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

Цели исследования

Цель работы

Создать метод построения моделей оптимальной сложности для задач обучения с учителем.

Проблема

Большая часть алгоритмов, имеющих небольшую вычислительную мощность, имеют сильную зависимость сходимости от начальных параметров.

Метод решения

Во-первых, использовать априорную информацию, производимую более сложной моделью - учителем. Во-вторых, научиться грамотно строить композицию 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. 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.Schlolkopf, V.Vapnik, D.Lopez-Paz, L.Bottou. ICLR, 2016. - Обобщение подходов Вапника и Хинтона к привилегированному обучению.

Постановка задачи

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 многоклассовой классификации.

Когда $y_i = [y_i^1, ..., y_i^r]$, при этом $\forall k : 0 \leq y_i^k \leq 1$, $\sum_{i=1}^{r} y_i = 1$, задача называется задачей классификации на k классов. Функцией ошибки мы выберем кросс-энтропию:

$$S(y_i, \hat{f}(x_i)) = -\sum_{i=1}^{r} y_i^k \log \sigma(\hat{f}(x_i)^k), \sigma(\hat{y})^k = \frac{\exp y^k}{\sum_{k'=1}^{r} \exp y^{k'}}$$

Постановка задачи

Задача декодирования

Если матрица ответов состоит из действительных векторов $y_i \in \mathbb{R}^r$, то задачам декодирования соответствуют следующие функции ошибки:

$$MAE(y_i, \hat{f}(x_i)) = ||y_i - \hat{f}(x_i)||_1,$$

$$MSE(y_i, \hat{f}(x_i)) = ||y_i - \hat{f}(x_i)||_2,$$

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

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.

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

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

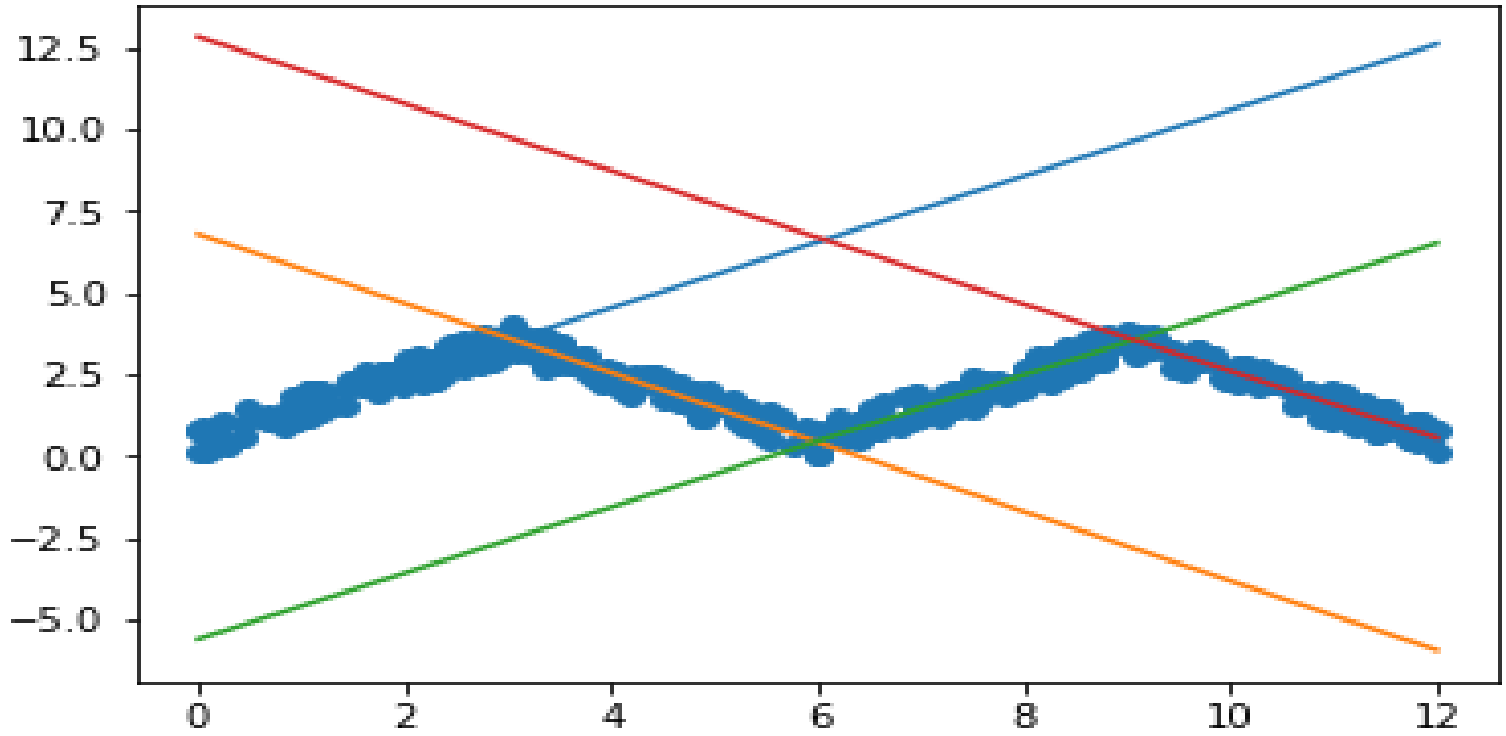
Смесь экспертов

Получив π_i, можно построить модель, задаваемую формулой

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

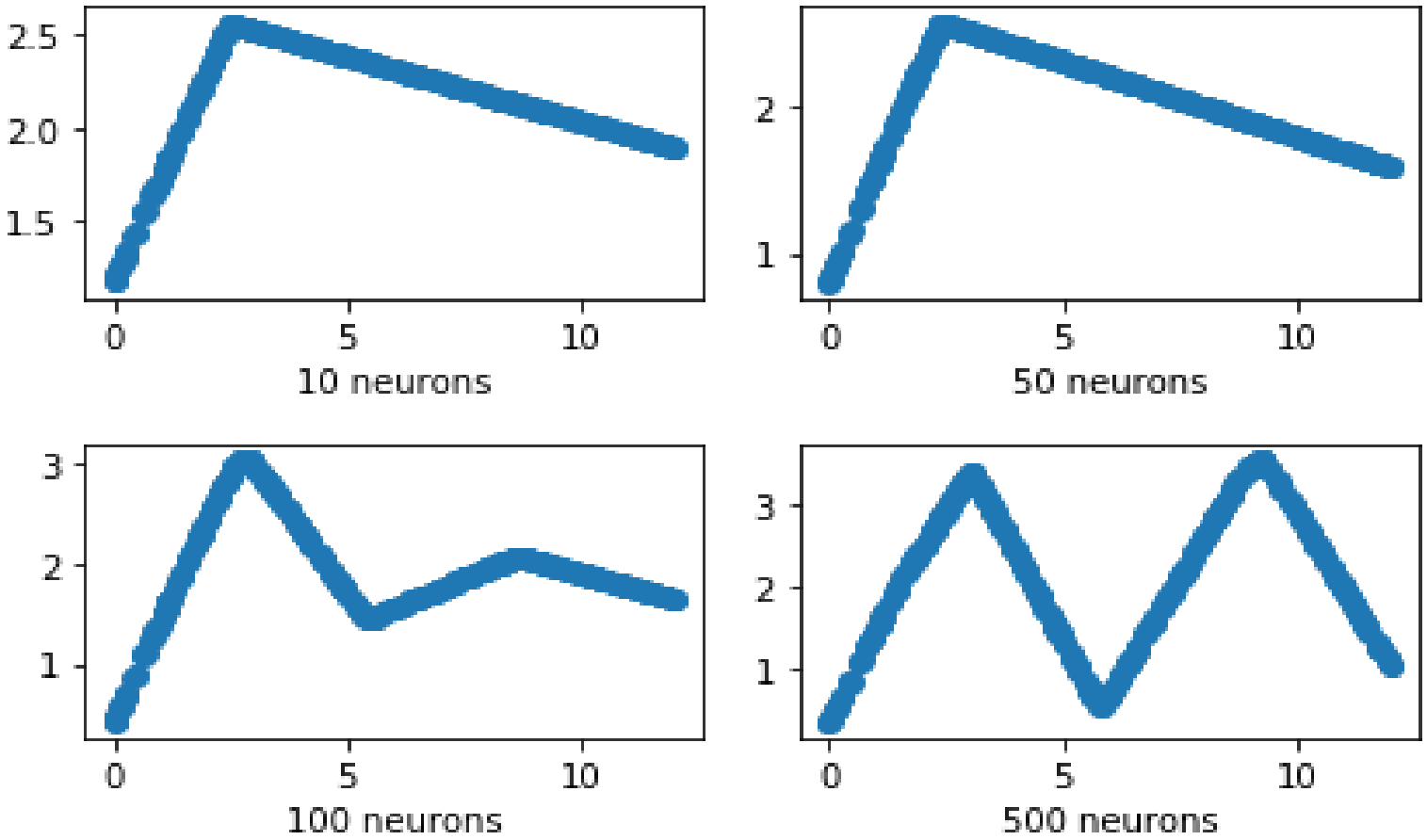
Цели эксперимента

Рассмотрим задачу аппроксимации кусочно - гладкой функции при помощи смеси экспертов, построенной на нескольких линейных моделях. Для начала рассмотрим случай, когда исходные модели описывают тот или иной участок, где наблюдается линейность.



Тривиальные модели

Классические нейронные сети приближают функцию только при достаточно большом количестве нейронов в слое - порядка 500.



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 электронами. Также выяснилось, что смесь экспертов выявляет незначимые модели f_{weak} $(\pi_{weak}(x, V))$, и при удалении этих моделей качество оценки повышается.

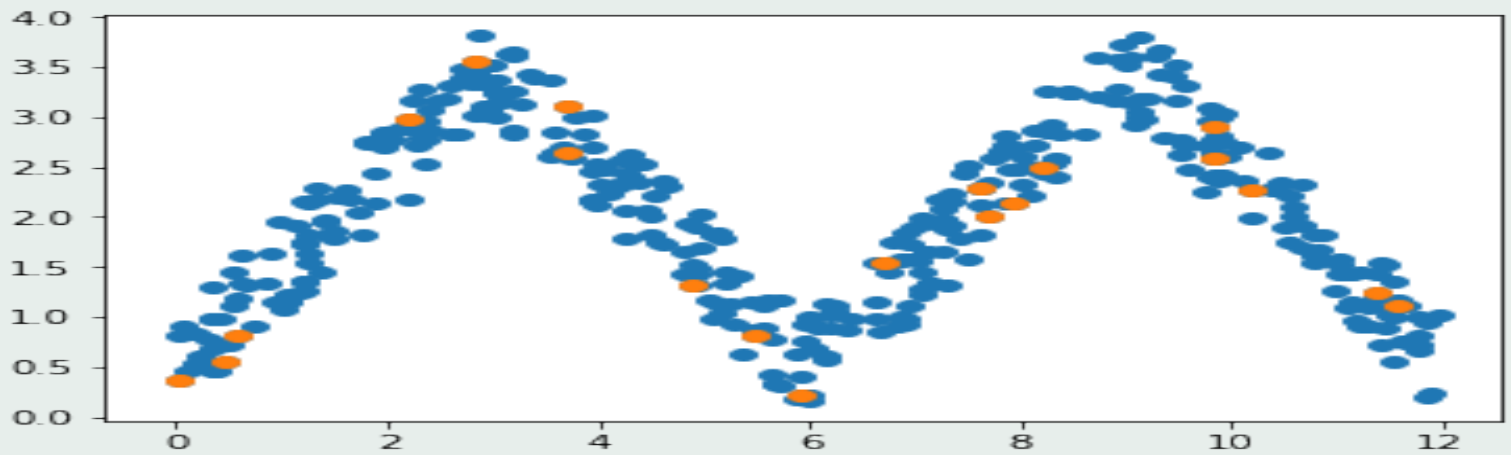
Зависимость от изначальной инициализации моделей

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.

Использование привилегированной информации для смеси экспертов

Решение проблемы слабой сходимости

Мы используем априорную информацию о принадлежности небольшого числа объектов конкретным экспертам, и именно на них обучаем простые модели.



Результат

В результате сходимость эксперимента увеличилась примерно до 70%.

Результаты эксперимента

Была построена модель, которая довольно сильно снижает вычислительную сложность алгоритма (в нашем случае сложность алгоритма считается зависящей от числа нейронов в сети). При этом было привлечено минимальное количество привилегированной информации.

Заключение

Выводы:

- При достаточно хорошо подобранных начальных параметрах исходных моделей использование мультимоделирования помогает весьма сильно оптимизировать сложность модели при той же точности
- В случае же, когда исходные модели не очень хорошо настроены, метод может просто не сойтись.
- При использовании некоторой априорной информации, полученной отдельными экспертами, качество оценки и эффективность модели сильно увеличилась по сравнению с тем случаем, когда исходные модели инициализировались протизвольно.