«Методы нормализации: нормализация данных и слои нормализации»

Аптуков Михаил
Марков Михаил
СПбПУ, группа 5030102/10201

Нормализация данных

Цель работы: исследовать методы нормализации данных, их применение и влияние на результаты машинного обучения.

Нормализация данных – это процесс приведения входных данных к общему масштабу без искажения различий в диапазонах значений, что облегчает сравнение, анализ и обработку данных.

Используемые библиотеки и инструменты

Для реализации нормализации и анализа данных использованы библиотеки Python, такие как NumPy, Pandas, Scikit-learn.

Типы нормализации

Нормализация данных – ключевой этап подготовки данных для машинного обучения.

Цель нормализации: улучшить производительность моделей машинного обучения, обеспечивая единую шкалу для всех признаков.

Методы:

- **Min-Max нормализация**
- **Standard нормализация**
- **Batch нормализация**

Min-Max нормализация

$$x' = \frac{(x - x_{min})}{(x_{max} - x_{min})}$$

Этот метод преобразует данные так, что они находятся в заданном диапазоне, чаще всего между 0 и 1, что позволяет устранить влияние масштабов различных признаков.

Основная цель – улучшить производительность и стабильность алгоритмов машинного обучения.

Полезен:

для алгоритмов, в которых расстояния между данными играют ключевую роль;

при обработке изображений нейронными сетями.

Standard нормализация

$$z = \frac{(x - \mu)}{\sigma}$$

где x - исходное значение признака, μ - среднее значение признака, σ - стандартное отклонение.

Этот метод преобразует данные так, что они имеют среднее значение 0 и стандартное отклонение 1, что особенно полезно для гауссовских распределений.

Позволяет измерять все признаки в одних и тех же единицах, упрощая процесс обучения и делая его более эффективным, так как модель не будет предвзято относиться к признакам с большим масштабом.

Не подходит, если исходные данные имеют выбросы, которые могут исказить среднее значение и стандартное отклонение.

Batch нормализация

Этот метод используется для стабилизации и ускорения обучения нейронных сетей, нормализуя выходы каждого слоя.

Для каждого признака в мини-пакете Batch нормализация вычисляет среднее и стандартное отклонение и использует их для нормализации выходных данных слоя. Затем применяется масштабирование и сдвиг, параметры которых оптимизируются в процессе обучения.

Преимущества:

позволяет использовать более высокие скорости обучения, ускоряя сходимость;
сеть становится менее чувствительной к выбору начальных весов.

Недостатки:

менее эффективна при очень маленьких размерах мини-пакетов;
дополнительных обучаемых параметров увеличивает сложность модели.

ЭКСПЕРИМЕНТ
Описание датасетов

Использовались несколько стандартных **датасетов для анализа**:

- **CaliforniaHousing**: содержит данные о стоимости жилья и характеристиках домов в Калифорнии.
- **Diabetes**: содержит медицинские данные пациентов для прогнозирования прогрессирования диабета.
- **Iris**: содержит набор данных для классификации видов ирисов по их характеристикам.
- **Wine**: содержит химические анализы вина для классификации его типов.

Технологии:

- **Python** (NumPy, Pandas, Seabon, Matplotlib): для работы с данными и их визуализации.
- **Модели нейронных сетей**: использовалась библиотека Keras (TensorFlow) для создания моделей регрессии и классификации.
- **Скалирование данных**: применялись методы нормализации данных – Min-Max, Standard – для улучшения обучения моделей.

Методология эксперимента

1. Подготовка данных: данные разделены на тренировочные и тестовые выборки, применены методы нормализации.
2. Созданы нейронные сети с использованием Batch нормализации и Dropout для предотвращения переобучения.
3. Для задач регрессии и классификации построены и обучены разные модели.
4. Оценка производительности выполнена с использованием метрик mean_squared_error для регрессии и accuracy_score для классификации.

Результаты эксперимента

- ✓ Обучение проводилось на GPU, что значительно ускорило процесс.
- ✓ Для задач регрессии на датасете CaliforniaHousing наилучшее среднеквадратическое отклонение достигнуто при использовании нормализации данных.
- ✓ Для классификации на датасетах Iris и Wine точность достигла порядка 95%, при этом Batch нормализации и Dropout улучшили устойчивость модели к переобучению.

Представлен краткий обзор эксперимента, который подчёркивает использование различных подходов к обработке данных и обучению моделей.

Заключение

Важно учитывать **специфику данных** при выборе метода нормализации. Правильный выбор **метода нормализации** улучшит производительность модели.

При работе с данными, имеющими **различные масштабы**, следует использовать **Standard или Min-Max нормализацию** для обеспечения одинакового влияния всех признаков на обучение.

В глубоком обучении важно применять нормализацию внутри слоев, например, **Batch для улучшения сходимости и устойчивости обучения**.