УДК 004.855.5

ЭТАПЫ СОЗДАНИЯ МОДЕЛИ МАШИННОГО ОБУЧЕНИЯ

Кондратёнок Е. В.
Белорусский национальный технический университет,
Минск, Беларусь, elena_kondr@tut.by

Аннотация. Машинное обучение – это циклический многоэтапный процесс, в течении которого, исследователи, разработчики и инженеры разрабатывают, обучают, и обслуживают модель машинного обучения. Разработка модели машинного обучения принципиально отличается от традиционной разработки программного обеспечения и требует своего собственного способа решения практической задачи формированием набора данных и алгоритмическим построением статистической модели на основе набора данных.Полученная статистическая модель используется для решения практической задачи.

Введение.

В мире все чаще возникают новые предметные области. Каждая предметная область характеризуется знаниями, которые описывают объекты этой области и их свойства. В любой предметной области или отрасли генерируется огромное количество данных. Для эффективного использования этих данных применяются методы исследования данных и машинного обучения.

Согласно методологии Cross-IndustryStandardProcessforDataMining (CrispDM) [1] внедрение методов машинного обучения включает в себя постановку задачи, анализ данных и их подготовку, моделирование (построение прогнозирующей модели и подбор ее параметров), оценку полученной модели, внедрение и техническую поддержку.

Постановка задачи.

Машинное обучение применяется при наличии большого количества достоверных данных, которые связанны с предсказываемой величиной, а также в случае необходимости строить прогнозы в соответствии с динамично меняющимися условиями. В настоящее время существует широкий спектр задач для решения, которых можно применить алгоритмы машинного обучения.

Примерами таких задач являются различные рекомендации, прогнозирование результатов пользователей, распознавание голоса, распознавание графического контента, диагностика в медицине, оценивание заемщиков, предсказание ухода клиентов, прогнозирование потребительского спроса, предсказания рейтингов, анализ рыночных корзин, классификация объектов на основе данных о них, анализ текстовой информации, рубрикация текстов, ранжирование текстовых документов и другие [1].

При постановке задачи машинного обучения необходимо определить ключевые показатели модели, которые необходимо прогнозировать и метрики качества в соответствии с принципами SMART для определения успешности решения задачи.

Подготовка и анализ данных.

Задачей этапа подготовки и анализа данных является получении обработанного, высококачественного набора данных, который подчиняется некоторой закономерности. Этап состоит из четырех стадий: анализ данных, сбор данных, нормализация данных и моделирование данных.

Реальные данные для последующей обработки получают из различных источников, процессов, создают вручную или генерируют некоторым алгоритмом. При этом данные могут быть повреждены, пропущены, содержать ошибки, быть не надежными, избыточными или несогласованными. Модель, обученная на таких данных, выдаст неверные результаты при прогнозировании.

Несмотря на увеличение производительности оборудования, большой объем данных сложно передавать, обрабатывать, сохранять. Поэтому набор исследуемых данных распределяется по компьютерной сети в виде реляционных баз данных. Базы содержат данные в разных форматах и часто не структурированы. В связи с этим перед использованием в практических задачах данные преобразуют, упорядочивают и приводят в форму, эффективную для хранения и обработки алгоритмами машинного обучения [2].

Для машинного обучения такой набор данных называется датасет – это обработанная и структурированная информация в табличном виде. Строки такой таблицы являются объектами, а столбцы – признаками. Признаки делятся на предикторы (независимые переменные) и целевые (зависимые переменные). Целевые признаки вычисляются на основе одного или нескольких предикторов.

Качество данных необходимое условие для создания качественных моделей прогнозирования, т. к. влияет на способность моделей к обучению и на их эффективность.

Для определения и планирования необходимой предварительной обработки данных проводится исследовательский анализ полученных данных.

Стандартные методы проверки работоспособности данных учитывают их объем, количество пропущенных в них значений, формат, допустимые диапазоны значений. Для работы с пропущенными значениями данных существует ряд способов, зависящих от типа пропусков. При достаточно большом наборе данные с отсутствующими значениями могут быть удалены. В зависимости от библиотеки и конкретной реализации алгоритма допустимо использование алгоритма обучения, умеющего работать с отсутствующими значениями. Кроме этого существуют вычислительные методы восстановления данных. Например, метод замены отсутствующего значения средним или медианным значением, вычисленным по всему набору данных. При значительном отличии значения признака от типичных значений алгоритм обучения заменяет отсутствующее значение значением, выходящим за пределы диапазона нормальных значений, например, серединой диапазона, т. к. это значение не оказывает значительного влияния на прогноз [2]

Конструирование признаков.

Выделение признаков– это процедура удаления незначащих признаков из очищенной выборки перед загрузкой данных в алгоритм. Наличие большого количества признаков делает модель сложной, что увеличивает время выполне-

ния, снижает скорость вычисления и точность предсказания. Для алгоритмов машинного обучения важны только те данные, которые влияют на итоговый результат, т. е. связаны с целевым признаком. Для этого необходимо оценить связь между признаками. Конструирование признаков включает в себя учет, статистическую обработку и преобразование признаков, используемых в модели.

Существует множество способов отбора признаков для машинного обучения. Все они призваны показать их значимость и исключить некоторые из них на основании этой значимости. Под значимостью понимается набор метрик и диаграмм. Все методы отбора признаков делятся на несколько категорий. Методы фильтрации (filtermethods) основаны на теории вероятности и статистических подходах. Оценивается степень корреляции каждого признака с целевой переменной и признаки ранжируются по значимости. Методы фильтрации работают быстро и имеют низкую стоимость вычислений, но не имеют высокой точности, т.к. не учитывается взаимное влияние признаков друг на друга и на целевую переменную. Оберточные методы(wrappermethods) это поисковые алгоритмы. Они рассматривают признаки как входы, а эффективность модели как выходы, которые должны быть оптимизированы. Встроенные методы (embeddedmethods) выделяют признаки во время процесса расчета модели. Эти алгоритмы требуют больше вычислений, чем методы фильтрации, но меньше чем оберточные методы [2, 3].

Моделирование

На этапе моделирования происходит обучение модели. Важной частью этапа является выбор алгоритма машинного обучения. Для решения поставленной задачи можно использовать несколько алгоритмов обучения. Обучение модели имеет итерационный характер – выбираются различные модели, настраиваются необходимые гиперпараметры, сравниваются значения выбранной метрики с целью выбора лучшей комбинации. После каждой итерации результат модели сохраняется. На выходе получают результаты каждой модели и значения использованных в ней гиперпараметров. Для подбора гиперпараметров датасет делится на три выборки: обучающая выборка, валидационная выборка и тестовая выборка. Для непосредственного обучения модели используется обучающая выборка (англ. trainingset). По этой выборке производится настройка алгоритма, т. е. оптимизация его параметров. Валидационная выборка (англ. validationset) используется для расчета ошибки и выбора наилучшей модели. Тестовая выборка используется для тестирования выбранной модели. Методы формирования обучающей и оценочных выборок зависят от класса решаемой задачи. Например, для задач классификации данные делятся так, чтобы в полученных наборах численное соотношение объектов разных классов было равно численному соотношению объектов разных классов в исходной генеральной совокупности. В задачах регрессионного анализа в полученных наборах, которые будут использоваться для обучения и контроля качества, необходимо одинаковое распределение целевой переменной.

Самыми известными алгоритмами машинного обучения можно назвать линейную регрессию, логистическую регрессию, обучение дерева решений, метод опорных векторов, метод k-ближайших соседей [6].

Важным критерием является количество признаков, которое может обрабатывать алгоритм. Наибольшее количество признаков обрабатывают нейронные сети. Выбор алгоритма зависит и от того какие признаки преобладают в наборе данных, количественные или качественные. Ряд алгоритмов допускает добавление весовых коэффициентов. Некоторые модели классификации выводят из заданного набора признаков только класс, другие возвращают оценку от 0 до 1. Эту оценку можно считать степенью уверенности модели в прогнозе или вероятностью принадлежности образца к определенному классу. Ряд алгоритмов работают со всем набором данных и при появлении дополнительных размеченных данных необходимо обучить модель снова. А другие принимают данные потоками и обучаются итерациями. В этом случае можно использовать новые данные, чтобы обновлять модель. Ряд алгоритмов могут использоваться и для регрессии, и для классификации [2–5].

Полученные модели анализируются, сортируются по объективному или субъективному критерию и выбираются лучшие. Анализ моделей проводится с помощью метрик качества – функций, по которым валидируется качество обученной модели. В задачах классификации используются матрица ошибок (англ. confusionmatrix), аккуратность (англ. accuracy), точность (англ. precision), полнота (англ. recall), f-мера (англ. f-score), ROC-кривая, precison-recall кривая. В задачах регрессии используются средняя квадратичная ошибка (англ. MeanSquaredError, MSE), средняя абсолютная ошибка (англ. MeanAbsoluteError, MAE), коэффициент детерминации, средняя абсолютная процентная ошибка (англ. MeanAbsolutePercentageError, MAPE), корень из средней квадратичной ошибки (англ. RootMeanSquaredError, RMSE), симметричная MAPE (англ. Symmetric MAPE, SMAPE), средняя абсолютная масштабированная ошибка (англ. Mean absolute scaled error, MASE). Хорошим способом оценки модели является кросс-валидация (скользящий контроль или перекрестная проверка). В кросс-валидации проводится некоторое количество разбиений исходной выборки на две подвыборки: обучающую и контрольную. Для каждого разбиения выполняется настройка алгоритма по обучающей подвыборке и оценивается его средняя ошибка на объектах контрольной подвыборки. Оценкой скользящего контроля называется средняя по всем разбиениям величина ошибки на контрольных подвыборках. Кросс-валидация имеет несколько разновидностей: Валидация на отложенных данных (Hold-OutValidation), Полная кросс-валидация (Completecross-validation), k-fold кросс-валидация, t×k-fold кросс-валидация, Кросс-валидация по отдельным объектам (Leave-One-Out), Случайные разбиения (Randomsubsampling), Критерий целостности модели (Modelconsistencycriterion). При получении заданных критериев качества модель готова к внедрению [2–5].

После обучения и проверки модель тестируют с использованием реальных данных. Это позволяет убедиться, что модель действительно работает с большим набором данных, который не использовался ни для обучения, ни для про-

верки. Как и на предыдущих этапах, любые данные, которые используются на этом этапе, должны отражать проблемную область, с которой нужно взаимодействовать, используя модель машинного обучения.

Внедрение и техническая поддержка.

Внедрение модели предполагает ее доступность для других систем, которые могут отправлять ей данные и получать от нее прогнозы. На этапе внедрения необходимо предусмотреть систему мониторинга качества модели в реальном времени, которая бы отслеживала падение качества прогнозирования и сигнализировала о необходимости перенастройки модели. При устаревании модели или появлении новых данных предусматривается процесс пересчета ее параметров на новых данных. При существенном изменении условий эксплуатации модели, может потребоваться выбор новой, более подходящей модели.

Заключение.

Машинное обучение – наука об алгоритмах, которые самостоятельно настраиваются на полученных данных. В основном машинное обучение используется в задача прогнозирования, где по входным данным необходимо предсказать выходные данные. Преимущество машинного обучения в том, что прогнозирующую функцию не обязательно задавать в явном виде, а достаточно определить ее общий параметризованный вид, автоматически настроив параметры по обучающей выборке.

Применение машинного обучения подразумевает выполнение определенной последовательности действий. Успешность полученной модели зависит и от выбора алгоритма обучения, и от правильного выполнения каждого этапа.

Многое зависит от мощности вычислительной машины. При наличии персонального компьютера или сервера с многоядерным процессором, вычисления можно выполнять параллельно, что увеличит скорость обучения или прогнозирования. Кроме того, необходимо учитывать объем памяти компьютера, на котором будет обучаться модель.

Литература

1. Pete Chapman (NCR), Julian Clinton (SPSS), Randy Kerber (NCR), Thomas Khabaza (SPSS), Thomas Reinartz (DaimlerChrysler), Colin Shearer (SPSS) and Rüdiger Wirth (DaimlerChrysler), CRISP-DM 1. 2000 SPSS Inc. CRISPMWP-1104.

2. Сейновски, Т. Антология машинного обучения. Важнейшие исследования в области ИИ за последние 60 лет / Т. Сейвонски; Massachusetts Institute of Technology. – 2018.

3. Бринк, Х.Машинное обучение / Х. Бринкс, Дж. Ричардс, М. Феверолф. – СПб.: Питер, 2017. – 336 с.: ил. – (Серия «Библиотека программиста»).

4. Бурков, А. Машинное обучение без лишних слов / А. Бурков.– СПб.: Питер, 2020. – 192 с.: ил. – (Серия «Библиотека программиста»).

5. Флах, П. Машинное обучение. Наука и искусство построения алгоритмов, которые извлекают знания из данных / П. Флак; пер.с англ. А.А.Слинкина. –М.: ДМК Пресс, 2015. – 400 с.

6. Кондратёнок, Е. В. Машинное обучение как инструмент анализа данных / Е. В. Кондратёнок, С. Н. Макареня // Информационные технологии в образовании, науке и производстве: IX Международная научно-техническая интернет-конференция, 20–22 ноября 2021 года / сост. Е. А. Хвитько. – Минск: БНТУ, 2022. – С. 304–309.