Алгоритмическая статистика

Алексей Милованов
almas239@gmail.com

3 октября 2016 г.

1 Об алгоритмической статистике вкратце

Допустим, у нас имеются некоторые экспериментальные данные, для которых требуется найти какое-нибудь "хорошее объяснение". Как такое объяснение найти, и вообще: какое объяснение следует считать хорошим? На последний вопрос можно ответить так: нужно найти такую модель, в которой вероятность произошедшего события была бы как можно больше. Однако такой ответ не следует считать удовлетворительным - мы всегда можем построить такую модель, в которой вероятность произошедшего события равняется единице (но такие модели редко являются приемлемыми с точки зрения практики). Поэтому следует добавить еще один критерий адекватности модели - она должна быть (по возможности) *простой*. Формализовать это требование удалось А. Н. Колмогорову с помощью алгоритмической теории информации в 1974 году. С этого момента и берет начало “алгоритмическая статистика”.

В качестве мере простоты модели предлагалось использовать ее *Колмогоровскую сложность* т.е. длину минимальной программы, которая порождает данную модель (другими словами, модель является простой, если ее можно хорошо заархивировать). С тех пор по данной теме было написано десятки статей, можно выделить следующих авторов: А. Х. Шень, Н. К. Верещагин, P. M. B. Vitányi. Одним из важнейших открытий было понятие *нестохастического* объекта, т.е. такого, для которого “хорошего” объяснения не существует в принципе.

По сути, один из важнейших принципов машинного обучения – Minimum description length – берет свои корни из алгоритмической статистики.

2 Ссылки на подробное изложение

Элементарным введением является видеолекция Н. К. Верещагина:
http://www.mathnet.ru/php/seminars.phtml?option_lang=rus&presentid=5097.

Более обстоятельно об этой науке рассказано в обзоре "Algorithmic statistics: forty years later" (https://arxiv.org/abs/1607.08077), см. также 14-ую главы книги "Колмогоровская сложность и алгоритмическая случайность" и 5-ую в "An Introduction to Kolmogorov complexity and its applications".

3 О возможных темах научно-исследовательских работ

Сразу следует отметить, что алгоритмическая статистика является теоретической наукой[1]. Она пытается ответить на вопрос "*Почему* методы машинного обучения оказываются столь эффективными на практике" (а не как сделать эти методы еще более эффективными). Возможными подтемами являются:

- алгоритмическая статистика и теория сложности вычислений;
- ограничивающие классы гипотез в алгоритмической статистике.

(Однако инициатива со стороны студента по поводу выбора конкретной темы только приветствуется.)

4 Обо мне

Я – аспирант механико-математического факультета МГУ, преподаватель МФТИ и ВШЭ. Моя страница: https://www.hse.ru/org/persons/176000050. Моими возможными помощниками в руководстве студентов могут быть Н. К. Верещагин и А. Х. Шень.

[1] Что не мешает использовать ее принципы для практических нужд, см., например, статью "Approximating Rate-Distortion Graphs of Individual Data: Experiments in Lossy Compression and Denoising "