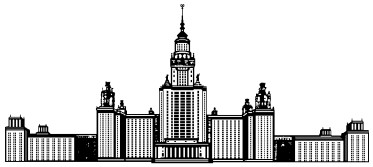
Московский государственный университет имени М. В. Ломоносова



Факультет Вычислительной Математики и Кибернетики
Кафедра Математических Методов Прогнозирования

ДИПЛОМНАЯ РАБОТА СТУДЕНТА 417 ГРУППЫ

″Система визуального анализа для активности головного мозга″

Выполнил:
студент 4 курса 417 группы
Тыцкий Владислав Игоревич

Научный руководитель:
к.ф-м.н., доцент
Майсурадзе Арчил Ивериевич

Москва, 2021

Аннотация

–В последнюю очередь–

Содержание

1 Введение

– В последнюю очередь –

2 Методы интерпретации моделей машинного обучения

Методы интерпретации моделей машинного обучения можно поделить на:

- Восприятие объекта:
  - Локальные – объясняют поведение модели для конкретного объекта.
  - Глобальные – объясняют работу модели в целом на всей выборке.
- Восприятие модели:
  - Агностичные к модели (Model Agnostic) – алгоритм интерпретации не опирается на внутреннюю структура модели. Все, что видит агностичная модель – это пару "вопрос" "ответ" или в терминах машинного обучения (x_i, y_i).
  - Специфичные для модели (Model Specific) – учитывая специфику изучаемой модели, мы можем придумать алгоритм, работающий лучше(качественней, быстрее) чем агностичные методы интерпретации.

2.1 Специфичные для модели (Model Specific)

Как упоминалось выше, методы интерпретации специфичные для модели опираются на ее внутреннюю структуру. Одним из простейших методов интерпретации для линейной модели может послужить рассмотрение весов для каждого признака – мы считаем, что чем больше(по модулю) вес у признака, тем более важным он является для предсказания. Более того, если вес, соответствующий признаку имеет положительное/отрицательное значение, можно говорить о положительной/отрицательной "корреляции" между этим признаком x_i и целевым значением y_i.

$$y_i = f(< x_i, w >), \text{где } f(x) = x \text{ или } f(x) = sign(x)$$

Важно отметить, что все признаки должны иметь одинаковое матожидание и дисперсию. Иначе веса для каждого признака могут зависеть от абсолютных значений признака, а не от его важности для модели. Можно сказать, что линейная модель интерпретируема сама по себе, мы считаем, что принцип ее работы настолько прост, что сама внутренняя структура модели (веса признаков) может объяснить ее поведение.

Достоинства линейных моделей:

- Вес модели в значительной мере определяет важность признака и этим можно пользоваться в качестве интерпретации
- Хорошо изучены и "понимаемы" сообществом
- Использование как "суррогатной" модели [1]

Недостатки линейных моделей:

- Обобщающей способности далеко не всегда достаточно для получения модели с надлежащим качеством
- Зависимость весов от других факторов (например от абсолютных значений признака)

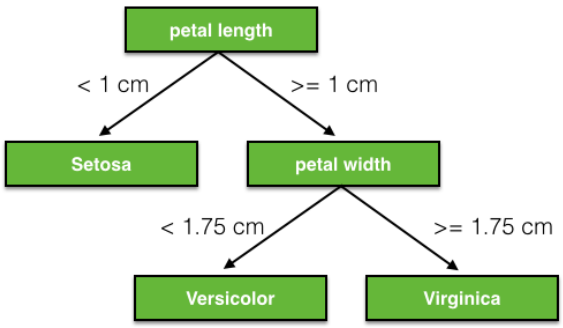


Рис. 1: Принцип работы решающего дерева

Еще одним примером интерпретации специфичной для модели может послужить интерпретация решающего дерева. Решающее правило f, по которому для x_i сопоставляется y_i весьма близко к тому, как рассуждает человек, когда принимает какие-либо решения. На каждом этапе мы рассматриваем конкретный признак и исходя из значения этого признака для объекта "спускаемся" в соответствующий узел.
Конкретный метод построения решающего дерева обычно определяется:[2]

- Видом предикатов в вершинах.
- Критерием информативности Q.
- Критерием останова.

[1] Далее мы приведем пример суррогатной модели (LIME)

[2] Мы не будем подробно останавливаться на особенностях построения деревьев решений

Достоинства деревьев решений:

- Разделяет данные на отдельные группы, что проще воспринимать, чем гиперплоскости больших размерностей
- Естественная визуализация метода
- Решающее правило сходно с тем, как рассуждает человек

Недостатки деревьев решений:

- Плохо справляются с линейными зависимостями
- Легко переобучаются(дерево, где в каждом листе один объект)
- Чем больше дерево, тем сложнее его интерпретировать

2.2 Агностичные модели (Model Agnostic)

Большим преимуществом методов интерпретации, не зависящих от модели, по сравнению с методами интерпретации, специфичными для конкретной модели, является их гибкость. Методы интерпретации могут быть применены к любой модели, поэтому исследователи могут свободно использовать любую модель, которая им нравится. Все, что основывается на интерпретации модели машинного обучения, например графический или пользовательский интерфейс, также становится независимым от базовой модели машинного обучения. Как правило, для решения задачи оценивается не один, а множество типов моделей машинного обучения, и при сравнении моделей с точки зрения интерпретируемости легче работать с объяснениями, не зависящими от модели, поскольку один и тот же метод может быть использован для любого типа модели.

Желаемые свойства агностического метода интерпретации:

- Гибкость модели – метод интерпретации может работать с любой моделью машинного обучения, такой как бустинги, случайные леса,и нейронные сети.
- Гибкость объяснения – мы не ограничены определенной формой объяснения. В некоторых случаях может быть полезно взглянуть на важности всех признаков, а в других выделить область изображения на картинке, являющийся причиной именно такого ответа.
- Гибкость представления – система должна иметь возможность работать с различным представлением объектов, будь то текст, картинка или табличные данные.

2.3 LIME

Локальные суррогатные модели(Local surrogate models) – это модели, которые можно использовать для объяснения отдельных прогнозов других моделей *"черных ящиков"*.[3]

В статье "Why I should trust you?"[1] – авторы предлагают конкретную реализацию локальных суррогатных моделей. Суррогатные модели обучаются для аппроксимации прогнозов "черного ящика". Цель локальных суррогатных моделей – понять, почему "черный ящик" сделал определенный прогноз на конкретном объекте. LIME изучает, что происходит с прогнозами, когда мы некоторым образом меняем объекты из обучающей выборки. LIME генерирует новый набор данных, состоящий из "возмущенных" выборок и соответствующих им прогнозов. На этом новом наборе данных обучается интерпретируемая модель, которая взвешивает объекты пропорционально близости к объекту, для которого мы объяснить поведение модели. Интерпретируемой моделью может быть что угодно, например линейная модель или дерево решений.

Математически локальные суррогатные модели выражаются следующим образом:

$$\text{explanation}(x) = \underset{g \in G}{\operatorname{argmin}} L(f, g, \pi_x) + \Omega(g)$$

$$L(f, g, \pi_x) = \sum_{x'} (f(x') - g(x'))^2 \pi_x(x, x')$$

где g (например, линейная модель), минимизирующая функцию потерь L (например MSE), которая измеряет, насколько близки прогнозы g к прогнозу исходной модели - "черному ящику" f, а $|Omega(g)$ – это сложность модели. G - это семейство возможных интерпретируемых моделей, например, всевозможные линейные модели. Мера близости $\pi(x)$ определяет окрестность вокруг объекта x, которую мы рассматриваем для построения объяснения.

Шаги для обучения локальных суррогатных моделей:

1. Выбрать интересующий вас объект, для которого нужно получить объяснение предсказания от модели - "черного ящика".
2. Получить предсказания на других объекта выборки (выборку можно создать искусственно).
3. Взвесить объекты в соответствии с их близостью к интересующему объекту.
4. Обучить взвешенную, интерпретируемую модель на этом наборе данных.
5. Объяснить прогноз, интерпретируя локальную модель.

[3]Термин "черный ящик" можно воспринимать как модель, о структуре которой мы ничего не знаем. Все, что мы можем: 1) Подать на вход модели объект из обучающей выборки, 2) Получить предсказание модели

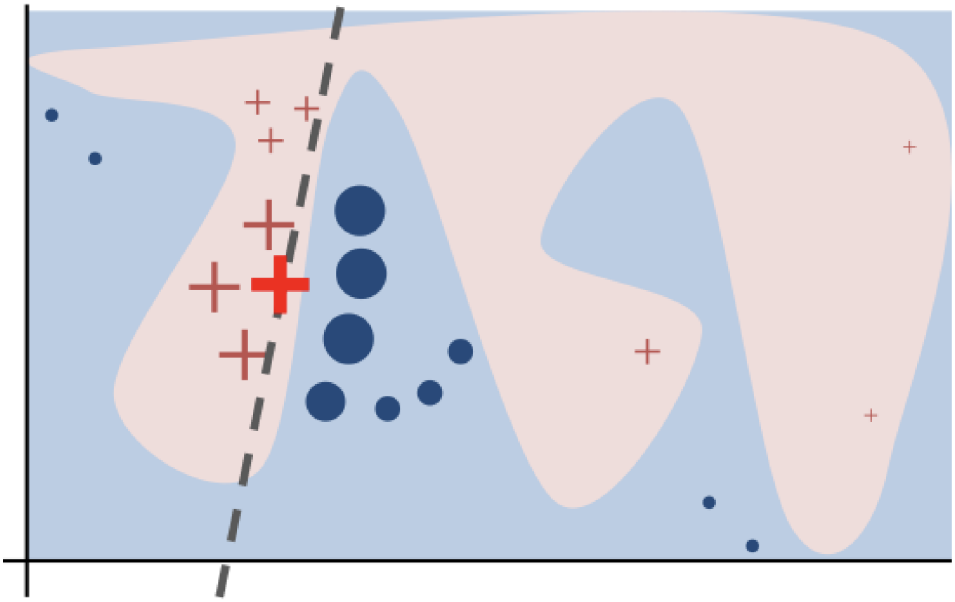


Рис. 2: Принцип работы LIME

Как вы можно получить "возмущенные" данные? Это зависит от типа объектов, которые могут быть текстовыми, графическими или табличными. Для текста и изображений решение состоит в том, чтобы включать или выключать отдельные слова или "суперпиксели". В случае табличных данных LIME создает новые выборки с помощью небольших изменений уже существующих объектов. К примеру, если все признаки объекта являются числовыми, то можно сгенерировать новые данных в некоторой окрестности объектов.

2.4 SHAP

– Понятно, что писать –

2.5 DEEP Lift/Gradient based методы

– Слышал, но пока что не изучал –

3 Существующие методы интерпретации временных рядов

Название не вполне удачное – часто интерпретацией временных рядов называют разложение его на компоненты: сезонность, тренд и т.д.

4 Предлагаемые методы

Задачи seq2seq, seq2label, быть может другие задачи

4.1 "Суперпризнаки"

– +- Понятно, что писать, но нужно проводить эксперименты –

4.2 CNN

– +- Понятно, что писать, но нужно проводить эксперименты –

4.3 Графовые нейросети

– Пока непонятно, что писать, нужно проводить эксперименты –

5 Применение к временным рядам

5.1 Активность головного мозга

– С данными туговато пока что –

Да и в принципе пока что не получается красной нитью вплести эту тему в текст

5.2 Другие примеры

– Здесь много разного можно напридумывать –

6 Заключение

–В последнюю очередь–

7 Полезные ссылки

- https://arxiv.org/pdf/1909.07082.pdf хорошая статья, приведен метод оценивания качества XAI
- https://arxiv.org/pdf/2009.13211.pdf Хорошая статья, предложен совершенно другой метод объяснения временных рядов
- https://proceedings.neurips.cc/paper/2020/file/2c29d89cc56cdb191c60db2f0bae796b-Paper.pdf просто обзор методов для разных областей

- https://boa.unimib.it/retrieve/handle/10281/324847/492202/Manuscript.pdf возможно что-то полезное, не читал пока что
- https://arxiv.org/pdf/2109.12935.pdf обзор методов для временных рядов

https://github.com/marcoancona/DeepExplain https://mne.tools/stable/auto_tutorials/preprocessing/40_artifact_correction_ica.html – красивые картинки с головами

7.1 Датасеты

ECG hearbeat dataset – sec2label
UCR Time Series Classification Archive – очень много разного
https://github.com/lmmentel/awesome-time-series – очень много разного
https://github.com/bhimmetoglu/time-series-medicine – три датасета
Human Activity Recognition – sec2label
Physionet MIT BIH Arrhythmia – ???
https://github.com/bhimmetoglu/time-series-medicine/tree/master/EEG – sec2label(EEG)
EEG Eye State DataSet – sec2label
https://github.com/effervescent-shot/Dream-Prediction – sec2sec (EEG)

Список литературы

[1] Ribeiro M. T., Singh S., Guestrin C. "Why should i trust you?"Explaining the predictions of any classifier //Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. – 2016. – С. 1135-1144.

[2] Fogel S. et al. Scrabblegan: Semi-supervised varying length handwritten text generation //Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. – 2020. – С. 4324-4333.

[3] Wang T. et al. Decoupled attention network for text recognition //Proceedings of the AAAI Conference on Artificial Intelligence. – 2020. – Т. 34. – №. 07. – С. 12216-12224.

[4] www.kaggle.com/constantinwerner/cyrillic-handwriting-dataset

[5] github.com/abdoelsayed2016/HKR_Dataset

[6] https://fki.tic.heia-fr.ch/databases/iam-handwriting-database