

Использование мультимоделирования и привилегированного обучения для построения моделей оптимальной сложности.

Нечепуренко Иван Олегович

Московский физико-технический институт
Факультет инноваций и высоких технологий
Кафедра анализа данных

Научный руководитель: В. В. Стрижов
Консультант: Р. Г. Нейчев
21 марта 2019

Цель работы

Создать метод построения моделей оптимальной сложности для задач обучения с учителем, использующий привилегированную информацию для улучшения сходимости.

Проблема

Существует зависимость сходимости алгоритмов, имеющих низкую вычислительную сложность, от параметров их инициализации.

Метод решения

Использовать априорную информацию, производимую более сложной моделью - учителем. Построить модель в виде композиции более простых моделей.

Смесь моделей

- Yuksel Seniha Esen, Wilson Joseph N., Gader Paul D. Twenty Years of Mixture of Experts // IEEE Transactions on Neural Networks and Learning Systems. 2012. 23, № 8. С. 1177–1193.
- Большая обзорная статья

Привилегированное обучение

- Learning using privileged information: Similarity control and knowledge transfer. V.Vapnik, R.Izmailov. JMLR, 2015. -
Использование привилегированного обучения применительно к SVM.
- Unifying distillation and privileged information. B.Scholkopf, V.Vapnik, D.Lopez-Paz, L.Bottou. ICLR, 2016. - Обобщение подходов Вапника и Хинтона к привилегированному обучению.

Общая задача

Набор объектов - \mathbf{X} . У каждого объекта есть набор признаков, лежащий в \mathbb{R}^m . Такие значения можно задать матрицей $\mathbf{X} = [\mathbf{x}_i]_{i=1}^n$, где \mathbf{x}_i - вектор признаков i -го объекта. Также есть матрица ответов $\mathbf{Y} = [\mathbf{y}_i]_{i=1}^n$. В общем случае задача - построить алгоритм \hat{f} , минимизирующий заданную целевую функцию $S(\mathbf{y}_i, \hat{f}(\mathbf{x}_i))$

Задача декодирования

Если матрица ответов состоит из действительных векторов $\mathbf{y}_i \in \mathbb{R}^r$, то задачу называют задачей декодирования. В нашем случае исследуется следующая функция ошибки:

$$\text{MAPE}(\mathbf{y}_i, \hat{f}(\mathbf{x}_i)) = \left\| \frac{\mathbf{y}_i - \hat{f}(\mathbf{x}_i)}{\mathbf{y}_i} \right\|_1,$$

Шлюзовая функция

Пусть имеются модели f_1, \dots, f_k . Для каждого объекта x определяется правдоподобие $\pi_k(x) \rightarrow [0, 1]$ i -й модели на нем.

$$\pi_k(x, V) = \sigma(g(x, \omega), V) = \frac{\exp v_k^T g(x, \omega)}{\sum_{i=1}^k \exp v_i^T g(x, \omega)}$$

Здесь $V = [v_1, \dots, v_k, \omega]$, σ - softmax, $g(x, \omega)$ - преобразование над x .

Смесь экспертов

Получив π_i , можно построить модель, задаваемую формулой

$$f_m(x) = \sum_{i=1}^k \pi_i(x, V) f_i(x)$$

Привилегированное обучение

Для некоторых объектов x доступна *привилегированная* информация x^* . Введем функции ученика $f_s \in \mathcal{F}_s$ (student) и учителя $f_t \in \mathcal{F}_t$ (teacher):

$$f_s : x \rightarrow y, f_t : x, x^* \rightarrow y$$

Дистилляция

Модель учителя сложнее, чем ученика, нет привилегированной информации. Обучение:

$$f_t = \arg \min_{f \in \mathcal{F}_t} \frac{1}{n} S(y_i, f(x_i)) + \Omega(\|f\|),$$

$$f_s = \lim \arg \min_{f \in \mathcal{F}_s} \frac{1}{n} \sum_{i=1}^n [(1 - \lambda) S(y_i, f(x_i)) + \lambda S(s_i, f(x_i))],$$

Обобщенная дистилляция

Учитель обучается на привилегированном наборе данных x^* , сложность модели учителя не ограничена. Затем предсказания учителя используются для обучения ученика.

Общий алгоритм выглядит примерно так:

- 1) Выбрать параметр имитации λ .
- 2) Выделить подмножество объектов, обладающих привилегированным описанием и найти оптимальную функцию учителя f_t
- 3) Используя функцию учителя f_t , построить сглаженные предсказания для всех объектов обучающей выборки.
- 4) Найти оптимальную функцию ученика.

Набор данных

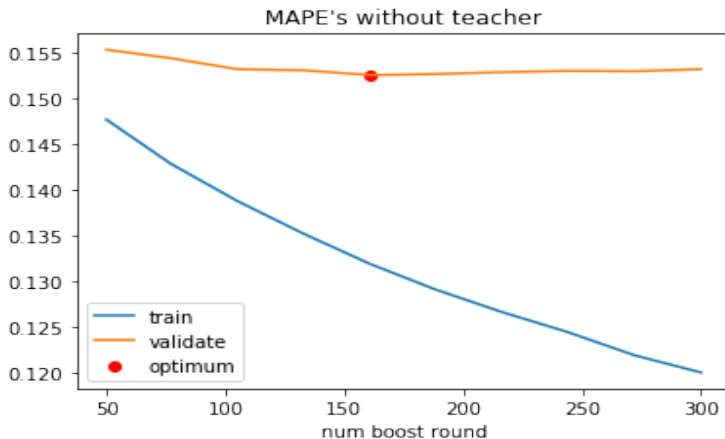
Решается задача определения стоимости недвижимости по её координатам и площади. В качестве привилегированной информации на тренировочной выборке доступна оценка качества архитектуры и состояния здания экспертами, и некоторые другие признаки.

Базовый алгоритм

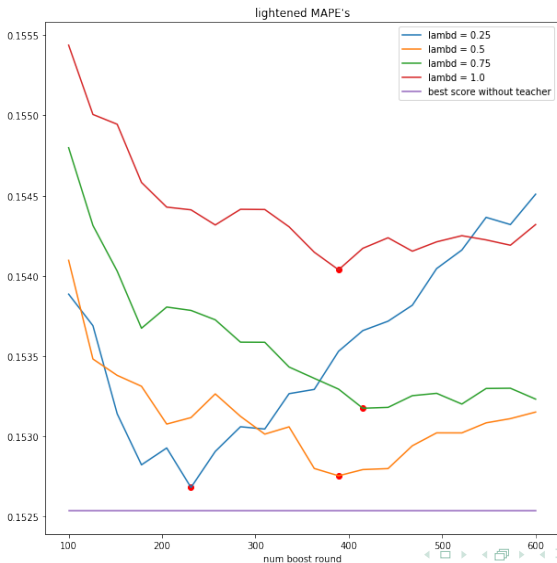
В качестве базового алгоритма используется градиентный бустинг. Единственный параметр, изменяемый при проведении эксперимента - число итераций алгоритма, служит показателем сложности алгоритма.

Результаты, полученные без привилегированного обучения

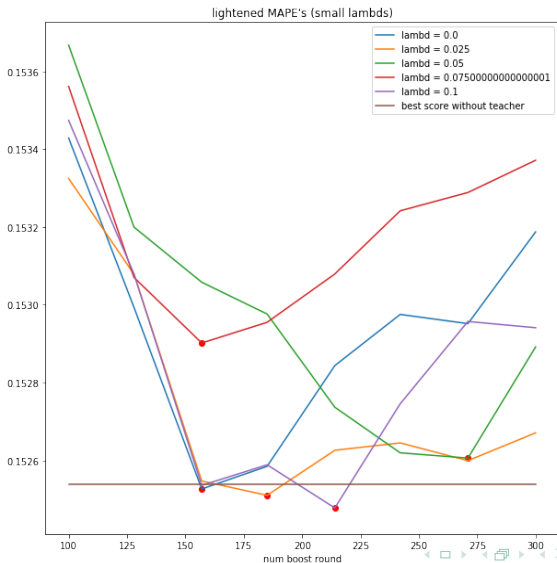
Без использования привилегированной информации MAPE
= 0.1525



Использование привилегированной информации для смеси экспертов



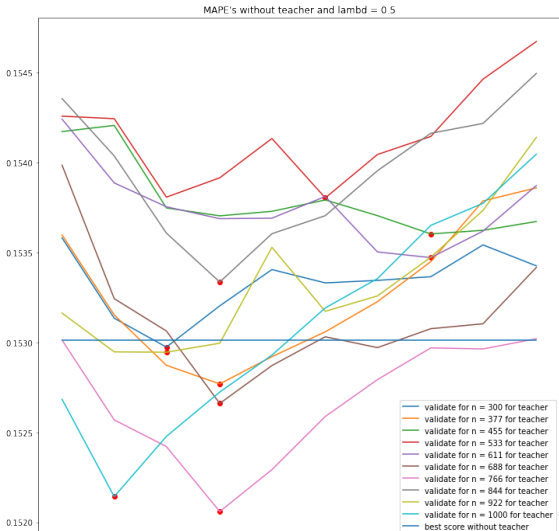
Использование привилегированной информации для смеси экспертов



Было взято несколько моделей, с различными степенями имитации, и числом итераций, оптимальным для них. В итоге алгоритм медленно сходится и дает большую ошибку: $MAPE = 19.6$ (против 15.3 изначально). Веса показывают, что при большом модели с большим параметром имитации отбрасываются.

Дополнительный эксперимент

Если брать не оптимальные модели учителя, а переобученные - качество оценки улучшается, но не значимо.



Основной эксперимент:

Модели, построенные при помощи дистилляции, не только имеют худшее качество оценки, но ещё и требуют большего числа ресурсов. Смесь экспертов, построенная на моделях с различным параметром дистилляции, также работает несколько хуже. Использование привилегированной информации не привело к положительному результату.

Дополнительно:

Если в алгоритме дистилляции использовать не оптимальные параметры учителя, а параметры, дающие небольшое переобучение, результат становится лучше - но не значимо.

Выводы:

- Показан пример задачи машинного обучения, для которой дистилляция увеличивает вычислительную сложность алгоритма, но при этом качество оценки не улучшается.
- Зависимость качества оценки от параметра дистилляции немонотонно, и имеет локальные минимумы, которые находятся только экспериментальным путем.
- Использование переобученного учителя в алгоритме дистилляции улучшает и качество оценки, и сложность модели
- Смесь экспертов - вычислительно сложный алгоритм, сходимость сильно от начальной инициализации.