Я.А. СЕРЕДА

Нижегородский государственный университет им. Н.И. Лобачевского
sereda@itmm.unn.ru

К МАТЕМАТИЧЕСКОЙ ПОСТАНОВКЕ ЗАДАЧИ ВЫБОРА АРХИТЕКТУРЫ СЕТИ [1*]

Предложен подход к проблеме выбора архитектуры глубокой сети, опирающийся на некоторые наработки[1] в статистической физике.

An approach is proposed to the problem of choosing a deep network architecture, based on some developments [1] in statistical physics.

Ключевые слова: принцип максимума энтропии, метод множителей Лагранжа, активное обучение, глубокие сети

В вероятностной постановке задачи машинного обучения обычно полагается, что генеральная совокупность наблюдаемой (многомерной) случайной величины X порождена неизвестной детерминированной функцией F(Z)=X, где Z – векторная скрытая (от измерения) переменная, имеющая неустранимо случайную природу. Размерность наблюдаемой с.в. X известна; размерность же с.в. Z может быть (и, скорее всего, является) бесконечной. Моделирование явления нейросетью имеет смысл тогда, когда в данных существует структура низкой размерности. Т.е. небольшое число «эффективных» компонент Z объясняют почти всю вариацию X. Остальные скрытые компоненты влияют мало, и так, что их совокупное влияние на X можно смоделировать аддитивным шумом (например, белым Гауссовским, предпосылку к чему дает ЦПТ).

В случае произвольной, несинтетической задачи машинного обучения заранее не известно ни число эффективных скрытых переменных, ни (тем более!) параметрическое семейство, из которого происходит F. Однако, state-of-the-art алгоритмы глубокого обучения, применяемые сегодня промышленно, предполагают однократное задание фиксированной структуры сети. Получается, что и размерность вектора скрытых

[1*] Данная работа выполнена при поддержке Минобрнауки РФ, проект № 14.Y26.31.0022.

переменных, и параметрическое семейство для неизвестной функции задаются эвристически и негибко. В результате даже успешная (в смысле выбранных метрик качества) сеть всегда избыточно параметризована[2]. Управлять специализацией её подсетей становится проблематично, и при таком подходе модель имеет проблемы с интерпретируемостью для человека.

Из описанной проблемы естественно возникает задача - модифицировать обучение глубоких сетей так, чтобы новые скрытые переменные добавлялись в глубокую сеть лишь по мере необходимости и на нужный уровень глубины.

В стат.физике уже рассматривалась задача о том, как получить наиболее информативное распределение для ненаблюдаемой случайной величины. Ненаблюдаемой (скрытой) случайной величиной назовем ту с.в., для которой не существует измерительного прибора. В качестве примера работы с такими величинами в физике можно рассмотреть задачу[1] поиска наиболее информативного распределения вероятностей по решетке кристалла: с какой вероятностью в каком узле содержится примесь? Поскольку эта с.в. не поддается прямому измерению, частотные прикидки ее распределения сделать нельзя. С другой стороны, можно разработать некоторые измеримые воздействия на кристалл, которые будут нести нечастотную информацию об интересующей нас скрытой случайной величине.

Например, можно пропускать через кристалл излучение, и мерить характеристики его рассеяния. Очевидно, местоположение примеси влияет на результаты подобных экспериментов, и, следовательно, результаты экспериментов несут информацию о местоположении примеси.

Чтобы найти распределение вероятностей для примеси, записывается функционал энтропии (относительно искомого распределения) и ищется его условный максимум. В качестве условий записываются результаты экспериментов. Эксперимент планируется так, чтобы его результат был измеримой функцией неизвестной с.в.. Это дает возможность составить ограничения типа «равенство» и решать задачу оптимизации с ограничениями.

Экстремальное распределение, полученное в ходе решения этой задачи, удовлетворяет требованию Лапласа: если нет необходимости считать одно значение с.в. более вероятным, чем другое, им назначена равная вероятность. Если энтропия экстремального распределения,

описанного выше, оказалась высока, значит ограничения (полученные из экспериментов) неинформативны. Иными словами, из меры неопределенности экстремального распределения можно делать выводы об информативности экспериментов.

В данной работе был предложен эскиз алгоритма добавления нового нейрона в глубокую сеть, основанный на решении задачи оптимизации энтропии в так называемой задаче «тензорного дополнения». Решение о необходимости добавления нового нейрона принимается на основе решения (несколько модифицированной) задачи условной оптимизации энтропии. А решение о том, с какими весами и куда его добавлять, предлагается принимать в ходе решения *обратной* задачи: оптимизируемый функционал известен, и местоположение экстремума тоже - нужно восстановить ограничения, которые сдвинули экстремум в наблюдаемую точку.

Глубокая сеть в предлагаемом подходе представляет собой набор измеримых функций от наблюдаемой векторной с.в. X. При наблюдении выборки(батча) данных каждая ф-ция принимают определенные значения, что может быть интерпретировано (по аналогии с физическим примером выше) как эксперименты с измеримым исходом, результаты которых можно подставлять в задачу оптимизации в виде ограничений. Задача решается методом множителей Лагранжа. Значения множителей, полученные в ходе решения показывают «значимость» соответствующих им активных ограничений. Это позволяет сформулировать другую задачу: из имеющихся ограничений выбрать небольшое подмножество «определяющих» ограничений (такое, что отбрасывание всех остальных ограничений не оказывает значимого влияния на вид экстремального распределения).

Если определяющих ограничений не найдено, значит сеть не может выдать *простой* выскоуровневой интерпретации наблюдаемого явления, только сложную – задействующую очень много ограничений, и оттого неинтерпретируемую. Если же такой набор ограничений нашелся – значит, ничего добавлять в сеть не нужно. Алгоритм добавления нейрона опирается на то, что каждый нейрон индуцирует разбиение мн-ва всех возможных наблюдений в «дополняемой» области на классы эквивалентности. Причем, чем в более глубоких слоях находится нейрон, тем более крупное разбиение с ним связано. Поэтому обратная задача решается при условии максимизации глубины слоя.

Список литературы

1. Jaynes E. T. Prior probabilities //IEEE Transactions on systems science and cybernetics. – 1968. – Т. 4. – №. 3. – С. 227-241..
2. Molchanov D., Ashukha A., Vetrov D. Variational dropout sparsifies deep neural networks //Proceedings of the 34th International Conference on Machine Learning-Volume 70. – JMLR. org, 2017. – С. 2498-2507..