Системы исскуственного интеллекта

приоритет2030^
лидерами становятся

# Подготовка данных для обучения

**Преподаватель:**

**Шарипов Ильдар Курбангалиевич**
**к.т.н., доцент, доцент кафедры электроснабжения и эксплуатации электрооборудования.**

# Тема 4. Подготовка данных для обучения.

## Вопрос 1. Основы подготовки данных.

## Вопрос 2. Подготовка модели.

## Вопрос 3. Основные проблемы при подготовке данных для машинного обучения.

## Вопрос 1. Основы подготовки данных

Машинное обучение становится все более популярным в современном мире, и все больше компаний и исследовательских организаций начинают применять его в своей работе. Однако, несмотря на все преимущества и потенциал, машинное обучение требует хорошо подготовленных данных для достижения максимальной эффективности и точности.

Подготовка данных для машинного обучения — это процесс, включающий несколько этапов, начиная от сбора данных и заканчивая их очисткой и преобразованием в удобный для обучения вид. Этот процесс является неотъемлемой частью любого проекта по машинному обучению и может существенно влиять на результаты исследования.

В этой теме мы рассмотрим 5 основных шагов, которые помогут вам подготовить данные для машинного обучения и достичь успеха в вашем проекте.

Содержание

Шаг 1: Понимание задачи машинного обучения

Перед тем как приступить к подготовке данных для машинного обучения, необходимо полностью понять задачу, которую мы хотим решить. Разработка моделей машинного обучения основана на анализе данных и нахождении закономерностей в них. Поэтому, чтобы получить качественные результаты, необходимо правильно поставить задачу.

Первый шаг в понимании задачи машинного обучения — определить, какую задачу мы хотим решить. Это может быть задача классификации, регрессии, кластеризации или прогнозирования. Каждая из этих задач имеет свои особенности и требует разных методов решения. Например, при классификации мы должны определить, к какому классу будет отнесен каждый объект, в то время как при регрессии мы должны предсказать численное значение.

Важно также определить цель нашей задачи. Что мы хотим достичь с помощью машинного обучения? Это может быть увеличение точности прогнозирования, автоматизация процессов, снижение затрат или другая конечная цель. Понимание цели поможет нам выбрать подходящую модель и определить метрики оценки качества.

Кроме того, на этом шаге мы также должны разобраться с доступными данными. Это включает понимание их структуры, формата, источников, а также проверку на наличие пропущенных значений, выбросов и других аномалий. Подробное изучение данных поможет нам определить необходимые преобразования и применить их на следующих шагах.

В итоге, понимание задачи машинного обучения является основой успешной подготовки данных. Хорошо поставленная задача поможет нам

выбрать правильные методы и модели, а также оценивать и интерпретировать результаты полученных моделей.

Шаг 2: Сбор и подготовка данных

В первую очередь, необходимо определить цель и задачу, которые вы хотите решить с помощью модели машинного обучения. Это поможет определить необходимые данные и критерии их качества.

Далее, следует сосредоточиться на сборе данных. Вы можете собирать данные самостоятельно, путем создания специальных форм для заполнения или путем автоматического сбора данных из различных источников. Важно убедиться, что данные собраны из надежных источников и отражают реальные ситуации и явления.

После сбора данных необходимо их подготовить для дальнейшего анализа и обучения модели. Этот этап может включать в себя различные операции, такие как очистка данных от выбросов и ошибок, заполнение пропущенных значений, преобразование данных в необходимый формат и нормализацию.

Изучение и визуализация данных также является важной частью подготовки данных. Это помогает понять особенности данных, взаимосвязи между признаками и выявить возможные зависимости.

Наконец, следует создать обучающую выборку, разделив данные на тренировочную и тестовую выборки. Это позволит оценить качество работы модели на новых данных.

В результате данного шага вы получите надежные и готовые к использованию данные, которые используете для обучения модели машинного обучения.

## Вопрос 2. Подготовка модели.

Шаг 3: Выбор и настройка модели

После предварительной обработки данных и разделения их на обучающую и тестовую выборки настало время выбрать и настроить модель машинного обучения. В этом шаге решается, какой алгоритм будет использоваться для обучения модели и как настроить его параметры.

Выбор модели зависит от цели задачи и типа данных, с которыми вы работаете. Существует множество алгоритмов машинного обучения, таких как линейная регрессия, решающие деревья, случайный лес, градиентный бустинг и другие. Каждый алгоритм имеет свои преимущества и недостатки, поэтому важно выбрать тот, который лучше всего соответствует вашей задаче.

После выбора модели необходимо настроить ее параметры. Это включает в себя определение гиперпараметров модели, таких как количество деревьев в случайном лесе или коэффициент скорости обучения в градиентном бустинге. Настройка параметров позволяет достичь лучшей производительности модели, учитывая особенности ваших данных.

Этот шаг также может включать выбор оптимального размера обучающей и тестовой выборок, а также алгоритма валидации модели. Выбор правильного способа валидации модели помогает оценить ее производительность и избежать переобучения или недообучения.

После завершения этого шага вы будете готовы обучить модель на обучающей выборке и произвести оценку ее производительности на тестовой выборке. Это позволит понять, насколько хорошо модель работает на новых данных и справляется с задачей машинного обучения.

Шаг 4: Обучение модели

После того как данные были предварительно обработаны и подготовлены, настало время перейти к обучению модели. В этом шаге мы позволяем нашей модели «увидеть» данные и выявить закономерности и связи между признаками и целевой переменной.

Для начала выберите алгоритм машинного обучения, который лучше всего соответствует вашей задаче. Это может быть классификация, регрессия или кластеризация, а может быть и другой подход, зависящий от ваших конкретных нужд и данных.

Затем разделите данные на тренировочный и тестовый наборы. Обучение модели будет проводиться на тренировочном наборе данных, а тестирование — на тестовом наборе. Это поможет вам оценить качество модели и ее способность обобщать данные.

После разделения данных приступайте к обучению модели. Используйте методы обучения, предусмотренные выбранным алгоритмом. Это может быть обучение с учителем (например, методы градиентного бустинга или нейронные сети) или обучение без учителя (например, методы кластеризации или ассоциативные правила).

Осуществляйте обучение модели, применяя выбранный алгоритм и используя тренировочный набор данных. В процессе обучения модель будет настраивать внутренние параметры и выравнивать их с целевой переменной.

После завершения обучения проанализируйте результаты. Оцените качество модели на тестовом наборе данных, используя метрики, соответствующие выбранной задаче (например, точность, полноту,

среднеквадратическую ошибку и т. д.). Это поможет понять, насколько хорошо модель работает и можно ли ее улучшить.

Важно помнить, что обучение модели — итеративный процесс. Возможно, потребуется провести несколько обучений, изменять параметры или применять другие методы машинного обучения, для достижения наилучших результатов. Также имейте в виду, что обучение модели может быть ресурсоемким процессом и может потребовать значительного времени и вычислительных ресурсов.

По окончании обучения модели, вы получите обученную модель, способную к предсказанию целевой переменной на новых, ранее неизвестных данных.

Шаг 5: Оценка и тестирование модели

После того, как модель машинного обучения обучена, необходимо оценить ее результаты и провести тестирование, чтобы убедиться в ее эффективности и точности.

Оценка модели включает в себя анализ различных метрик, таких как точность, полнота, f1-мера и другие, которые позволяют понять, насколько хорошо модель выполняет задачу, для которой она была разработана.

Для тестирования модели необходимо использовать отдельный набор данных, который не был использован в процессе обучения. Это позволяет проверить, насколько модель способна обобщать и применять полученные знания на новых данных. Тестирование помогает выявить проблемы в работе модели, такие как переобучение или недообучение, и внести необходимые корректировки для повышения ее качества.

| Метрика | Описание |
|---|---|
| Точность | Доля правильно классифицированных положительных образцов относительно всех положительных образцов. |
| Полнота | Доля правильно классифицированных положительных образцов относительно всех реально положительных образцов. |
| F1-мера | Численная характеристика, учитывающая как точность, так и полноту, и позволяющая оценить сбалансированность модели. |

После тестирования модели и анализа полученных результатов можно принять решение о том, является ли модель пригодной для использования в реальных условиях или требуется ее дальнейшая оптимизация и улучшение.

Важные аспекты подготовки данных для машинного обучения

1. Сбор и обработка данных: Начать следует с определения целевой переменной и сбора данных, которые будут полезны для ее предсказания. Очистка и обработка данных также являются неотъемлемыми шагами, включающими удаление дубликатов, заполнение пропущенных значений, нормализацию и преобразование данных в удобный для анализа формат.

2. Исследовательский анализ данных: Данные нужно изучить и проанализировать для определения взаимосвязей и корреляций между различными признаками. Это поможет выявить выбросы, аномалии и потенциальные проблемы, которые могут повлиять на работу модели.

3. Функциональное преобразование: В некоторых случаях данные нужно преобразовать, чтобы они стали лучше подходить для обучения

модели. Это может включать в себя создание новых признаков, применение математических функций или приведение данных к другому масштабу.

4. Выделение обучающей, проверочной и тестовой выборок: Для оценки модели необходимо разделить данные на обучающую, проверочную и тестовую выборки. Обучающая выборка используется для обучения модели, проверочная выборка оценивает ее эффективность и позволяет настраивать параметры модели, а тестовая выборка предсказывает результаты для оценки точности.

5. Учет дисбаланса классов: Если данные содержат дисбаланс классов, то модель может быть смещена в сторону более часто встречающегося класса. Это может привести к низкой точности модели на меньше представленных классах. Необходимо применить методы сбалансирования классов, такие как oversampling или undersampling, чтобы обеспечить равномерное представление классов в данных.

## Вопрос 3. Основные проблемы при подготовке данных для машинного обучения.

Отсутствие данных: Одной из основных проблем может быть отсутствие необходимых данных для обучения модели. Без данных модель не сможет выучить закономерности и принимать правильные решения.

Недостаточное количество данных: Даже если данные присутствуют, но их количество слишком мало, это может привести к недостаточной эффективности модели. Модель может стать переобученной и не сможет обобщить эти данные на новые примеры.

Нерепрезентативность данных: Если данные не отражают реальное состояние предметной области, то модель может помочь в принятии неправильных решений. Например, если данные не учитывают редкие случаи или несбалансированы по классам, модель может выдавать неверные результаты.

Некачественные данные: В данных могут присутствовать ошибки, выбросы, пропущенные значения и другие проблемы, которые могут негативно сказаться на качестве модели. Поэтому важно проводить предварительный анализ данных и устранять эти проблемы.

Неподходящий формат данных: Некоторые модели требуют определенного формата данных, например, числовой или категориальный формат. Если данные не соответствуют требуемому формату, то модель не сможет их корректно обработать.

Решение этих проблем требует тщательного анализа данных, их обработки и очистки. Необходимо обратить внимание на качество данных и принять меры для улучшения этого качества. Следование этим шагам поможет увеличить точность и эффективность модели машинного обучения.

Подготовка данных для машинного обучения в Python: лучшие практики

Импорт и изучение данных: Первым шагом является импорт данных в Python. После импорта следует изучить данные: проверить общую структуру, типы данных, наличие пропущенных значений и выбросы.

Обработка пропущенных значений: Встречающиеся пропущенные значения могут существенно повлиять на работу алгоритма машинного обучения. Необходимо решить, как обрабатывать их: удалить строки или столбцы с пропущенными значениями, заполнить их наиболее вероятными значениями или использовать другие методы обработки пропусков.

Кодирование категориальных признаков: Многие алгоритмы машинного обучения работают только с числовыми данными. Поэтому необходимо кодировать категориальные признаки в числовые, например, с помощью метода «one-hot encoding» или «label encoding».

Масштабирование признаков: Признаки разных масштабов могут негативно повлиять на работу алгоритмов машинного обучения. Необходимо привести признаки к одному масштабу, например, с помощью стандартизации или нормализации.

Правильная подготовка данных позволяет существенно улучшить результаты работы алгоритмов машинного обучения. Следуя описанным лучшим практикам, вы сможете достичь более точных и стабильных моделей.