol{H_i}$,

$\boldsymbol{P(A)}$ – полная вероятность события $\boldsymbol{A}$.

В спам-фильтрах определяют вероятность того, что письмо спам, при условии, что в нем встречаются определенные слова.

Популярным методом классификации в машинном обучении является метод опорных векторов. Он использует гиперплоскость, чтобы классифицировать данные на 2 класса. Для простой задачи с 2 параметрами гиперплоскость может быть линией, однако, если параметров больше – уже нет.

Метод опорных векторов позволяет спроецировать ваши данные в пространство большей размерности, используя для этого новые признаки объектов, а когда данные спроецированы, метод определяет гиперплоскость, которая делит данные на 2 класса наилучшим образом.

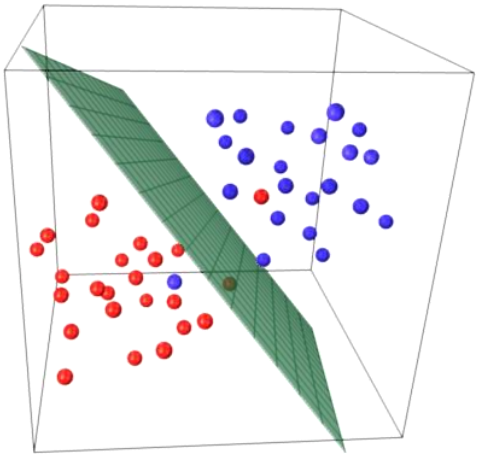

Рисунок 2. Визуализация разделения данных на два класса.

Суть метода в том, что он максимизирует отступ так, чтобы гиперплоскость находилась примерно на одинаковом расстоянии от 2 типов данных – это снижает шанс ошибок классификации.

У данного метода есть хорошее качество – поиск аномалий. Когда какой-либо признак объекта сильно отличается от среднего значения, программа выделяет этот участок как аномальный. Данная особенность нашла применение в медицине (компьютер подсвечивает врачу все подозрительные области МРТ или выделяет отклонения в анализах) и не только.

Литература

1. (2010) Классификация данных методом опорных векторов//Сайт https://habr.com. 29 сентября. (https://habr.com/ru/post/105220/) Просмотрено: 15.04.2019.
2. (2016) Метод опорных векторов (SVM)//Сайт http://datascientist.one. 15мая. (http://datascientist.one/support-vector-machines/) Просмотрено: 15.04.2019.
3. Udiprod (2007) SVMwithpolynomialkernelvisualizationвекторов//YouTube. 5 февраля. (https://www.youtube.com/watch?v=3liCbRZPrZA) Просмотрено: 16.04.2019.

4. (2018) Машинное обучение для людей//Сайт https://vas3k.ru. 22июля. (https://vas3k.ru/blog/machine_learning/) Просмотрено: 14.04.2019.
5. (2015)6 простых шагов для освоения наивного байесовского алгоритма//Сайт http://datareview.info. 23 сентября. (http://datareview.info/article/6-prostyih-shagov-dlya-osvoeniya-naivnogo-bayesovskogo-algoritma-s-primerom-koda-na-python/) Просмотрено: 15.04.2019.
6. Veritasium(2017) The Bayesian Trap//YouTube. 5апреля. (https://www.youtube.com/watch?v=R13BD8qKeTg) Просмотрено: 16.04.2019.
7. (2017) Теорема Байеса: из-за чего весь сыр-бор?//Сайт https://habr.com. 16июня. (https://habr.com/ru/post/404633/) Просмотрено: 16.04.2019.