АНАЛИЗ ПРОЦЕССА ПРИНЯТИЯ РЕШЕНИЯ В МОДЕЛИ НЕЙРОСЕТЕВОЙ КЛАССИФИКАЦИИ МЕТОДОМ ЛОКАЛЬНЫХ ЛИНЕЙНЫХ АППРОКСИМАЦИЙ

Курочкин А. В., Садов В. С.
Кафедра интеллектуальных систем, Белорусский государственный университет, Факультет радиофизики и компьютерных технологий
Минск, Республика Беларусь
E-mail: alex.v.kurochkin@gmail.com, sadov@bsu.by

Одним из существенных ограничений нейросетевых моделей для классификации является тот факт, что процесс принятия решений в обученной нейронной сети очень сложно поддаётся анализу, из-за чего невозможно оценить смысловую корректность результатов и сформировать на их основе новые экспертные знания. Для решения этой проблемы предлагается использовать метод локальных линейных аппроксимаций обученной модели, позволяющий оценить влияние отдельных признаков одного экземпляра данных на полученный результат в понятной форме.

Введение

Искуственные нейронные сети прямого распространения на сегодняшний день являются одним из самых распространенных инструментов для построения моделей классификации на основе известных данных (обучающей выборки). Относительная простота конфигурации, обширный программный инструментарий и возможность тонкой настройки модели делают нейросетевую классификацию очень эффективным средством выведения зависимостей на основании произвольных данных в табличном виде. При наличии достаточно репрезентативной обучающей выборки алгоритм обучения нейронной сети может подобрать параметры моделей таким образом, чтобы сформировать разнообразные совокупности линейных и нелинейных комбинаций входных признаков, которые в последующем используются для получения итогового результата.

С точки зрения систем поддержки принятия решений нейронные сети часто противопоставляются экспертным системам на базе логического вывода (например, нечеткой логики). В таких экспертных системах процесс принятия решения описывается формально, на основании некоторого описательного представления экспертных знаний, и не учитывает статистические характеристики существующих данных. В то же время, нейросетевые модели не включают в себя представление экспертных знаний, а вместо этого выстраивают собственный процесс принятия решений путём подбора параметров модели в процессе обучения таким образом, чтобы минимизировать ошибку модели на некотором объеме существующих данных. С одной стороны, это позволяет устанавливать сложные зависимости в некотором объеме данных без необходимости их детального смыслового анализа. С другой стороны, обученная нейросетевая модель является «черным ящиком» – процесс принятия решений описывается комбинацией параметров модели и, как правило, крайне сложен для интерпретации [1].

I. Анализ обученных нейросетевых классификаторов

Для интерпретации обученных нейросетевых моделей в первую очередь необходимо определить, какое представление процесса принятия решения может быть интуитивно понятно для восприятия человеком.

Нейросетевая модель M для решения задачи бинарной классификации по совокупности из n вещественных признаков представляет собой параметрическую функцию:

$$M(\vec{x}, \vec{\theta}) : (\mathbb{R}^n \times \Theta) \to [0; 1], \qquad (1)$$

где $\vec{x} \in \mathbb{R}^n$ – n-мерный вещественный вектор значений входных признаков, $\vec{\theta} \in \Theta$ – вектор параметров модели, $\mathbb{R}$ – множество вещественных чисел, Θ – множество значений параметров модели. Как правило $\Theta \subset \mathbb{R}^m$, где m – количество параметров модели [1].

Параметрами модели в нейронной сети прямого распространения являются веса связей между формальными нейронами соседних слоёв. Каждый нейрон реализует некоторую функцию активации, на вход которой подаётся взвешенная сумма выходов нейронов предыдущих слоёв. Увеличение количества нейронов и количества слоёв позволяет нейронной сети оперировать более сложными совокупностями признаков. Интуитивно ожидается, что в процессе обучения нейронная сеть сможет выделить такие совокупности значений признаков, которые совместно предоставляют более сложное абстрактное описание входных данных. Например, в задаче распознавания символов печатного шрифта при использовании значений пикселей как входных параметров один из нейронов второго слоя

может активизироваться по совокупности значений горизонтальных пикселей, что соответствует абстрактному описанию «горизонтальная черта в центре»; в дальнейшем, такой нейрон может иметь значительный вес при связи с выходными нейронами, соответствующими буквам «A», «B», «E», «F», «H», «P», «R». В то же время, для принятия решения могут использоваться более сложные нелинейные совокупности признаков, которые не будут иметь простой интерпретации [2].

II. Построение интерпретаций на основе линейных аппроксимаций

В процессе анализа установлено, что наиболее универсальными и интуитивно понятными являются интерпретации процесса принятия решения на основе порогового значения одного из признаков, т.е. в виде линейной модели по одному из входных значений [3]. Таким образом, для некоторой совокупности входных признаков $\vec{x^*} \in \mathbb{R}^n$ необходимо построить набор из n взвешенных описаний, каждое из которых будет характеризовать линейное локальное поведение нейросетевой модели в некоторой окрестности совокупности $\vec{x^*}$ по каждому из входных признаков.

Для построения таких описаний для элемента $\vec{x^*}$ осуществляется анализ пространственной окрестности полученной модели на некотором расстоянии. Для этой окрестности осуществляется сэмплирование – выборочная генерация некоторого количества точек, которые выступают в качестве небольшой локальной обучающей выборки. Для этой выборки методом линейной регрессии строится линейная модель $m[\vec{x^*}, \vec{\theta^*}] : (\mathbb{R}^n) \to [0; 1]$. Параметры этой модели $\vec{\theta^*}$ позволяют в n-мерном пространстве входных признаков построить гиперплоскость, которая разделяет исходное пространство на 2 части. При этом для каждого из n признаков можно рассмотреть проекцию этой гиперплоскости на ось, соответствующую этому признаку. Точка пересечения и угол наклона определяют, соответственно, пороговое значение и его влияние на конечный результат. Если гиперплоскость перпендикулярна оси признака, то этот признак оказывает определяющее влияние, поскольку принятие решения осуществляется только на основании его порогового значения. Если гиперплоскость параллельна оси признака или проходит под очень малым углом наклона, влияние этого признака на результат минимально. Таким образом, для определения влияния по углу наклона гиперплоскости относительно оси может использоваться косинус.

Другими словами, для совокупности из значений n признаков $\vec{x^*} = \{x_i^*\}$ и некоторой обученной классификационной нейросетевой модели $y = M(\vec{x}), y \in [0; 1]$ линейные аппроксимации строятся следующим образом: по окрестности точки $\vec{x^*}$ в соответствии с выходом модели строится линейная модель, задающая некоторую гиперплоскость T^*. При проекции гиперплоскости T^* на оси признаков x_i^* можно определить точки пересечения $\overline{x_i^*}$, а также углы наклона $\overline{\varphi_i^*}$.

Полученные значения могут быть интерпретированы следующим образом: косинус угла наклона $\cos\overline{\varphi_i^*}$ показывает, насколько на полученный выходной результат $y^* = M(\vec{x^*})$ влияет тот факт, что i-й признак x_i принял значение, большее или меньшее, чем пороговое $\overline{x_i^*}$. Таким образом, для любого входного значения и любого признака может быть получен набор линейных аппроксимаций, которые могут использоваться для оценки влияния конкретного значения признака на выходной результат. Кроме того, пороговое значение $\overline{x_i^*}$ для этого признака отражает, что анализируемая модель будет иметь схожее поведение в ближайшей окрестности при изменении значения i-го признака в пределах от x_i^* до $\overline{x_i^*}$.

Полученные линейные аппроксимации позволяют по линеаризации модели в окрестности исследуемой точки понять, какие именно признаки оказали наибольшее влияние на полученное значение.

Заключение

В работе представлен метод линейных аппроксимаций, позволяющий анализировать поведение сложных нейросетевых классификационных моделей. Хотя такой подход не предоставляет полное описание процесса принятия решения, проводимый анализ позволяет, для некоторой совокупности входных признаков, получить более интуитивное представление о том, какие из этих признаков являлись определяющими в вычислении итогового значения по этой модели. Полученные результаты могут использоваться для формирования новых экспертных знаний, проверки корректности модели в граничных случаях, а также для экспертной оценки состоятельности полученной модели.

Список литературы

1. Goodfellow, I. Deep learning / I. Goodfellow, B. Yoshua, C. Aaron. – Cambridge, MT: The MIT Press, 2016. –775 p.
2. Ioannou, Y. Decision Forests, Convolutional Networks and the Models in-Between / Y. Ioannou, D. Robertson, D. Zikic, et. al. // arXiv:1603.01250[cs] [Электронный ресурс]. – 2016. – Режим доступа: https://arxiv.org/abs/1603.01250 – Дата доступа: 08.10.2019.
3. Lundberg, S. M. A Unified Approach to Interpreting Model Predictions / S. M. Lundberg, S.-I. Lee // Advances in Neural Information Processing Systems. – 2017. – № 30. – pp. 4765–4774.