Вики

Выделение признаков (Feature engineering)

> Синонимы: Конструирование признаков, Генерация признаков, Feature extruction
> Разделы: Алгоритмы

В машинном обучении и статистическом моделировании этап предобработки данных играет важную роль. На нем сырые необработанные данные преобразуются в обучающее множество, содержащее признаки, которые описывают наблюдения и пригодны для построения обучаемых моделей.

Выделение признаков (feature engineering) не следует путать с их отбором (feature selection), так как последний этап не формирует новые признаки, а выбирает наиболее значимые из уже имеющегося набора, что позволяет снизить размерность задачи. В то же время выделение признаков создает их из, например, текста, изображений и других видов неструктурированных или слабоструктурированных данных. Этот процесс может включать преобразование уже существующих признаков в форму, которая обеспечивает более эффективное обучение.

Например, данные о стоимости недвижимости могут быть представлены ее площадью и ценой. На основе этих признаков может быть сгенерирован новый признак — стоимость за квадратный метр, который позволит по-новому взглянуть на проблему и сделать прогнозные оценки более стабильными.

Другой пример: если в базе данных для каждого клиента указаны даты сделанных им покупок, то новым интересным признаком может стать среднее время между покупками для каждого потребителя за некоторый период. Так, если для какого-то из них данное время начинает увеличиваться, это дает повод задуматься о снижении лояльности.

Для построения моделей машинного обучения используется признаковое описание объектов. Поэтому обучающий набор данных должен быть структурированным, т.е. в строках содержать наблюдения, а в столбцах — признаки. Однако не все источники могут «поставлять» структурированную информацию, пригодную для признакового описания, например, текст или изображения.

Методы извлечения признаков существуют в различных формах: от статистических инструментов, таких как метод главных компонент для уменьшения размерности, до подходов, специфичных для предметной области, которые позволяют извлекать соответствующую информацию из текста, изображений или других типов данных.

Наиболее простым методом извлечения признаков из текста является «мешок слов».