РАЗРАБОТЧИКАМ (HTTPS://ID-LAB.RU/CATEGORY/POSTS/DEVELOPERS/)

Вопросы и ответы из Data Science интервью

Освежите ваши статистические знания перед интервью с работодателем

Наша очередная публикация вопросов и ответов из интервью на позицию data scientist. Для плодотворной подготовки к интервью советуем посмотреть наши предыдущие публикации с задачами из интервью Google (https://id-lab.ru/posts/developers/zadachi-iz-intervyu-google-data-science/) и Facebook (https://id-lab.ru/posts/developers/zadachi-iz-intervyu-facebook-data-science/), а также полезные материалы от Теренс Шин (https://towardsdatascience.com/@terenceshin).

10 минут чтения

Фото с сайта admiralmarkets.com

Здесь собраны наиболее распространенные вопросы по статистике и теории вероятности из интервью для data scientists.

Необходимо учитывать, что фундаментальные знания в области статистики необходимы для успешной работы на позиции data scientist.

В. Чем отличается Байесовский подход к теории вероятности от классического? Сформулируйте теорему Байеса. Аналогом какого метода в классической теории она является? В чем заключаются преимущества Байесовской модели для Машинного Обучения?

Ключевое отличие заключается в том, как трактовать случайность. В классическом подходе случайная величина – это величина, значение которой мы принципиально не можем предсказать, это некоторая объективная неопределенность.

В Байесовском подходе случайная величина является детерминированным процессом, просто часть факторов, которые определяют исход этого процесса, для нас неизвестны.

Теорема Байеса позволяет определить вероятность (https://ru.wikipedia.org/wiki/%D0%92%D0%B5%D1%80%D0%BE%D1%8F%D1%82%D0%BD%D0%BE%D1%81%D1% какого-либо события при условии, что произошло другое статистически взаимозависимое (https://ru.wikipedia.org/wiki/%D0%A1%D1%82%D0%B0%D1%82%D0%B8%D1%81%D1%82%D0%B8%D1%87%D0%E с ним событие. Другими словами, по формуле Байеса можно более точно пересчитать вероятность, взяв в расчет как ранее известную информацию, так и данные новых наблюдений

Формула Байеса:

P(A|B) = P(B|A) * P(A) / P(B)

P(A) – априорная вероятность гипотезы *A*;

P(A|B) – вероятность гипотезы *A* при наступлении события *B* (апостериорная вероятность);

P(B|A) – вероятность наступления события *B* при истинности гипотезы *A*;

P(B) – полная вероятность наступления события *B*.

Теорема Байеса является аналогом метода максимального правдоподобия в классической теории вероятности

Особенность теоремы Байеса заключается в том, что для её практического применения требуется большое количество расчетов, вычислений, поэтому байесовские оценки стали активно использовать только после революции в компьютерных и сетевых технологиях.

Основным достоинством Байесовского подхода является возможность его применения для неполных, не размеченных или не полностью размеченных данных.

Байесовский метод позволяет использовать данные, которые поступают последовательно (не требует больших выборок – больших n, как в случае классического подхода)

Другое преимущество – это возможность использовать априорное распределение, которое позволяет избегать излишней настройки параметров, то есть избегать возможного переобучения модели

В. Напишите формулу распределения Пуассона. Приведите примеры задач, соответствующих распределению Пуассона

$P(x=k) = \lambda^k \exp(-\lambda) / k!$ – вероятность того, что случайная величина x равняется k, $\lambda > 0$, k = 0, 1, 2...

Распределение Пуассона описывает задачи-счетчики, например, число использования конкретного слова в тексте; число автобусов, проезжающих за час мимо автобусной остановки; число радиоактивных распадов, улавливаемых счетчиком Гейгера

В. Как вы оцените статистическую значимость наблюдения?

Вы должны выполнить проверку гипотез, чтобы определить статистическую значимость. Во-первых, надо сформулировать нулевую гипотезу и альтернативную гипотезу.

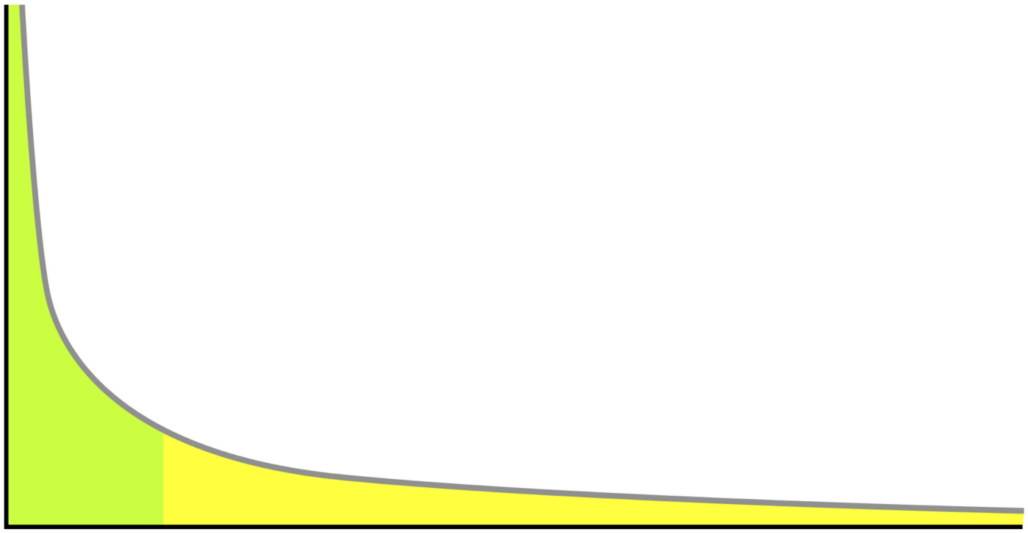
Во-вторых, надо вычислить уровень значимости p-value (вероятность получения наблюдаемых результатов теста, предполагая, что нулевая гипотеза верна).

Наконец, необходимо установить уровень значимости (альфа – обычно 0,05), и если p-value меньше альфа, то нулевая гипотеза отклоняется в пользу альтернативной. Другими словами, результат является статистически значимым.

Необходимо также осознавать ограниченность этого метода и его недостатки. Одна из некорректных практик заключается в принятии альтернативной гипотезы для любого p-value, номинально меньшего 0,05 без других подтверждающих доказательств. Хотя p-value полезны при оценке того, насколько несовместимы данные наблюдения с рассматриваемой статистической моделью, необходимо также учитывать контекст исследования.

В. Объясните, что такое распределение с длинным «хвостом», и приведите три примера соответствующих задач. Почему такие распределения важны в задачах классификации и регрессии?

Пример распределения с длинным «хвостом»:



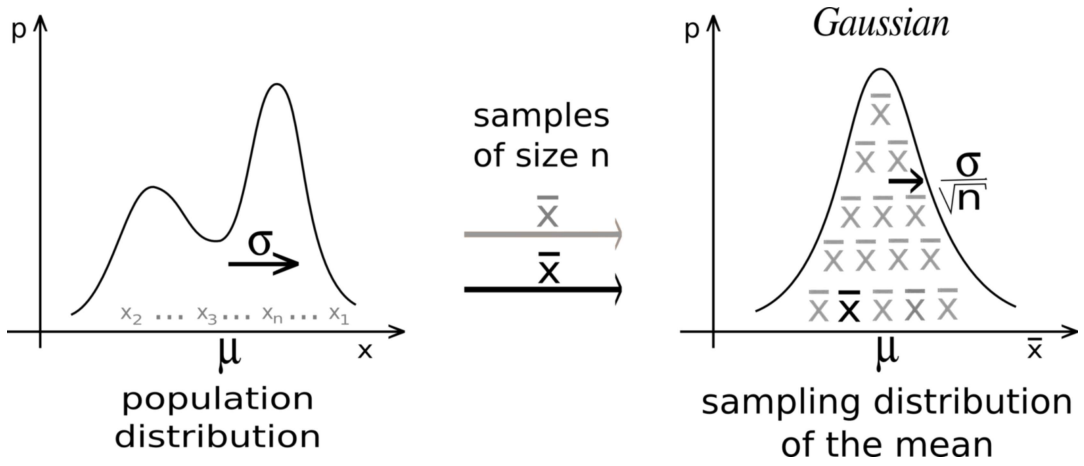
Длиннохвостое распределение представляет собой тип распределения, который имеет «хвост» (или «хвосты»), которые постепенно асимптотически уменьшаются.

– Принцип Парето (более известный как правило 80/20), когда в бизнесе 20% клиентов дают 80% торгового оборота

– Продажи продукта имеют длиннохвостовое распределение, когда 20% ассортимента формирует 80% прибыли

Важно помнить о длиннохвостых распределениях в задачах классификации и регрессии, потому что менее часто встречающиеся значения составляют основу выборки. В конечном итоге это может изменить способ работы с выбросами, а также вступить в противоречие с некоторыми методами машинного обучения, основанными на предположении, что данные имеют нормальное распределение.

В. Что такое центральная предельная теорема? Почему она важна?



Центральная предельная теорема утверждает, что распределение выборки среднего значения рассматриваемой выборки приближается к нормальному распределению, по мере увеличения размера выборки независимо от формы распределения рассматриваемой выборки.

Центральная предельная теорема важна, потому что она используется при проверке гипотез, а также для вычисления доверительных интервалов.

Она также является аналогом теоремы Байеса в классической теории вероятности.

В. Что такое статистическая мощность?

Статистическая мощность в математической статистике (https://ru.wikipedia.org/wiki/%D0%9C%D0%B0%D1%82%D0%B5%D0%BC%D0%B0%D1%82%D0%B8%D1%87%D0 вероятность отклонения основной (или нулевой) гипотезы при проверке статистических гипотез (https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%BE%D0%B2%D0%B5%D1%80%D0%BA%D0%B0_%D1%81%D1 в случае, когда конкурирующая (или альтернативная) гипотеза верна.

Статистическая мощность относится к силе бинарной гипотезы, то есть вероятности того, что тест отвергает нулевую гипотезу, учитывая, что альтернативная гипотеза верна.

P = P(отклонить нулевую гипотезу | альтернативная гипотеза верна)

В. Объясните, что такое смещенный / предвзятый отбор данных (систематическая ошибка отбора)? Почему важно этого не допускать? Как процедуры работы с данными, такие как обработка отсутствующих данных, могут ухудшить ситуацию?

Смещенный отбор данных (систематическая ошибка отбора) – это явление отбора отдельных лиц, групп или данных для анализа таким образом, что надлежащая рандомизация не достигается. Это в итоге приводит к выборке, которая не является репрезентативной для изучаемого распределения.

- **Смещенная выборка** : выборка, сделанная неслучайным образом
- **Интервал времени** : выбор определенного периода времени, который поддерживает желаемый вывод. Например, проведение анализа продаж под Рождество.
- **Воздействие** : включает предвзятость, индикацию.
- **Данные** : выбор конкретных данных (первой или последней точек, выбросов), подавляющий доказательства или ошибочность доказательств.
- **Истощение** : предвзятость истощения аналогична предвзятости выживания, когда в анализ включаются только те, кто «пережил» длительный процесс, или предвзятость неудачи, где включены только те, кто «провалился»
- **Отбор наблюдателей** : связан с антропным принципом, который является философским соображением, согласно которому любые данные, которые мы собираем о вселенной, фильтруются тем фактом, что для того, чтобы ее можно было наблюдать, она должна быть совместима с сознательной и разумной жизнью, которая ее наблюдает

Обработка пропущенных данных может усугубить смещение выборки, поскольку различные методы влияют на данные по-разному. Например, если вы заменяете нулевые значения средним значением данных, вы добавляете смещение в том смысле, что вы заранее предполагаете распределение данных.

В. Приведите простой пример того, как план эксперимента может помочь ответить на вопрос о поведении исследуемого объекта/явления. Как экспериментальные данные контрастируют с данными наблюдений?

Данные наблюдений поступают из наблюдений за исследуемыми явлениями/объектами, когда вы наблюдаете определенные переменные и пытаетесь определить имеющиеся закономерности.

Экспериментальные данные поступают из экспериментальных исследований, когда вы контролируете определенные переменные и держите их постоянными, чтобы определить имеющиеся зависимости.

Примером планового эксперимента является следующее: разбить исследуемую группу на две части. Контрольная группа живет своей жизнью нормально. Испытуемой группе рекомендуется выпивать стакан вина каждую ночь в течение 30 дней. Затем можно провести исследование, чтобы увидеть, как вино влияет на сон.

В. Является ли заполнение отсутствующих данных средними значениями приемлемой практикой? Почему да или почему нет?

Среднее вменение – это практика замены отсутствующих (NaN, None, NA) значений в наборе данных на среднее значение данных.

Среднее вменение, как правило, является плохой практикой.

Во-первых, оно не учитывает возможную зависимость признаков. Например, представьте, что у нас есть таблица, показывающая возраст и показатель физической подготовки. И представьте, что восьми летнему ребенку недостает этого показателя. Если мы взяли среднюю оценку физической подготовленности группы в возрасте от 15 до 80 лет, то восьми летний ребенок будет иметь гораздо более высокую оценку физической подготовленности, чем он на самом деле должен.

Во-вторых, среднее вменение уменьшает дисперсию данных и увеличивает смещение наших данных. Это приводит к менее точной модели и более узкому доверительному интервалу из-за меньшей дисперсии.

В. Вы работаете с ветеринарной клиникой, которая занимается здоровьем лошадей. Им необходимо знать, потребуется ли лошади операция, основываясь на различных медицинских показателях. Они отправили вам набор данных, где, как вы заметили, много пропусков. Как вы поступите? Всегда ли отсутствие данных означает отсутствие информации?

yes	adult	39.2	88	20	NA	slight
no	adult	38.3	40	24	normal	none
yes	young	39.1	164	84	cold	severe

Варианты действий

– убрать строки с NA

– убрать столбцы с NA

– поговорить с заказчиком, понять значения признаков и их значимость

– выяснить, что такое none

– заменить NA на среднее, медиану или предыдущее значение

Не всегда признак со значение NA надо исключать. Например, если конкретное лицо не владеет автомобилем, то другой признак для даты регистрации автомобиля будет содержать значение NaN, поскольку информации для заполнения нет.

В. Какие есть способы обработки недостающих данных? Какие методы вы рекомендуете?

Есть несколько способов обработки отсутствующих данных:

- Удалить строки с отсутствующими данными
- Удалить столбцы с отсутствующими данными (отказаться от части признаков)
- Заменить их на среднее / медиану
- Заменить константой (например, нулем)
- Попытаться предсказать недостающие значения
- Использовать алгоритм, который работает с пропущенными значениями. Например, Random forest

Наилучшим методом является удаление строк с отсутствующими данными, поскольку это гарантирует, что смещение или отклонение не будет добавлено или удалено, и в конечном итоге приведет к созданию надежной и точной модели. Однако это можно рекомендовать только в том случае, если есть достаточно данных и процент пропущенных значений невелик.

В. Что такое выброс (outlier)? Объясните, как можно обнаружить выброс и что бы вы сделали, если бы нашли их в своем наборе данных? Кроме того, объясните, что такое неявный выброс (inlier) и как вы можете их отфильтровать и что бы вы сделали, если бы нашли их в своем наборе данных?

Выбросом являются данные, которые существенно отличаются от других наблюдений.

Причиной выброса может быть:

- Ошибки измерения.
- Необычная природа входных данных. Например, если наугад измерять температуру предметов в комнате, получим цифры от 18 до 22 °C, но радиатор отопления будет иметь температуру 70°.
- Выбросы могут быть и частью распределения — так, в нормальном распределении (https://ru.wikipedia.org/wiki/%D0%9D%D0%BE%D1%80%D0%BC%D0%B0%D0%BB%D1%8C%D0%BD%D0%
каждое 22-е измерение будет выходить из «двух сигм (https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%B0%D0%B2%D0%B8%D0%BB%D0%BE_%D1%82%D1%8
и каждое 370-е — из трёх.

Z-оценка / стандартное отклонение: в этом случае 99,7% набора данных находятся в пределах трех стандартных отклонений. Мы можем рассчитать стандартное отклонение, умножить его на 3 и найти данные, которые находятся за пределами этого диапазона. Аналогично, мы можем вычислить z-показатель для данной точки, и если он равен +/- 3, то это выброс.
Обратите внимание: что при использовании этого метода необходимо учитывать несколько обстоятельств; данные должны быть нормально распределены, это не работает для небольших наборов данных, и наличие слишком большого количества выбросов делает z-показатель неприменимым.

Межквартильный диапазон (IQR): IQR – концепция, используемая для построения диапазонов отклонений, также может быть использована для выявления выбросов. IQR равен разнице между 3-м квартилем и 1-м квартилем. Таким образом можно определить, является ли точка выбросом, если она меньше Q1–1,5 * IQR или больше Q3 + 1,5 * IQR. Это соответствует приблизительно 2,698 стандартных отклонений.

Другие методы определения выбросов, это критерии Шовене (https://en.wikipedia.org/wiki/Chauvenet%27s_criterion), Пирса (https://en.wikipedia.org/wiki/Peirce%27s_criterion) и некоторые аналогичные подходы. Также возможно использовать методы кластеризации, такие как, например, DBScan (https://ru.wikipedia.org/wiki/DBSCAN).

Неявный выброс (inlier) это данные, которые лежат в пределах основного набора данных, но при этом являются необычными или ошибочными. Поскольку они находятся внутри набора данных, то их сложнее идентифицировать, чем выброс. Для их идентификации требуются дополнительные внешние данные.

Найденные неявные выбросы обычно удаляют из набора данных для устранения их влияния на проводимые исследования.

1. Алгоритм определения выбросов:

– определить максимум и минимум, в пределах которых лежит основной массив данных:

среднее +/- 1.5 * стандартное отклонение

или через межквартильное расстояние:

$[x_{25} – 1.5*(x_{75}-x_{25}), x_{75} + 1.5*(x_{75}-x_{25})] = [x_{median} – 2*(x_{75}-x_{25}), x_{median} + 2*(x_{75}-x_{25})]$

– выбрать все строки со значениями вне интервала минимум – максимум

2. Python – язык для работы с данными, поэтому оптимален для задач статистического анализа. Как альтернатива, можно использовать язык R

3. Причины выбросов:

- Из-за ошибки измерения.
- Из-за необычной природы входных данных. Например, если наугад измерять температуру предметов в комнате, получим цифры от 18 до 22 °C, но радиатор отопления будет иметь температуру в 70°.
- Выбросы могут быть и частью распределения — так, в нормальном распределении (https://ru.wikipedia.org/wiki/%D0%9D%D0%BE%D1%80%D0%BC%D0%B0%D0%BB%D1%8C%D0%BD%D0%
каждое 22-е измерение будет выходить из «двух сигм (https://ru.wikipedia.org/wiki/%D0%9F%D1%80%D0%B0%D0%B2%D0%B8%D0%BB%D0%BE_%D1%82%D1%8
и каждое 370-е — из трёх.

В. Что такое метод стохастического градиентного спуска? Назовите его основные достоинства и недостатки.

Градиентный спуск — метод (https://ru.wikipedia.org/wiki/%D0%93%D1%80%D0%B0%D0%B4%D0%B8%D0%B5%D0%BD%D1%82%D0%BD%D
нахождения *локального* экстремума (https://ru.wikipedia.org/wiki/%D0%AD%D0%BA%D1%81%D1%82%D1%80%D0%B5%D0%BC%D1%83%D0%BC) (минимума или максимума) функции (https://ru.wikipedia.org/wiki/%D0%A6%D0%B5%D0%BB%D0%B5%D0%B2%D0%B0%D1%8F_%D1%84%D1%83%D
с помощью движения вдоль градиента (https://ru.wikipedia.org/wiki/%D0%93%D1%80%D0%B0%D0%B4%D0%B8%D0%B5%D0%BD%D1%82).

Метод градиентного спуска оказывается очень медленным, особенно в случае большой размерности признакового пространства. Поэтому, часто в машинном обучении используют стохастический градиентный спуск, где каждый шаг вычисляется по градиенту одного случайно выбранного параметра.

Достоинством стохастической модификации градиентного спуска является его доказанная сходимость к тому же экстремуму, что и при градиентном спуске. При этом он намного более практичен и может использоваться на данных с большим числом признаков. Метод может быть обобщен для нелинейных моделей, использован с большим набором данных, а также с самыми разными функциями потерь.

К недостаткам метода можно отнести возможность сходимости метода к локальному, а не абсолютному экстремуму. Возможна также расходимость или очень медленная сходимость, поэтому нужно знать, какими способами можно ускорить сходимость этого метода. Наконец, в линейных моделях возможно переобучение из-за неприятного эффекта, который называется мульти-коллинеарностью.

В. У вас есть данные о продолжительности звонков в колл-центр. Создайте план того, как вы будете анализировать эти данные. Объясните вероятный сценарий того, как может выглядеть распределение этих длительностей. Как вы можете проверить, даже графически, оправдались ли ваши ожидания?

Обычно такие данные должны следовать логарифмически нормальному распределению

Пример логнормального распределения

Для графического подтверждения этого предположения возможно использовать график Q-Q (https://en.wikipedia.org/wiki/Q%E2%80%93Q_plot). Это позволит подтвердить, соответствует ли длительность вызовов логнормальному распределению или нет.

В статистике график Q – Q (квантиль-квантиль) – это график вероятности, который представляет собой графический метод для сравнения двух распределений вероятности путем построения их квантилей друг против друга. Сначала выбирается набор интервалов для квантилей. Точка (x, y) на графике соответствует одному из квантилей второго распределения (координата y), нанесенному на тот же квантиль первого распределения (координата x). Таким образом, линия является параметрической кривой с параметром, который является номером интервала для квантиля.
Если сравниваемые два распределения похожи, точки на графике Q – Q будут приблизительно лежать на линии y = x. Если распределения линейно связаны, точки на графике Q – Q будут приблизительно лежать на линии, но не обязательно на линии y = x.

B. Объясните различия между административными наборами данных и наборами данных, собранными в результате экспериментальных исследований. Какие проблемы характерны для административных данных? Как экспериментальные методы помогают облегчить эти проблемы? Какие проблемы они создают в свою очередь?

Административные наборы данных обычно представляют собой исторически накопленные наборы данных, собираемые правительствами или другими организациями по различным нестатистическим причинам.

Административные данные обычно представляют типичный образец big data, они весьма обширны и охватывают значительные периоды времени. Использование административных данных обычно проще и экономичнее, чем проводить экспериментальные исследования. Они также регулярно обновляются, если предположить, что организация, связанная с набором административных данных, активна и функционирует.

В то же время административные наборы данных могут не охватывать все данные, которые могут потребоваться, и могут не иметь желаемого формата. Другие недостатки это их обычно невысокое качество и отсутствующие записи.

Экспериментальные данные поступают из экспериментальных исследований, когда вы заранее фокусируетесь на нужных параметрах, а также контролируете определенные переменные и держите их постоянными, чтобы определить имеющиеся зависимости.

Примером экспериментального исследования является следующее: исследуемая группа людей делится на две части. Контрольная группа живет своей обычной жизнью. Испытуемой группе рекомендуется выпивать стакан вина каждую ночь в течение 30 дней. Затем можно провести исследование, чтобы увидеть, как вино влияет на сон.

B. Вы составляете отчет о пользовательском контенте, загружаемом каждый месяц, и отмечаете всплеск загрузок в октябре. В частности, всплеск загрузок изображений. Как вы думаете, что может быть причиной этого, и как вы это проверите?

Существует несколько возможных причин скачка загрузок фотографий:

1. Возможно, в октябре была реализована новая функция, которая связана с загрузкой фотографий и которая получила большую популярность среди пользователей. Например, функция, которая дает возможность создавать фотоальбомы.
2. Точно так же возможно, что процесс загрузки фотографий ранее не был интуитивно понятным и был улучшен в октябре.
3. Возможно, имело место вирусное движение в социальных сетях, которое занималось загрузкой фотографий, которые длились весь октябрь.
4. Возможно, что всплеск произошел из-за того, что люди выкладывают свои фотографии в костюмах к Хэллоуину.

Метод тестирования зависит от причины скачка. В общем случае вы должны провести проверку гипотез, чтобы выяснить причину всплеска. Надо задать нулевую и альтернативную гипотезы, порог уровня значимости (альфа – обычно 5%), и посчитать уровень значимости (p-value). Если p-value меньше альфа, то нулевая гипотеза отклоняется в пользу альтернативной.

Это позволит определить, что является действительной причиной всплеска загрузок фотографий.

To be continued...

Поделиться...

7 привычек плохих программистов (https://id-lab.ru/posts/news/7-privychek-plohih-programmistov/)

Читать полностью » (https://id-lab.ru/posts/news/7-privychek-plohih-programmistov/)

(https://id-lab.ru/posts/developers/4-nestandartnyh-tryuka-python-kotorye-polezno-znat/)

4 нестандартных трюка Python, которые полезно знать (https://id-lab.ru/posts/developers/4-nestandartnyh-tryuka-python-kotorye-polezno-znat/)

Читать полностью » (https://id-lab.ru/posts/developers/4-nestandartnyh-tryuka-python-kotorye-polezno-znat/)

(https://id-lab.ru/posts/news/preimushhestva-i-nedostatki-iskusstvennogo-intellekta-i-mashinnogo-obucheniya/)

Преимущества и недостатки Искусственного Интеллекта и машинного обучения (https://id-lab.ru/posts/news/preimushhestva-i-nedostatki-iskusstvennogo-intellekta-i-mashinnogo-obucheniya/)

Читать полностью » (https://id-lab.ru/posts/news/preimushhestva-i-nedostatki-iskusstvennogo-intellekta-i-mashinnogo-obucheniya/)

РАЗРАБОТЧИКАМ

КОНТАКТЫ

БЛОГ

НОВОСТИ

РАЗРАБОТЧИКАМ

КОНТАКТЫ

+7 985 7840521()

info@id-lab.ru(mailto:info@id-lab.ru)