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 избыточную вычислительную сложность, от параметров их иницилизации.

Метод решения

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

Литература

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. C. 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.Schlolkopf, V.Vapnik, D.Lopez-Paz, L.Bottou. ICLR, 2016. - Обобщение подходов Вапника и Хинтона к привилегированному обучению.

Общая задача

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

Задача декодирования

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

$$\mathrm{MAPE}(y_i, \hat{f}(x_i)) = ||\frac{y_i - \hat{f}(x_i)}{y_i}||_1,$$

Постановка задачи

Шлюзовая функция

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

$$\pi_k(x, V) = \sigma(g(x, \omega), V) = \frac{expv_k^T g(x, \omega)}{\sum\limits_{i=1}^{k} expv_k' g(x, \omega)}$$

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

Смесь экспертов

Получив π_i, можно построить модель, задаваемую формулой

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

Использование смеси экспертов

Способность мультимодели к фильтрации

В качестве $\pi(x, V)$ Была использована нейросеть с 50 нейронами для достижения примерно того же качества, как у простой нейросети с 50 - ю нейронами. Также выяснилось, что смесь экспертов выявляет незначимые модели f_{weak} ($\pi_{weak}(x, V)$), и при удалении этих моделей качество оценки повышается.

Зависимость от изначальной иницилизации моделей

Если в том же эксперименте параметры "простых"моделей подбирать случайно, только в около 10% случаев метод вообще сходится. Значит, требуется как-то использовать дополнительную информацию.

Постановка задачи

Привилегированное обучение

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

$$\mathbf{f}_s : \mathbf{x} \to \mathbf{y}, \mathbf{f}_t : \mathbf{x}, \mathbf{x}^* \to \mathbf{y}$$

Дистилляция

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

$$\mathbf{f}_t = \arg\min_{\mathbf{f} \in \mathcal{F}_t} \frac{1}{n} S(\mathbf{y}_i, \mathbf{f}(\mathbf{x}_i)) + \Omega(||\mathbf{f}||),$$

$$\mathbf{f}_s = \lim \arg\min_{\mathbf{f} \in \mathcal{F}_s} \frac{1}{n} \sum_{i=1}^{n} \left[(1-\lambda) l(\mathbf{y}_i, \boldsymbol{\sigma}(\mathbf{f}(\mathbf{x}_i))) + \lambda l(\mathbf{s}_i, \boldsymbol{\sigma}(\mathbf{f}(\mathbf{x}_i)))\right],$$

Цели эксперимента

Обобщенная дистилляция

Учитель обучается на привилегированном наборе данных $\mathbf{x}^*$, 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 $\mathbf{f}_t$
3) Используя функцию учителя $\mathbf{f}_t$, построить сглаженные предсказания для всех объектов обучающей выборки.
4) Найти оптимальную функцию ученика.

Цели эксперимента

Набор данных

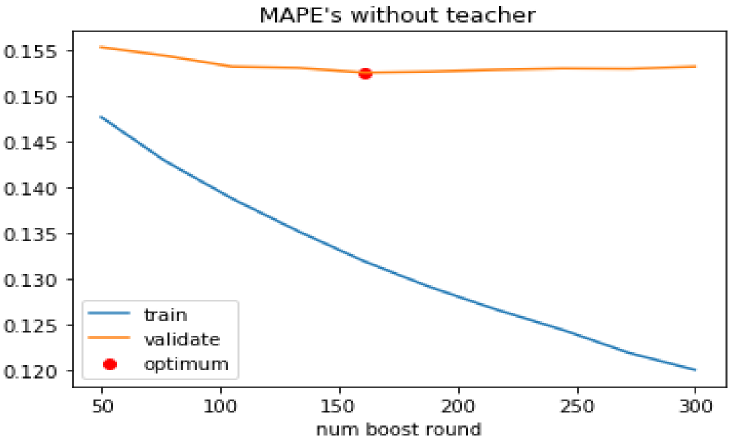
Решается задача определения стоимости недвижимости по её координатам и площади. В качестве привилегированной информации на тренировочной выборке доступна оценка качества архитектуры и состояния здания экспертами, и некоторые другие признаки.

Базовый алгоритм

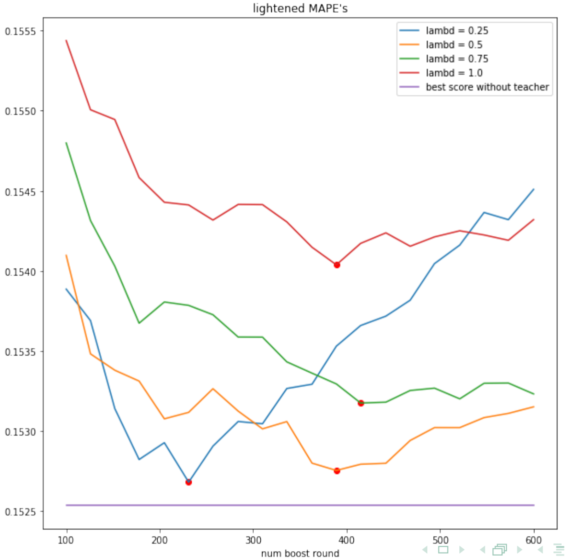
В качестве базового алгоритма используется градиентный бустинг. Единственный параметр, изменяемый при проведении эксперимента - число итераций алгоритма, служит показателем сложности алгоритма.

Результаты, полученные без привилегированного обучения

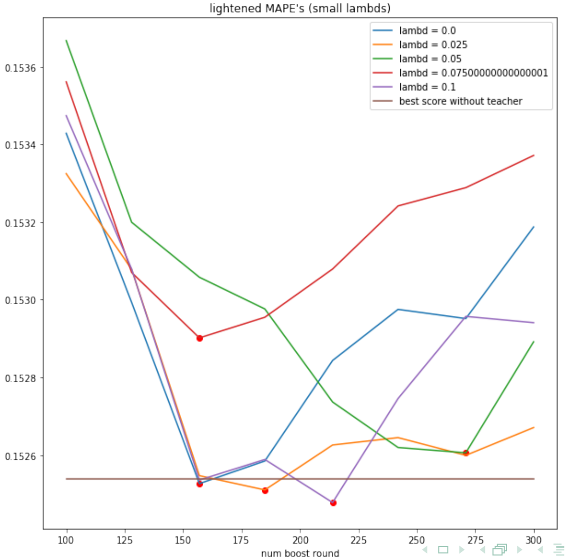
Без использования привилегированной информации MAPE $= 0.1525$



Использование привилегированной информации для смеси экспертов



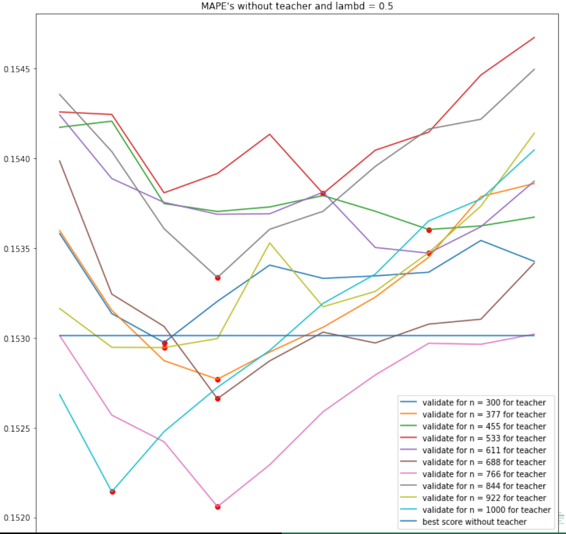
Использование привилегированной информации для смеси экспертов



Построение 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: $\mathrm{MAPE} = 19.6$ (против 15.3 изначально). Веса показывают, что при большом модели с большим параметром имитации отбрасываются.

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



Результаты эксперимента

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.

Заключение

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 иницилизации.