УДК 004 ИНФОРМАЦИОННЫЕ ТЕХНОЛОГИИ

НАЗВАНИЕ ТЕЗИСА ДОКЛАДА

Автоматический выбор моделей предобработки данных для задач машинного обучения

Введение. В докладе будут рассмотрены подходы к автоматическому выбору моделей предобработки данных для задач машинного обучения. Традиционно предобработка данных выступает в качестве первых шагов в пайплайнах (pipleines) машинного обучения и оказывает существенное влияние на итоговый результат для типовых классов задач: классификации и регрессии. Для решения задачи предлагается использовать два альтернативных подхода: на основе ручного подбора моделей и процедуры кросс-валидации, либо на основе методов автоматического машинного обучения (AutoML).

Основная часть.

Процесс построения моделей машинного обучения и выбора лучшей модели является достаточно трудоемким, и одну из ключевых задач в этом процессе играют методы предобработки данных (data preprocessing). При этом в настоящее время наблюдается тенденция к автоматизации типовых операций машинного обучения, включая операции предобработки. К таким операциям относятся кодирование признаков, заполнение пропусков, нормализация данных, отбор признаков, снижение размерности. В зависимости от конкретной задачи данный список может как уменьшаться, так и расширяться. Выбор данных методов и их параметров может быть осуществлен как вручную, путем их программной реализации на низком уровне с использованием популярных библиотек, так и с использованием средств AutoML, в том числе на базе высокоуровневых платформ для машинного обучения. На практике выбирается метрика качества для данной задачи, а далее явно или неявно реализуется процедура кросс-валидации, на основе которой выполняется выбор одной процедуры предобработки или комбинации нескольких процедур. Объектом разработки являются методы, которые позволяют с наименьшими вычислительными затратами получить последовательность оптимальных моделей предобработки данных в рамках единого пайплайна машинного обучения для конкретной задачи.

Выводы. Предлагаемый подход к автоматическому выбору моделей предобработки данных для задач машинного обучения предполагает автоматизацию процесса выбора наилучших методов предобработки данных на основе процедур кросс-валидации и автоматического машинного обучения. Ручная реализация данных процедур для каждой конкретной задачи является весьма трудоемкой как по времени разработки, так и по времени вычислений, поэтому предлагаемый подход в целом является актуальным и позволит повысить качество

моделей машинного обучения, разрабатываемых в том числе и неопытными специалистами по данным.

Бунэхас Сабир (автор)

Иванов Сергей Владимирович. (научный руководитель)