Литература:



**Лекция 1-2**

Основы машинного обучения

1. Основные понятия и определения в области машинного обучения
2. Базовые этапы машинного обучения

**Основные понятия и определения в области машинного обучения**

Машинное обучение – одна из форм искусственного интеллекта, включающая робототехнику, обработку лингвистической информации и системы машинного зрения. Машинное обучение относится к специализированной совокупности знаний и связанным с ней техником. Определение Тома Митчела: компьютерная программа обучается, если ее производительность при выполнении определенной задачи, выражаемая в измеряемых единицах, увеличивается по мере накопления опыта.

В свое время, компания Кэгул объявила конкурс на разработку алгоритма, максимально точно отличающего собак от кошек. Для тренировки участникам предоставили 25т изображений с метками, указывающими, кто именно изображен на фото. После обучения, каждый алгоритм должен был классифицировать 12.5т не имеющих меток тестовых изображений.

Для обучения и обобщения, люди используют различные данные из примеров, включая формы, цвета, текстуры, пропорции и другие характеристики. МО также применяет множество стратегий в различных комбинациях в зависимости от поставленной задачи. Эти стратегии нашли воплощение в наборе алгоритмов, разработанных в самых разных дисциплинах: от статистики, робототехники и прикладной математики до поиска в Интернет, развлекательной сферы, цифровой рекламы и переводов с 1 языка на другой.

Алгоритмы крайне разнообразны, имеют сильные и слабые стороны. Некоторые относят объекты к определенному классу, другие предсказывают числовые значения. Существуют алгоритмы, определяющие сходства и различия допускающих сравнения сущностей. При этом, все алгоритмы обучаются на примерах (т.е. опыте), т.е. способны к обобщению.

Заявленные на конкурс распознавания программы, на этапе обучения пытались корректно выполнить классификацию, используя множество алгоритмов. На каждой из миллионов обучающих итераций, программа производила классификацию, измеряла полученных результат и затем, хотя бы немного, корректировала процесс в поисках постепенного улучшения. Победитель этого конкурса смог корректно распознать 98,9% ранее не демонстрировавшихся тестовых изображений. Это очень хороший результат, если учесть, что у людей частота 7% ошибок. Процесс МО анализирует изображения с метками и строит модель, которая в свою очередь используется процессом воспроизведения, или предсказания, для классификации новых изображений. В данном случае, мы рассматриваем т.н. **обучение с учителем.** Но существуют и другие виды МО.

МО может применяться к широкому кругу экономических задач: от обнаружения мошенничества до выбора целевой аудитории и рекомендаций товара, наблюдения за производством в реальном времени, анализа тональности текстов и мед. Диагностики. МО решает задачи, которые невозможно выполнить вручную из-за огромного количества подлежащих обработке данных. В случае больших данных (биг дата), МО иногда обнаруживает неочевидные зависимости, которые невозможно распознать при ручном рассмотрении.

**Самые распространенные техники обучения с учителем и варианты их практического применения.**

**Первая задача** – задача классификации. Описание этой задачи на основе данных определяется дискретный класс для каждого объекта. Примеры применения: фильтрация спама, анализ тональности текстов, обнаружение мошенничества, рассылка целевых рекламных объявлений, прогнозирование оттока клиентов, обработка заявок на тех поддержку, выявление производственных дефектов, сегментация потребителей, обнаружение событий, изучение геномов и эффективности лекарственных средств.

**Вторая задача** – регрессия. Описание этой задачи: на основе данных предсказывается фактическое значение параметра. Примеры применения: прогнозы на рынке ценных бумаг, прогноз спроса, цены, оптимизация аукциона рекламных объявлений, управление рисками, активами, прогнозы погоды, спортивные предсказания и тд

**Третья задача –** рекомендация. Описание: предсказывается альтернатива, которую предпочтет пользователь. Примеры: предложения продуктов, подбор персонала, предложение контента и тд

**4** – задача заполнения пропусков. Описание: вывод значений отсутствующих входных данных. Примеры: неполные истории болезней, отсутствующая информация о клиентах, данные переписей.

**В каких случаях стоит применять МО**

Рассмотрим реальную экономическую задачу, решение которой упрощается с помощью МО. Эта задача о принятии решения для предоставления кредита физ. Лицам для открытия малого бизнеса в неблагоприятных районах. *Традиционные подходы к анализу данных при рассмотрении заявлений:* анализ вручную (каждое заявление смотрится и по критериям сотрудник принимает решение) и бизнес-правила (фильтры, помогающие отсеивать клиентов, не попадающих под условия организации). Основной недостаток: система автоматически не учится на предоставляемых ей данных. Системы, управляемые данными, от простых статистических моделей до более проработанных обучающих рабочих процессов позволяют избежать таких проблем.

В случае рассматриваемой задачи, МО хорошо подходит, тк его природа позволяет поддерживать нужный темп как бы быстро не возрастал приток заявлений от клиентов. Более того, система извлекает оптимальные решения непосредственно из поступающих данных, не заставляя заранее жестко программировать выбираемые бизнес-правила. Такой переход от процесса анализа на основе бизнес-правил к МО означает, что точность решений не только возрастет, но и будет увеличиваться по мере выдачи новых кредитов. При МО данные служат основой для более глубоко представления управления. Чтобы определить наилучший порядок действий по отношению к каждой новой кредитной заявке, ML-система использует данные за прошлые периоды, на которых она обучалась.

Для запуска МО берутся данные о тысячи выданных кредитов. Это сведения из заявлений заемщиков и информации о возврате кредитов. Сведения из заявлений, в свою очередь, состоят из набора признаков. Это численные или категориальные показатели, фиксирующих значимые характеристики каждого заявления. К признакам относятся: оценка кредитоспособности заявителя, его пол и род занятий. После этого, модель с МО определяет, как можно использовать данные каждого заявления для наилучшего прогнозирования. Обнаруживая в наборе обучающих данных закономерности и применяя их, процедура МО создает модель. Пока ее можно рассматривать как черный ящик. Эта модель прогнозирует поведение заемщика на основе предоставленной им информации.

Следующим шагом является выбор алгоритма. Виды МО варьируются от простых статистических моделей до нетривиальных подходов. Мы сравним 2 примера: простую параметрическую модель и непараметрических ансамбль деревьев классификации. Большинство традиционных статистических экономических моделей попадает в первую категорию. В параметрических моделях соотношение между результатом и входными данными выражается через простые фиксированные уравнения. Данные в этом случае применяются для определения оптимальных значений неизвестных частей уравнения. В эту категорию попадают: модель линейной регрессии, логистической регрессии и модель авто регрессии с «лаговым оператором L»(?).

Для моделирования процесса одобрения кредита, возьмем логистическую регрессию. В ней логарифм отношения шансов каждого конкретного кредита на погашения рассматривается как линейная функция входных признаков. Будем полагать, что каждая новая заявка содержит 3 значимых признака: кредитный лимит заявителя, уровень его образования и возраст.

Тогда, логистическая регрессия попытается предсказать логарифм отношения шансов на невозврат кредита при помощи след уравнения:

Y= B0 + B1\*кредит.лайн + B2\*educat.lvl + B3\*h(возраст)

B – бета

B0, B1, B2 – некие коэф-ты.

Отношения шансов в этом случае – один из способ выражения вероятности. Оптимальное значение всех коэфов уравнения (в нашем рассматриваемом случаеB0-B3 ) определяется на основе 1тысячи примеров обучающих данных. Когда соотношение между результатом и входными данными удается выразить формулой, результат (y) можно легко предсказать по значениям признаков. Остается только определить, какие значения коэфов дают наилучший результат, воспользовавшись данными за прошедшие периоды.

Непараметрические ML-модели дают высокий уровень точности при работе со сложными многомерными реальными наборами данных. В их число попадают такие методы МО, как, к примеру, метод К-ближайших соседей, ядерное сглаживание, метод опорных векторов, деревья принятия решений и композиционное обучение.

Преимущество линейных алгоритмов, связанные с тем, что их проще интерпретировать, они быстрее обсчитываются и прозе масштабируются при увеличении набора данных.

**Пример тот же.** Для задачи выдачи кредита больше всего подойдет непараметрическая ML-модель. Она может обнаружить правила, которые вводились вручную, хотя, возможно, они будут выглядеть немного иначе для оптимизации статистических преимуществ. Скорее всего, модель самостоятельно выведет другую, более глубокую корреляцию входных параметров и желаемого результата. В этом случае, можно не только автоматизировать рабочий процесс, но и увеличить его точность, из чего непосредственно вытекает рост коммерческой стоимости.

Предположим, непараметрическая МЛ-модель дает прогноз с точностью на 25% большей, чем у алгоритма логистической регрессии. Это означает, что при обработке новых заявок, будет допущено меньше ошибок. В результате возрастет средняя доходность по кредитам, что позволит выдать больше кредитов и получить более высокий доход.

**5 преимуществ МО**

**1)Точность**. МО использует данные для создания принимающей решения программы, оптимизированной под поставленную задачу. По мере накопления данных, автоматически возрастает точность прогнозов.

**2)Автоматизация.** По мере подтверждения и отбрасывания ответов, МЛ-модель может автоматически обнаруживать новые шаблоны. Это позволяет встраивать МО в автоматизированные рабочие процессы.

**3) Скорость.**  МО дает ответы за доли секунды после поступления новой информации, позволяя системам реагировать в реальном времени.

**4) Возможность настройки.** Многие задачи, управляемые данными, можно решить с помощью МО. Модели строятся на базе собственных данных организации и допускают настройку под любую систему мер, принятых в бизнесе.

**5) Масштабируемость.** При росте бизнеса, МО-модель легко приспосабливается к увеличивающимся объемам данных. Некоторые алгоритмы можно использовать для обработки множества данных на разных ВМ в облаке.

Одновременно с преимуществами, есть и сложности:

1. Специалисты по работе с данными тратят на их подготовку примерно 80% времени.
2. При коммерческой деятельности фиксируется огромное количество данных, которые нельзя напрямую использовать в качестве входных данных для МО систем. Для извлечения из этих массивов полезных данных требуется большая работа. Достаточно сложно сформулировать проблему таким образом, чтобы к ней можно было применить методы МО и получить имеющие практическую ценность и измеримые результаты
3. Когда нужен сложный результат, выбор алгоритма и способы его применения сам по себе требует больших усилий.
4. Модель, которая идеально работает на обучающих данных, может демонстрировать полную неспособность к достоверным прогнозам на основании неизвестных ранее данных. Причиной, в большинстве случаев, становится переобучение

**Базовые этапы машинного обучения**

1. **Сбор и подготовка данных**. Сбор и подготовка данных для систем МО влечет за собой их представление в виде таблица. Большинство алгоритмов МО требуют именно такого формата. В этом случае, столбцы данных содержат данные одного типа, данные в строках принадлежат разным объектам. Может выполняться приведение некоторых типов данных к другим типам в зависимости от конкретного алгоритма МО. Реальные данные мб «запутаны» разными способами. Например, на этапе сбора данных нельзя измерить какое-то значение для признака объекта, нельзя вернуться назад, чтобы найти недостающий фрагмент информации. В подобных случаях некоторые ячейки таблицы останутся незаполненными, что усложнит как построение модели, так и последующее прогнозирование. Иногда сбор данных осуществляется вручную и допускаются ошибки при выполнении повторяющихся задач. В результате, часть данных оказывается некорректной. В этих случаях, алгоритм МО должен предусматривать обработку отсутствующих или недостоверных данных.
2. **Этап обучения модели на данных**. Первый этап построения успешной системы МО – это формулировка вопроса, ответ на который должны дать данные. Например, простая таблица с личными данными позволяет построить МО модель, предсказывающую семейное положение заявителя. Такая информация может использоваться, например, при выборе демонстрируемой пользователю рекламы. При этом переменная «семейное положение» будет использоваться как целевая, или target, или как метка label, а все прочие переменные станут признаками. МО алгоритм должен понять, каким образом набор входных признаков позволяет успешно предсказать значение целевой переменной. После этого можно использовать построенную модель для предсказания семейного положения заявителя, когда оно, по каким-то причинам не указано. На данном этапе, МО алгоритм можно представить в виде черного ящика, который выполняет отображение выходных признаков на выходные данные, но для построения работающей модели нужно более 2 рядов. Одним из преимуществ алгоритмов МО, в сравнении с другими распространенными методами, является умение обрабатывать множество признаков. Некоторые алгоритмы в достаточной степени не восприимчивы к неинформативным признакам, в то время, как другие дают более точные предсказания, когда такие признаки убираются из рассмотрения. Ценные сведения часто могут быть извлечены и из неинформативных на первый взгляд признаков. Такой вариант усовершенствования данных называется **извлечением признаков.** Предсказанная целевая переменная возвращается в той форме, в которой она фигурировала во взятых для обучения моделей исходных данных. Прогнозирование с помощью модели, по сути, является заполнением пустого столбца новыми значениями. Некоторые МО алгоритмы часто включают в результат своей работы связанные с каждым классов вероятности. Для нашего примера с семейным положением заявителей, вероятностная модель может выдавать для каждого нового заявителя 2 значения: вероятность, что человек состоит в браке и вероятность, что не состоит. **Итого**: любая система, связанная с МО, занимается созданием моделей и их применением для получения прогнозов.
3. **Оценка производительности модели**. Системы с МО практически никогда не используются пока не будет проверена их производительность. Всегда необходимо понимать, насколько готовая модель справляется со стоящими перед ней задачами. Возьмем набор данных и представим, что целевая переменная неизвестна. Можно построить на их основе модель и использовать данные в качестве тестовых для нескольких прогнозов. После этого можно сравнить результаты прогнозов с известными истинными значениями и оценить точность модели.
4. **Оптимизация производительности модели**. (увеличение точности прогнозов даваемых моделью). Увеличение точности достигается 3 способами:
   1. **Редактирование параметров модели.** Каждый МО-алгоритм обладает набором параметров, оптимальные значения которых заччастую зависят от типа и структуры данных. При этом на производительность модели может влиять значение как любого отдельного параметра, так и их произвольной комбинации
   2. **Выбор подмножества признаков.** Часто задачи, связанные с МО, включают множество признаков и вносимые ими помехи мешают алгоритму обнаружить верную закономерность даже если сами по себе эти признаки являются информативными. Тк заранее никогда не известно, как это скажется на эффективности модели, следует тщательно выбирать признаки, которые дадут наиболее универсальную и точную модель.
   3. **Предварительная обработка данных.** Реальные данные, используемые алгоритмами МО, часто подлежат очистке и обработке. Этот процесс называется **выпасом данных**. Набор данных может включать имена, которые пишутся по-разному, хотя и относятся к одной и той же сущности, а также отсутствующие или недостоверные значения. Все это негативным образом сказывается на эффективности модели.

**Усовершенствованные способы повышения эффективности.**

1. **Предварительная обработка данных и проектирование признаков.** В любой предметной области требуются специальные знания, позволяющие решать, какие именно данные подлежат сбору. Именно эти знания применяются для извлечения из собранных данных к ценной информации, которая добавляется к признакам строящейся модели. Этот процесс называется **проектированием признаков.** Несколько важных примеров проектирования признаков:
   1. **Дата и время**. Эти переменные часто встречаются в наборах данных, но для МО алгоритмов, которым, как правило, требуются необработанные числа или категории, сами по себе они бесполезны. Но в некоторых случаях из них можно извлечь ценную информацию. Например, при выборе демонстрируемой рекламы важно знать время, день недели и время года. Благодаря проектированию признаков, эти сведения извлекаются из переменных даты и времени и делаются доступными для модели. Возникают такие переменные и при наблюдении за повторяющейся активностью. Например, за повторными визитами посетителя на сайт в течение месяца или года. В этом случае, они позволяют вычислить промежутки между визитами, что может дать инфу для прогнозирования.
   2. **Местоположение**. В некоторых наборах данных присутствуют координаты в виде широты и долготы или названия места. Такая инфа позволяет извлечь из нее доп параметры, необходимые для решения специфических задач. Например, для прогнозирования результатов выборов требуется извлечь информацию о плотности населения, среднем доходе и тд
   3. **Цифровые средства коммуникации**. В эту группу попадают такие данные, как тексты, документы, изображения и видео. Изображение дают информацию о контурах, формах и спектре цветов. С помощью математических преобразований, они классифицируются, давая набор признаков, пригодных для алгоритмов классификации.
2. **Непрерывное совершенствование моделей**. В большинстве случаев, традиционные модели МО статические и перестраиваются редко. Но во многих случаях прогнозы возвращаются обратно в систему и желательно, чтобы модель постепенно совершенствовалась и адаптировалась к изменениям в этих данных. Существуют МО алгоритмы, поддерживающие данный тип динамического обучения. В этом случае, непрерывное переучивание встраивается в рабочий процесс МО.
3. **Масштабирование моделей.** Современные наборы данных быстро увеличиваются в размерах. Наборы для методов МО с учителем, в которых целевые ответы входят в обучающую выборку были традиционно небольшими, тк для получения ответов прибегали к помощи человека. В настоящее время, множество данных, включая ответы, продуцируются непосредственно измерительными элементами или ЭВМ. И для обработки таких объемов требуются уже масштабируемые МО алгоритмы.

**ЛЕКЦИЯ 3-4**

Тема: Этапы машинного обучения

При обучении с учителем используются данные, чтобы научить автоматизированные системы принимать решения. Алгоритмы МО проектируются таким образом, что умеют обнаруживать в данных за прошедшие периоды закономерности и ассоциации. Они учатся на этих данных и по результатам строят модели, которое будет предсказывать важные атрибуты для новых данных. Соответственно, тренировочные данные являются фундаментом МО. Качественные данные позволяют выявить особенности и корреляции данных и построить на их основе высокоточную прогнозирующую систему.

В то же время плохое качество обучающей выборки сводит на нет работу даже лучших МО алгоритмов.

**1 Этап. Сбор данных.**

Вначале необходимо корректно сформулировать вопрос, который будет решаться с помощью МО. Большинство реальных задач сводятся к предсказанию целевой переменной или переменных.

Во-вторых, в каждом вопросе четко прописана желательная форма ответа. Иногда, это бинарный ответ, иногда ответ принадлежит к набору классов, иногда принимает численное значение. В Статистике и теории ВС целевая переменная называется откликом или зависимой переменной.

В-третьих, для каждой из поставленных задач можно получить набор данных за прошедшие периоды с известной целевой переменной. В дополнение к известным целевым значениям, файлы с данными за прошедшие периоды будут содержать информацию для каждого экземпляра, который был доступен для изучения на момент создания прогноза. Эти входные признаки, также известные как объясняющие переменные, или независимые переменные. Например, для предсказания оттока клиентов какой-то компании может использоваться история использования продукта каждым клиентом вместе с демографическими характеристиками и сведениями об учетной записи.

Эти входные признаки вместе с известными значениями целевой переменной составят **обучающую выборку (training set).** И в финале каждый из поставленных вопросов подразумевает некоторые действия для определенных значений целевой переменной. Например, сведения о будущем спросе на продукт заставят организовать канал поставок для его удовлетворения.

Во всех описываемых случаях, МО алгоритм должен по обучающей выборке определить, каким способом набор входных признаков позволяет наиболее точно предсказать целевую переменную. Результат этого определения кодируется в виде модели МО, а признаки новых образцов с неизвестной целевой переменной скармливаются МО модели, которая генерирует для них прогнозы, давая конечному пользователю возможность предпринять действия.

Кроме этого, МО модель позволяет получить представление о том, как выглядит зависимость между входными признаками и целевой переменной.

**Рассмотрим задачу предсказания оттока клиентов из телекоммуникационной компании**. В данном случае экземпляром является абонент. Целевая переменная предполагает двоичный (бинарный) результат. Входные признаки могут включать в себя информацию об абоненте, которую можно узнать на начало месяца. Например, срок существования учетной записи, детали тарифного плана, а также сведения о пользовании услугами.

Сам процесс сбора данных сильно варьируется в зависимости от предметной области. Но при накоплении сведений для обучающей выборки есть ряд типовых вопросов. Их 4:

1. Какие входные признаки следует включить?
2. Как получить известные значения целевой переменной
3. Сколько обучающих данных требуется?
4. Как оценить качество обучающей выборки?

В задачах, связанных с МО, как правило, присутствуют десятки признаков, которые можно использовать для предсказания целевой переменной. В примере с оттоком абонентов ТК-компании на эту роль подходят все виды связанных с клиентом атрибутов. Это демографические атрибуты, тарифный план и его детали, данные о пользовании услугой.

Существуют два ограничения на использование сведений в качестве входного признака:

1. Значение признака должно быть известно на момент прогноза.
2. Признак должен быть численным или категориальным.

Простое эмпирическое правило гласит, что признак имеет смысл использовать только если он каким-либо образом связан с целевой переменной. Поскольку, целью МО с учителем является предсказание целевой переменной признаки, не умеющие с ней ничего общего, необходимо исключить.

Например, идентификатор абонента нет смысла использовать в качестве входного признака для прогнозирования ухода. Такие бесполезные признаки мешают отличать истинную взаимосвязь или сигналы от случайных помех данных (шума). Чем больше не информативных признаков, тем ниже соотношение сигнал-шум и, соответственно, менее точна в среднем МО-модель. Но при этом, точности модели может повредить исключение признака, о связи которого с целевой переменной изначально ничего не было известно.

При отборе входных признаков приходится идти на компромисное решение. С одной стороны, добавляя в модель все, есть риск заглушить группу признаков, которая содержит хоть какой-то сигнал, подавляющим все шумом. От этого пострадает точность модели, тк она перестанет отличать реальные шаблоны от случайной флуктуации. С другой стороны, тщательно выбирая только немногие признаки, гарантированно связанные с целевой переменной, можно не использовать (исключить из рассмотрения) полезную информацию. И это снова негативно повлияет на точность МО модели, тк о дополнительных признаках модель не узнает.

Исходя из этого, **на практике поступают следующим образом**:

1. Нужно включить все признаки, которые кажутся хоть как-то связанными с целевой переменной. Здесь же выполняется обучение МО модели. Если точность прогноза устраивает, то происходит остановка.
2. В противном случае, расширяется набор признаков путем добавления в него признаков, связь которых с целевой переменной менее очевидна. Снова выполняется обучение модели и оценка ее точности. И, если точность прогноза устраивает, происходит остановка.
3. Если точность модели не удовлетворительна, для расширенного набора признаков запускается алгоритм отбора, позволяющий выбрать оптимальное сильнее всего влияющее на процесс прогнозирования подмножество.

Алгоритмы отбора признаков ищут наиболее точную модель на основе подмножества из набора признаков, сохраняя сигнал и отбрасывая шум. Они крайне ресурсоемки, но дают очень большой прирост производительности модели.

Признак можно добавить в набор даже если он присутствует не у каждого экземпляра.

.

**Наблюдаемое значение целевой переменной.**

Одна из самых больших трудностей на 1 этапе МО с учителем это формирование обучающей выборки с известными значениями целевых переменных. Зачастую, для этого требуется работа существующей, условно оптимальной системы до тех пор, пока не будут собраны все данные. Например, при построение МО-решения, предсказывающего отток абонентов, нужно в течение достаточно длительного времени наблюдать за тем, как уходят одни абоненты и появляются другие. Получив достаточную для построения точной МО-модели обучающую выборку, можно задействовать непосредственно МО. В каждом случае процесс получения и оценки наблюдаемых значений целевой переменной будет разным. Примеры:

1. Рассылка целевых рекламных объявлений. В течение выбранного времени проводится кампания, чтобы определить, какие пользователи реагируют на рекламные баннеры, а какие нет.
2. Распознавание мошенничества. Исследование данных за прошлые периоды позволяет выявить недобросовестных пользователей.
3. Прогнозирование спроса. Журналы с информацией об управлении поставками позволяют узнать спрос за последние месяцы или годы.

Существуют и другие способы получения достоверных значений целевой переменной:

1. Нанять аналитика, который будет вручную просматривать текущие данные и данные за прошлые периоды с целью определения или оценки значений целевой переменной.
2. Привлечь к оценке значений целевой переменной широкий круг лиц (т.е. задействовать коллективный разум)
3. По результатам проведенных мероприятий опрашивать клиентов и проводить с ними другие практические эксперименты.
4. Запускать управляемые эксперименты и следить за ответами.

Все эти стратегии являются трудоемкими. Но процесс обучения можно ускорить, а время на сбор обучающей выборки сократить, если рассматривать целевые переменные только для экземпляров, оказывающих максимальное влияние на МО-модель.

Примером является метод активного обучения. При наличии существующей (маленькой) обучающей выборки и большого набора данных с неизвестной переменной отклика (целевой переменной), активное обучение идентифицирует в этом наборе подмножество экземпляров, включение которых в обучающую выборку позволяет получить максимально точную МО модель. В этом смысле, активное обучение ускоряет процесс построения модели МО, концентрируя собираемые вручную ресурсы.

**Какой размер обучающей выборки достаточен для построения и запуска МО модели.**

Существует ряд факторов, от которых зависит количество необходимых данных:

1. Сложность задачи. Можно ли описать связь между входными признаками и целевой переменной простым шаблоном или же она запутана и не имеет линейной зависимости.
2. Требование точности. Если достаточно всего 60% успешных результатов, можно обойтись меньшей обучающей выборкой.
3. Размерность пространства признаков. Например, если доступны всего 2 таких признака, обучающих данных потребуется меньше, чем при наличии 2000 таких признаков.

Существует важный принцип. По мере обучения обучающей модели, модели в среднем становятся более точными. Большая обучающая выборка обеспечивает более высокую точность из-за того, что МО-модели базируются именно на данных. Взаимосвязь между признаками и целевой переменной выводится целиком из обучающей выборки. Соответственно, чем большо размер этой выборки, тем выше способность модели распознавать и фиксировать менее выраженные шаблоны и соотношения.

Для примера задачи прогнозирования оттока клиентов, рассмотрим стратегию, позволяющую определить требуется ли дальнейшее увеличение обучающей выборки. Тренировочный набор данных ТК компании состоит 3333 экземпляров (абонентов), каждый из которых содержит 19 признаков, а также двоичный результат в виде ушедших или продливших обслуживание абонентов.

1. В имеющихся тренировочных данных выбираются подвыборки разных размеров для экспериментов. Например, если набор состоит из 3333 экземпляров, то сетка размеров может выглядеть так: 500 1000 1500 2000 2500 3000
2. Для каждой подвыборки случайным образом определяется указанное количество экземпляров из обучающей выборки
3. Для каждого подмножества тренировочных данных строится МО-модель и оценивается ее точность.
4. Рассматривается изменение точности как функция от размера выборки. Если при больших размерах эта функция выравнивается, существующего набора тренировочных данных скорее всего достаточно. Но, если точность продолжает расти, нужно увеличивать количество тренировочных данных.

**Репрезентативность обучающей выборки**

Кроме размера обучающей выборки, фактором, влияющим на точность прогнозов МО-модели, является репрезентативность этой выборки. Тк целью МО с учителем является генерация точных предсказаний для новых данных, крайне важно, чтобы обучающая выборка включала в себя экземпляры тех типов, для которых будет создаваться прогноз. Выборка с экземплярами, не имеющими отношения к будущим данным, называется **систематической ошибкой отбора, или ковариационным сдвигом**.

Отсутствие репрезентативности в обучающей выборке может быть обусловлено рядом факторов:

1. Получить экспериментальные значения целевой переменной удалось только для определенного содержащего ошибку подмножества данных.
2. Свойства экземпляров со временем менялись.
3. Набор входных признаков меняется со временем.

В каждом из вышеуказанных случаев, МО-модель, хорошо работающая с обучающей выборкой, не сможет хорошо выполнить экстраполяцию на новых данных. Чтобы избежать подобных проблем, обучающая выборка должна максимально репрезентативно представлять будущие данные. Значит, требуется процесс структурирования набора тренировочных данных, убирающий систематические ошибки

**2 этап. Подготовка данных к моделированию**

Самые простые МО-данные имеют числовой формат, но часто в набор входят и признаки других типов, например, категориальные переменные. Часть значений может просто отсутствовать. Иногда признаки приходится проектировать или рассчитывать.

Бывают численные признаки, требующие нормализации, чтобы их можно было сравнить друг с другом и привести в соответствие.

**Категориальные признаки.** Признак считается таковым, если его значения можно отнести к какой то группе. При этом не важен их порядок. Некоторые алгоритмы МО используют категориальные признаки в исходном виде. Но, обычно, данные требуется представить в численной форме. Можно присвоить каждой категории номер. Но такие зашифрованные данные нельзя использовать в качестве истинных категориальных признаков, тк при этом добавляется случайным образом выбранный пользователем порядок категорий. Лучше преобразовать каждую категорию в отдельный двоичный признак, имеющий значений 1 для экземпляров, попадающий в категорию и 0 для непопадающих. В результате, каждый категориальный признак преобразуется в набор двоичных признаков, по одному на категорию. Сконструированные таким способом признаки называют **вспомогательными переменными**.

Такая техника работает для большинства МО-алгоритмов. Но есть и алгоритмы, например, с деревьями принятия решений такие, как «случайный лес», которые используют категориальные признаки в исходном виде.

**Отсутствующие данные**

Если данные представлены в табличном виде, то отсутствующие данные принимают вид пустых ячеек или ячеек со знанием Nan (not a number). Как правило, это артефакты процесса сбора данных, возникающие в том случае, когда по какой-то причине конкретное значение для экземпляра данных измерить невозможно. Обработка двух основных типов отсутствующих данных происходит по-разному:

1. Иногда, сам факт отсутствия данных может быть информацией, полезной для МО алгоритма. Разумеется, если измерение значений было невозможно, этот факт не несет никакого глубинного смысла. Рассмотрим сначала информативные отсутствующие данные.

Если нам кажется, что в данных не хватает какой-то информации, значит мы хотим, чтобы МО алгоритм смог ею воспользоваться и теоретически увеличить точность предсказаний. Для этого отсутствующее значение нужно привести к формату остальных значений в столбце. Для числовых столбцов выбирается число с одного из концов числового спектра значений. При этом нельзя выбирать вариант из центральной части диапазона значений.

Для категориальных столбцов с потенциально информативными отсутствующими данными можно создать новую категорию, после чего новый категориальный признак обрабатывается обычным способом. Если же, отсутствие значений отдельных экземпляров не несет информационной нагрузки порядок действий будет другим. Добавлять специальное число или категорию уже нельзя, тк есть риск, что значение окажется некорректным.

Некоторые МО-алгоритмы игнорируют пропуски. Если же алгоритм этого не умеет, данные следует предварительно обработать, убрав отсутствующие значения или заменив их неким оценочным значением. Такой подход к обработке отсутствующих данных называется **заполнением пропусков**.

В случае большого набора данных с небольшим количеством отсутствующих значений, самое простое это отбросить их. Если же данные отсутствуют в большой части наблюдений, потеря данных уменьшит прогностическую силу модели. Более того, если распределение наблюдений с отсутствующими значениями по набору данных не является случайным отбрасывание может внести неожиданную систематическую ошибку.

Есть и другой подход. Можно предположить, что экземпляры данных обладают некоторой временной упорядоченностью и вставить на место пропусков значения из предшествующей строки того же столбца. В данном случае предполагается, что при переходе от одного экземпляра к другому результат измерений не изменился. Это некорректно.

Когда это возможно, лучше использовать большие фрагменты существующих данных для приблизительной оценки отсутствующих значений. Их можно заменить средним значением или медианой столбца. При отсутствии другой информации предполагается, что среднее значений ближе всего к истине. Но, в зависимости от распределения значений в столбце, иногда лучше использовать медиану. Среднее значение чувствительно к выбросам. Заменяя все отсутствующие значения одним новым, снижается наблюдаемость потенциальной корреляции с другими переменными, которая может помочь алгоритму распознать имеющиеся закономерности.

**Основы проектирования признаков**

Это использование существующих признаков для создания новых, увеличивающих ценность исходных данных путем применения наших знаний о данных или рассматриваемой предметной области. Некоторые МО-алгоритмы требуют, чтобы данные были нормализованы. Это означает, что каждый признак обрабатывается с целью его подгонки под единую числовую шкалу. Диапазон значений признака может влиять на его важность относительно прочих признаков.

Зачастую, нормализация сводится к переводу диапазона от 0 до 1 или от -1 до 1. Для получения значимых результатов любые новые данные, например, для предсказаний, требуют нормализации тем же способом, что и ранее нормализованные данные. Это означает, что пользователь выполняющий МО-моделирование, должен запоминать способы нормализации конкретных признаков и сохранять соответствующие значения.

**3 этап. Визуализация данных**

Этап предназначен для проверки работоспособности обучающих признаков и целевой переменной. Простые техники визуализации позволяют увидеть, как входные признаки связаны с целевой переменной. Эта информация служит ориентиром при построении модели и помогает лучше понять модель и генерируемые ею прогнозы.

Кроме этого, визуализация зачастую дает представления о степени репрезентативности обучающей выборки и о том, экземпляров каких типов там, возможно, не хватает.

Выделяют 4 техники визуализации:

1. Мозаичные диаграммы
2. Диаграммы размаха
3. Графики плотности
4. Диаграммы рассеивания

Каждая техника предназначена для конкретной комбинации типов (числовых или категориальных), входных признаков и целевой переменной.

1. Мозаичные диаграммы визуализируют связи между двумя и более категориальными переменными. ПО для создания таких диаграмм доступно на R, Python и прочих языках, предназначенных для статистической обработки данных
2. Диаграммы размаха, или ящики с усами. Это стандартная статистическая техника визуализации распределения численных переменных. Для одной переменной, такая диаграмма показывает квартили ее распределения. Диаграмма размаха одной переменной дает представление о центре, разбросе и асимметрии в распределении значений, а также о значении выбросов. Кроме этого, диаграммы размаха можно использовать для сравнения распределений параллельных трафиков. В частности, они позволяют визуализировать разницу в распределении числового признака как функцию от различных категорий категориальной переменной отклика.

Важно понимать ограниченность техник визуализации. Визуализация не заменят МО-моделирование. Модели с МО могут обнаруживать и использовать скрытые глубоко внутри данных взаимосвязи, которые не проявляются при простой визуализации. Не следует автоматически исключать из рассмотрения признаки только потому, что на диаграмме не видно их четкой связи с целевой переменной. Такая связь может обнаруживаться при их рассмотрении в совокупности с другими входными признаками.

1. Графики плотности. Отображает распределение одной переменной более детально, чем диаграмма размахов.

Во-первых, рассчитывается сглаженная оценка распределения вероятности переменной, обычно для этого используется техника, называемая **ядерным сглаживанием.**

Во-вторых, распределение отображается в виде кривой, показывая значения, которые скорее всего будет иметь переменная.

Создав график плотности для переменной плотности(?) из каждой категории входного признака, можно визуализировать любые расхождения значений для разных категорий. Графики плотности аналогичны гистограммам, но их сглаженная природа упрощает процесс визуализации набора распределений на одном рисунке.

1. Диаграмма рассеивания. Диаграмма рассеивания дает простую визуализацию соотношения между двумя численными переменными. Это один из самых популярных графических инструментов. Значение признака на такой диаграмме отмечается относительно значения переменной отклика. И каждый экземпляр представляется в виде точки. Несмотря на свою простоту, диаграммы рассеивания обнаруживают как линейные, так и нелинейные соотношения между входной переменной и переменной отклика.

**ЛЕКЦИЯ 5-6**

**Моделирование и прогнозирование**

**Основы моделирования с МО**

Цель МО – обнаружение закономерностей и взаимосвязей данных и практическое применение полученной информации. Процесс обнаружения реализуется различными методами, но все методы подчиняются одной задаче – оценить функциональное соотношение между входными признаками и целевой переменной



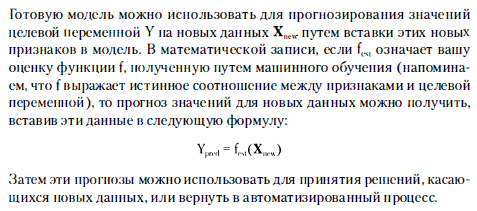
Рассмотрим моделирование в случае МО на примере набора данных Auto MPG. Набор данных содержит параметры: год выпуска, вес, мощность и тд. Общая задача: обнаружить соотношение между входными признаками и расходом топлива NPG. Входные признаки обозначаются X1, X2 … Xn. Целевая переменная обозначается символом Y. Соотношение между входными параметрами Х и результатом Y можно представить след формулой:

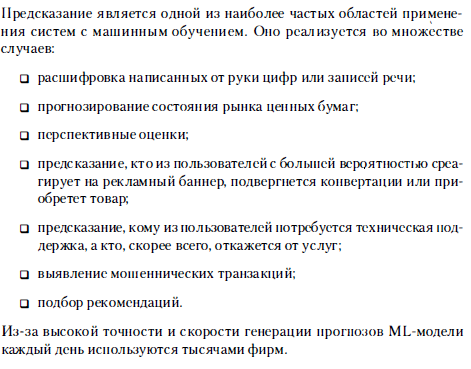


Цель МО моделирования – точно оценить функцию f, используя данные. Сложность МО состоит в том, чтобы использовать данные для определения настоящего сигнала, игнорируя при этом шум. В примере с набор данных АЛТОНПГ функция f описывает истинный расход топлива каждого авто как функцию от множества его входных признаков. Точно зная вид этой функции, можно получить информацию о расходе топлива любого автомобиля. Но при этом есть ряд источников шума:

1. Несовершенные результаты измерений расхода топлива в каждой машине, обусловленные неточностями в измерительных приборах – помехи при измерениях
2. Флуктуации производственного процесса, приводящие к тому, что каждый выпущенный автомобиль будет иметь немного различающиеся показатели расхода топлива – шум от производственного процесса
3. Шум при измерениях таких входных признаков, как вес и мощность автомобиля в лошадиных силах
4. Отсутствие доступа к более широкому набору признаков, который позволил бы точно определить расход топлива

Тк работать приходится с зашумленными данными, полученных измерением сотен автомобилей, техники МО способны дать всего лишь хорошую оценку функции f. Результирующая оценка и называется МО моделью.





**Типы методов моделирования**

