В России придумали способ определять оптимальное количество данных для обучения нейросетей

Учёные из МФТИ разработали два новых метода для определения нужного объёма данных для обучения моделей машинного обучения. Это важно, так как слишком мало данных может снизить точность модели, а слишком много — привести к лишним затратам. Используя технику бутстрэпа, специалисты предложили критерии, которые помогают лучше понять, когда данных достаточно для построения качественной модели.

В их подходах основное внимание уделяется функции правдоподобия, которая показывает, насколько вероятны наблюдаемые данные для заданных параметров модели. В частности, они предложили два критерия: D-достаточность, которая проверяет стабильность модели при использовании разных подвыборок данных, и M-достаточность, которая оценивает, перестаёт ли модель улучшаться при добавлении новых данных. Эти методы, по словам исследователей, могут быть применимы в самых разных областях машинного обучения и не требуют строгих статистических допущений.

Технология бутстрэпа, которая используется для оценки стабильности правдоподобия, позволяет многократно создавать подвыборки из исходных данных и проверять, как меняется модель. Эксперименты показали, что предложенные методы эффективно работают как на синтетических, так и на реальных данных, включая известные наборы данных для регрессии и классификации. Эти результаты открывают новые возможности для планирования экспериментов и более эффективного использования данных в различных областях, таких как медицина, финансовый анализ и социология.