ОПЫТ УЧАСТИЯ В МЕЖДУНАРОДНЫХ СОРЕВНОВАНИЯХ ПО АНАЛИЗУ ДАННЫХ

Никулин В.Н., Багаев И.В., Канищев И.С.

Вятский государственный университет, ф-т экономики и менеджмента, кафедра ММЭ, Россия, 610000, г.Киров, ул.Московская, д.36, E-mail: vnikulin.uq@gmail.com

Анализ данных (data mining) или вычислительная статистика является сравнительно новой научной областью. Важность анализа данных (АД) обусловлена огромными базами данных, которые порождаются передовыми технологиями. В настоящее время, вычислительная статистика является одним из наиболее динамичных и перспективных научных направлений. Успешные результаты в области АД базируются на глубоком понимании статистических закономерностей и на свободном владении языками научного программирования. Отметим, что в отличии от теоретической статистики, вычислительная статистика ориентирована на обработку и анализ реальных данных. Соответственно, центральным вопросом является выбор оценивающего критерия, согласно которому производится сравнение различных методов и алгоритмов.

Международные соревнования по АД приобрели значительную популярность за последние 3-5 лет. Этот факт не удивителен поскольку эти соревнования включают в себя два очень важных качества. С одной стороны, это независимая оценка при использовании данных, которые были не доступны для построения модели. С другой стороны, это возможность участия в эксперименте десятков и сотен независимых команд. Представляется нереалистичным ожидать, что одна группа учёных может владеть всем разнообразием и всей многоплановостью знаний и опытом множества команд из различных стран. Формирование достаточно больших баз данных для экспериментов, основанных на реальных наблюдениях, является чрезвычайно важной и трудоёмкой задачей. Организаторы соревнований располагают всеми необходимыми специфическими качествами для успешного выполнения этой задачи. Область приложения здесь не ограничена и включает экономику, финансы, медицину, экологию, спорт и образование. Практический опыт является лучшим способом обучения, а участие в соревнованиях по вычислительной статистике может быть очень полезно для научных работников, прикладников и, в особенности, для студентов.

В наших курсах по АД мы используем данные Credit (финансовый риск кредитования) и MNIST (распознавание рукописных цифр). Эти и многие другие данные могут быть получены с платформы Kaggle (https://www.kaggle.com/). Мы делим наблюдения случайным образом на три части (триплет): 1) тренировка, 2) валидация (самоконтроль) и 3) тестирование, где третий случай используется для проверки знаний и навыков, поэтому метки (ответы) не предоставляются. Студенты (в составе небольших групп) анализируют отдельный триплет. Оценка всех произведённых решений осуществляется в автоматическом режиме. Процесс обучения проходит в форме локального соревнования, что особенно стимулирует студентов изучать наиболее передовые методы машинного обучения. В наших курсах мы используем следующие платформы и языки программирования: R, Matlab, Python и JAVA/Weka.