УДК 004.85

КОНТРОЛИРУЕМОЕ ОБУЧЕНИЕ

Ерошевич К.В., ассистент кафедры ИиАПС
Кузбасский государственный технический университет
имени Т.Ф. Горбачева
г. Кемерово

Контролируемое обучение, также известное как контролируемое машинное обучение, является подкатегорией машинного обучения и искусственного интеллекта. Это определяется использованием помеченных наборов данных для обучения алгоритмов, которые позволяют точно классифицировать данные или прогнозировать результаты. По мере ввода входных данных в модель, она корректирует свои веса до тех пор, пока модель не будет подобрана соответствующим образом, что происходит в рамках процесса перекрестной проверки. Контролируемое обучение помогает организациям решать различные масштабные проблемы реального мира, такие как классификация спама в отдельной папке из вашего почтового ящика.

Контролируемое обучение использует обучающий набор для обучения моделей, чтобы получить желаемый результат. Этот обучающий набор данных включает входные данные и правильные выходные данные, которые позволяют модели обучаться с течением времени. Алгоритм измеряет свою точность с помощью функции потерь, корректируя ее до тех пор, пока ошибка не будет достаточно сведена к минимуму.

Контролируемое обучение можно разделить на два типа проблем при интеллектуальном анализе данных — классификацию и регрессию:

- Классификация использует алгоритм для точного распределения тестовых данных по определенным категориям. Он распознает конкретные объекты в наборе данных и пытается сделать некоторые выводы о том, как эти объекты должны быть помечены или определены. Распространенными алгоритмами классификации являются линейные классификаторы, машины опорных векторов

(SVM), деревья решений, k-ближайший сосед и случайный лес, которые более подробно описаны ниже.

- Регрессия используется для понимания взаимосвязи между зависимыми и независимыми переменными. Он обычно используется для составления прогнозов, например, для определения выручки от продаж для данного бизнеса. Линейная регрессия, логистическая регрессия и полиномиальная регрессия являются популярными алгоритмами регрессии.

Контролируемые алгоритмы обучения

В контролируемых процессах машинного обучения используются различные алгоритмы и вычислительные методы. Ниже приведены краткие объяснения некоторых наиболее часто используемых методов обучения, обычно рассчитываемых с помощью таких программ, как R или Python:

- Нейронные сети. В первую очередь используемые для алгоритмов глубокого обучения, нейронные сети обрабатывают обучающие данные, имитируя взаимосвязь человеческого мозга через слои узлов. Каждый узел состоит из входных данных, весов, смещения (или порога) и выходных данных. Если это выходное значение превышает заданный порог, оно “срабатывает” или активирует узел, передавая данные на следующий уровень в сети. Нейронные сети изучают эту функцию отображения посредством контролируемого обучения, корректируя ее на основе функции потерь в процессе градиентного спуска. Когда функция затрат равна или близка к нулю, мы можем быть уверены в точности модели, чтобы получить правильный ответ.
- Naive Bayes. Naive Bayes - это классификационный подход, который использует принцип условной независимости класса от теоремы Байеса. Это означает, что наличие одного признака не влияет на наличие другого в вероятности данного результата, и каждый предиктор оказывает равное влияние на этот результат. Существует три типа наивных байесовских классификаторов: многочленный Наивный Байес, Наивный Байес Бернулли и Наивный Байес Гаусса. Этот метод в основном используется в системах классификации текстов, идентификации спама и рекомендаций.
- Линейная регрессия. Линейная регрессия используется для определения взаимосвязи между зависимой переменной и одной или несколькими независимыми переменными и обычно используется для прогнозирования будущих результатов. Когда существует только одна

независимая переменная и одна зависимая переменная, это называется простой линейной регрессией. По мере увеличения числа независимых переменных это называется множественной линейной регрессией. Для каждого типа линейной регрессии он стремится построить линию наилучшего соответствия, которая вычисляется с помощью метода наименьших квадратов. Однако, в отличие от других моделей регрессии, эта линия является прямой при нанесении на график.

- Логистическая регрессия. В то время как линейная регрессия используется, когда зависимые переменные являются непрерывными, логистическая регрессия выбирается, когда зависимая переменная является категориальной, что означает, что они имеют двоичные выходные данные, такие как "истина" и "ложь" или "да" и "нет"." В то время как обе регрессионные модели направлены на понимание взаимосвязей между входными данными, логистическая регрессия в основном используется для решения задач двоичной классификации, таких как идентификация спама.
- Машина опорных векторов (SVM). Машина опорных векторов - популярная модель контролируемого обучения, разработанная Владимиром Вапником, используемая как для классификации данных, так и для регрессии. Тем не менее, он обычно используется для решения задач классификации, построения гиперплоскости, в которой расстояние между двумя классами точек данных максимально. Эта гиперплоскость известна как граница принятия решений, разделяющая классы точек данных (например, апельсины против яблок) по обе стороны плоскости.
- K - ближайший сосед. K-ближайший сосед, также известный как алгоритм KNN, представляет собой непараметрический алгоритм, который классифицирует точки данных на основе их близости и связи с другими доступными данными. Этот алгоритм предполагает, что похожие точки данных могут быть найдены рядом друг с другом. В результате он стремится вычислить расстояние между точками данных, обычно через евклидово расстояние, а затем присваивает категорию на основе наиболее частой категории или среднего значения. Простота использования и малое время вычислений делают его предпочтительным алгоритмом для специалистов по обработке данных, но по мере роста тестового набора данных время обработки увеличивается, что делает его менее привлекательным для задач

классификации. KNN обычно используется для систем рекомендаций и распознавания изображений.

- Случайный лес. Случайный лес - это еще один гибкий контролируемый алгоритм машинного обучения, используемый как для целей классификации, так и для целей регрессии. "Лес" ссылается на набор некоррелированных деревьев решений, которые затем объединяются вместе, чтобы уменьшить дисперсию и создать более точные прогнозы данных.

Обучение без присмотра по сравнению с обучением под наблюдением или полу-наблюдением.

Неконтролируемое машинное обучение и контролируемое машинное обучение часто обсуждаются вместе. В отличие от контролируемого обучения, неконтролируемое обучение использует немаркированные данные. На основе этих данных он обнаруживает закономерности, которые помогают решать проблемы кластеризации или ассоциации. Это особенно полезно, когда эксперты по предмету не уверены в общих свойствах набора данных. Распространенными алгоритмами кластеризации являются иерархические, k-средние и гауссовы смешанные модели.

Полууправляемое обучение происходит, когда помечена только часть заданных входных данных. Обучение без присмотра и под присмотром может быть более привлекательной альтернативой, поскольку полагаться на знания в предметной области для надлежащей маркировки данных для обучения под присмотром может быть трудоемким и дорогостоящим.

Список литературы

1. Машинное обучение – это легко [Электронный ресурс] // Хабр: Сообщество IT- специалистов – Режим доступа: https://habr.com/ru/post/319288/ – (Дата обращения: 01.02.2022).
2. Применение машинного обучени и Data Science в провышленности [Электронный ресурс] // Хабр: Сообщество IT- специалистов – Режим доступа: https://habr.com/ru/company/mailru/blog/462769/ – (Дата обращения: 01.03.2022).
3. Обзор задач компьютерного зрения в медицине [Электронный ресурс] // Хабр: Сообщество IT- специалистов – Режим доступа: https://habr.com/ru/post/309152/ – (Дата обращения: 05.01.2022).