Машинное обучение: Объяснительный ИИ (Explainable AI (XAI) or ML)

Объяснительный ИИ (Explainable AI or ML)

Уткин Л.В.

Санкт-Петербургский политехнический университет Петра Великого

Термины и составляющие

1. Прозрачность (transparency): рассматривает подход на основе машинного обучения
2. Интерпретируемость (interpretability): рассматривает модель машинного обучения совместно с данными
3. Объяснимость (explainability): рассматривает модель, данные и участие человека

Прозрачность (1)

- Подход МО прозрачен, если его разработчик может описать процессы, которые извлекают параметры модели из обучающих данных и генерируют метки тестовых данных.
- Прозрачность подхода МО касается его различных компонентов: общей структуры модели, отдельных элементов модели, алгоритма обучения и того, как конкретное решение получается с помощью алгоритма.

Прозрачность глубоких нейронных сетей (пример)

- Модель прозрачна, так как соотношение вход-выход и ее структура могут быть записаны в мат-ких терминах.
- НО: слои - их число, размер или функции активации выбираются эвристически и не определяются знанием, поэтому эти решения не являются прозрачными.
- Алгоритм обучения прозрачен, например, градиентный спуск.
- НО: выбор параметров (скорость обучения, размер пакета и другие) эвристический, непрозрачный.
- Нескольких локальных минимумов -> решение часто не воспроизводимо -> не является (полностью) алгоритмически прозрачным.

Интерпретируемость (1)

- Цель интерпретируемости - представить человеку некоторые свойства модели ML в понятных терминах.
- В идеале - ответить на вопрос: “Можем ли мы понять, на чем алгоритм ML основывется в своем решении?”

Интерпретируемость (2)

- Интерпретация может быть получена посредством прокси-моделей, которые аппроксимируют предсказания исходной сложной модели.
- Деревья решений, линейные модели.
- Пример - LIME (Local Interpretable Model-Agnostic Explanations) (Ribeiro et al., 2016): линейная прокси-модель в окрестности данного.
- В отличие от прозрачности, для достижения интерпретируемости **всегда** используются данные.

Критерии для методов интерпретации

1. Intrinsic or post hoc
2. Model-specific or model-agnostic
3. Local or global

Критерии (intrinsic or post hoc)

- Внутренняя интерпретируемость - использование модели МО, которая интерпретируема (линейные модели, модели на основе деревьев).
- Post hoc интерпретируемость - выбор и обучение модели черного ящика (композиции, нейронные сети) и применение методов интерпретируемости после обучения (значимость признаков).

Критерии (model-specific or model-agnostic)

- Специфичные для модели методы интерпретации специфичны зависят исключительно от каждой модели. Это могут быть коэффициенты, p-values, оценки AIC, правила из дерева решений и так далее.
- Независимые от модели методы интерпретации (агностические) могут использоваться для любой модели МО. Работают путем анализа (и возмущений входов) признаков пар вход-выход.
- Эти методы не имеют доступа к каким-либо внутренним компонентам модели, таким как веса, ограничения или допущения.

Критерии (local or global)

- Эта классификация интерпретации говорит о том, объясняет ли метод интерпретации одно предсказание или все поведение модели.

Глобальная интерпретация

- Ответы на вопросы:
  - Как модель делает прогнозы?
  - Как подмножества модели влияют на решения модели?
- Глобальная интерпретируемость - это способность объяснять и понимать решения модели на полном наборе данных.

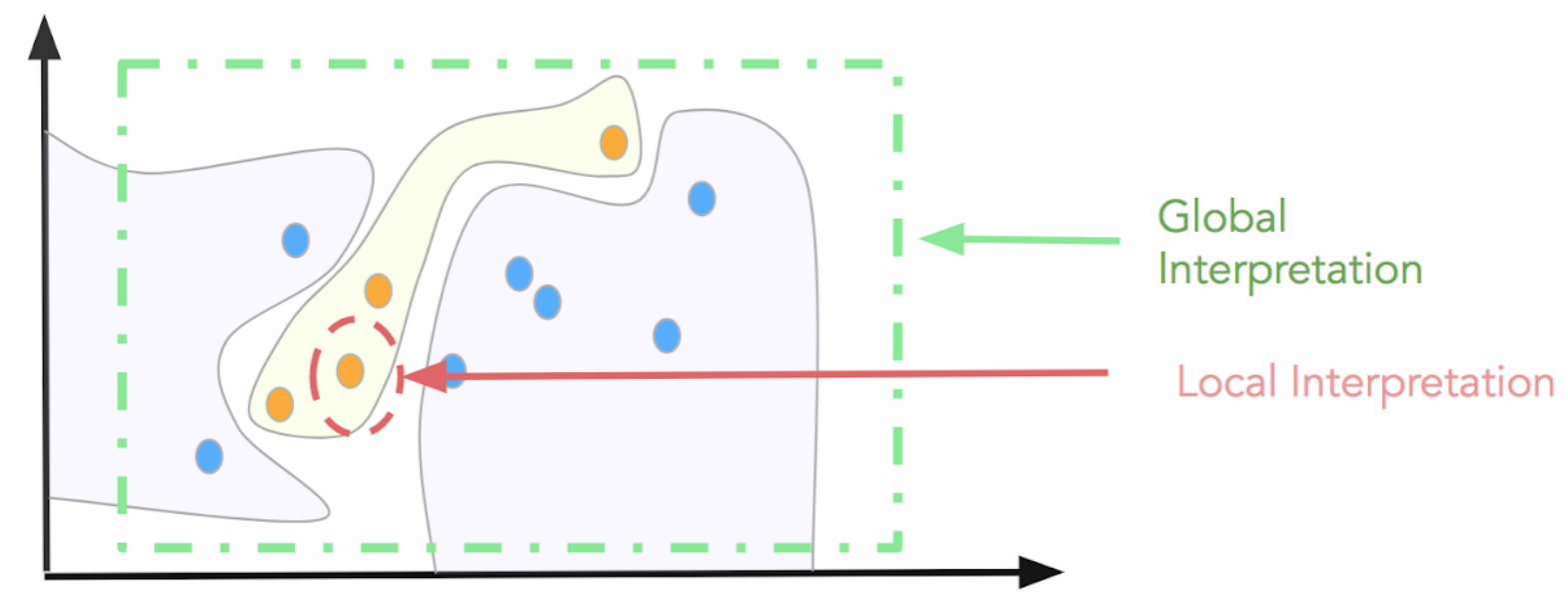
Локальная интерпретация

- Ответы на вопросы:
  - Почему модель принимает конкретное решение для одного примера?
  - Почему модель принимает конкретные решения для группы примеров?
- Для локальной интерпретируемости рассматриваем модель как черный ящик.

Локальная интерпретация (2)

- Для понимания решений по прогнозированию для одной точки данных рассматриваем локальную область вокруг этой точки.
- Локальные распределения данных и пространство признаков могут вести себя совершенно иначе и давать более точные объяснения в отличие от глобальных интерпретаций.
- Используют комбинацию глобальной и локальных интерпретаций, чтобы объяснить решения для группы примеров.

Локальная и глобальная интерпретации



Объяснимость

- Объяснение - это набор признаков интерпретируемой области, которые внесли вклад в данный пример для принятия решения (Montavon et al., 2018)
- Интерпретация может быть объяснением только при наличии дополнительной контекстуальной информации, основанной на **знании предметной области** (domain knowledge) и **цели анализа**, т.е. объяснимость не может быть достигнута чисто алгоритмически.
- Сама по себе интерпретация модели для отдельных данных может не дать объяснения для понимания решения. Например, наиболее выжные признаки могут быть одинаковыми для нескольких примеров.
- Объяснение зависит от основной цели анализа.

Противоречивые объяснения (1)

- Обычно спрашивают, не почему был сделан определенный прогноз, а почему этот прогноз был сделан вместо другого прогноза.
- Для прогноза стоимости дома человека может интересовать, почему прогнозируемая цена была выше по сравнению с более низкой ценой, которую он ожидал.
- Когда заявка на кредит отклонена, меня не интересует, почему отказ. Меня интересуют факторы моей заявки, которые должны измениться, чтобы она была принята.
- Противоречивые объяснения легче понять, чем полные объяснения.

Противоречивые объяснения (2)

- Врач задается вопросом: «Почему лечение не сработало на пациенте?»
- Полное объяснение, почему лечение не работает, включает: пациент болеет с 10 лет, 11 генов сверхэкспрессированы, что делает болезнь более тяжелой, организм пациента разрушается, лекарство неэффективно
- Сравнительное объяснение - отвечает на вопрос по сравнению с другим пациентом, для которого препарат работал, может быть проще: у пациента есть комбинация генов, которые делают лекарство неэффективным, по сравнению с другим пациентом
- Лучшее объяснение - это то, что подчеркивает наибольшую разницу между объектом интереса и эталонным объектом

Объяснения снова

- Люди не ожидают, что объяснения охватят полный список причин события. Выбирается одна или две причины из огромного числа возможных причин в качестве объяснения.
- Люди больше фокусируются на аномальных причинах, чтобы объяснить события. Если один из признаков был аномальным и он влиял на прогноз, его следует включить в объяснение, даже если другие «нормальные» признаки имеют такое же влияние на прогноз как аномальный.
- Хорошие объяснения подтверждаются в реальности. Когда говорим, что три балкона увеличивают цену дома, это должно быть верно для других домов.

Глобальная интерпретация - feature importance

- Значимость признаков - какие признаки оказывают наибольшее влияние на прогнозируемые значения?
- Алгоритм: значимость перестановок (permutation importance)
  1. Получить обученную модель и записать признаки в виде таблицы (столбец - признак)
  2. Перемешать значения в одном столбце, сделать прогнозы, используя полученный набор данных. Снижение точности - значимость признака, который перемешали.
  3. Вернуться к исходной таблице (отмена перемешивания из шага 2). Повторить шаг 2 со следующим столбцом в таблице, пока не будут найдены значимости каждого столбца.

Глобальная интерпретация - Partial Dependence Plot

- **Значимость признаков** показывает, **какие** признаки больше всего влияют на прогнозы, **график частичной зависимости** показывает, **как** признак влияет на прогнозы
- График частичной зависимости показывает, какая зависимость между признаком и выходом: линейнеая, монотонная или более сложная

График частичной зависимости (пример 1)

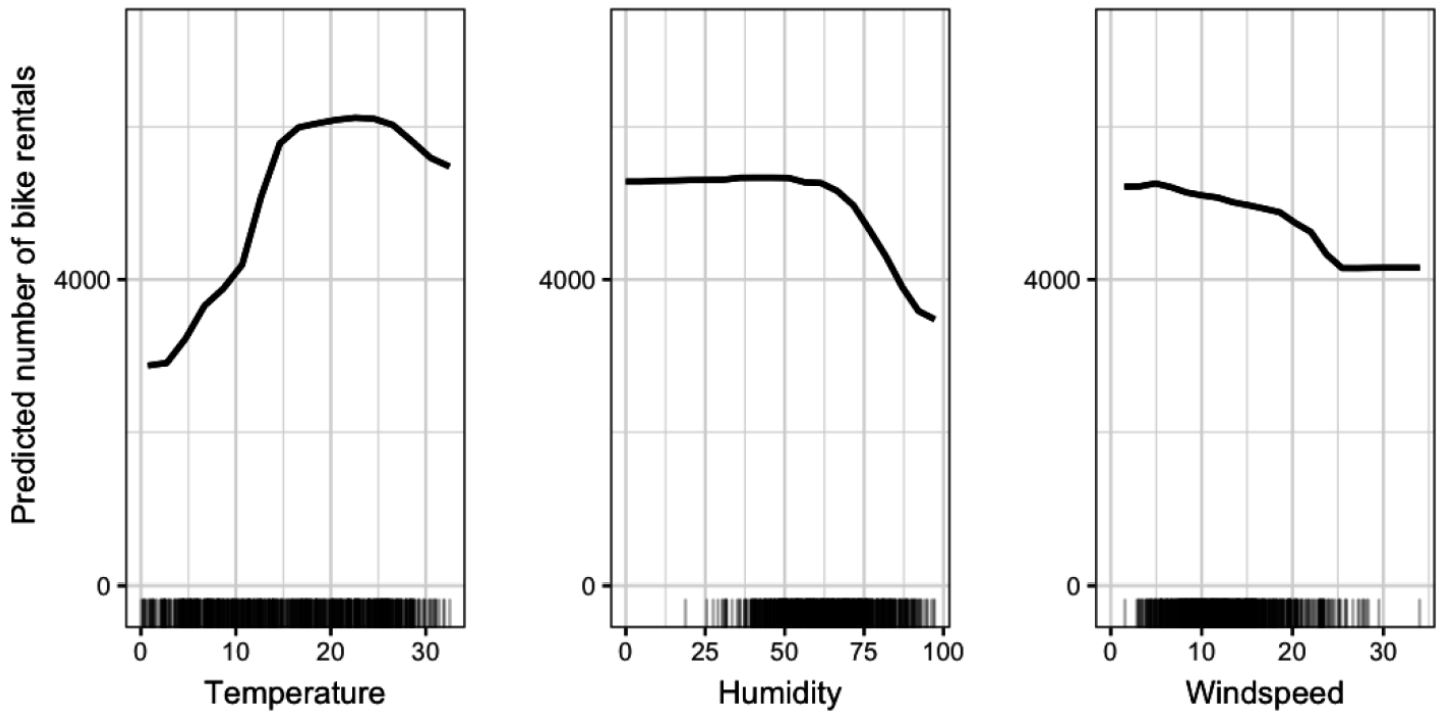


График частичной зависимости (пример 2 - взаимодействие признаков)

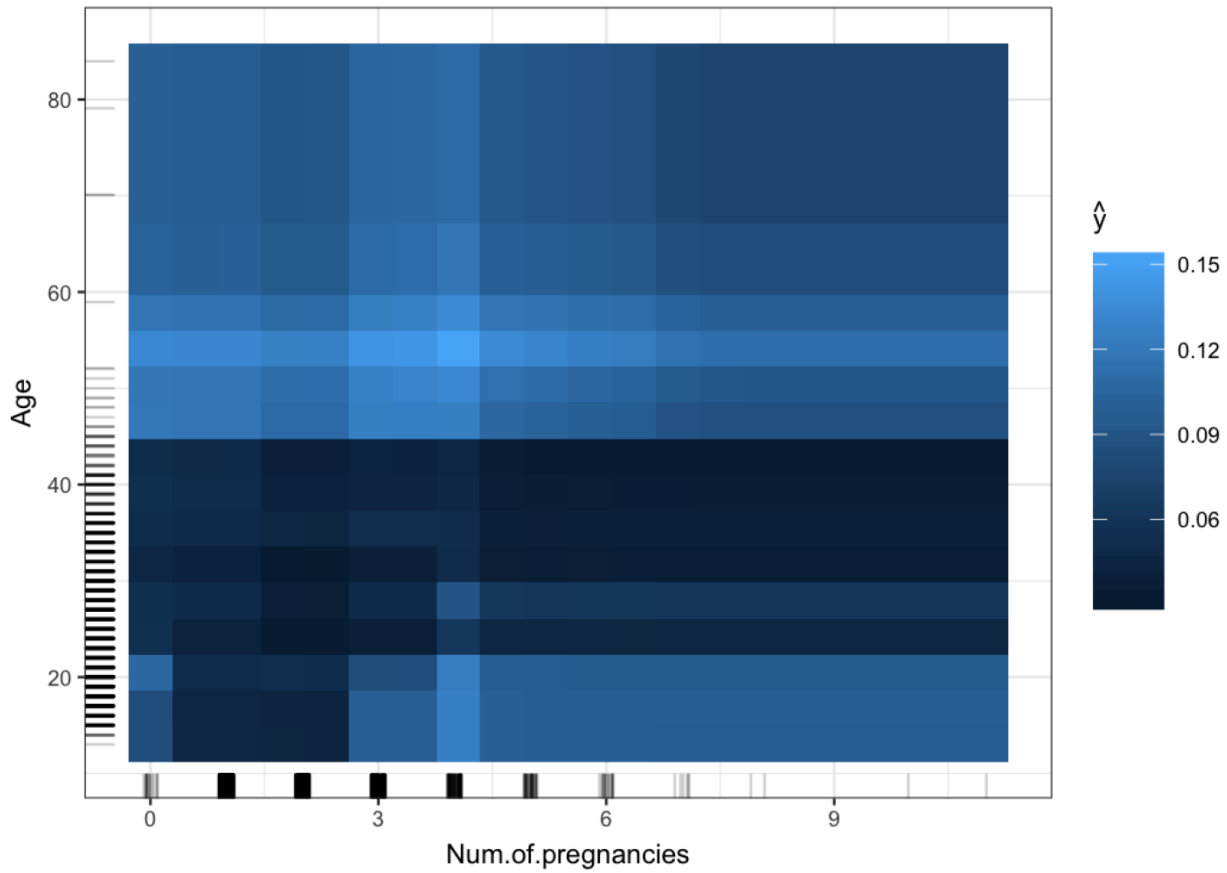


График частичной зависимости

- Частичная зависимость:

$$f_{x_S}(x_S) = \mathbb{E}_{x_C}\left[f(x_S, x_C)\right] = \int f(x_S, x_C) d\mathbb{P}(x_C)$$

- x_S - множество признаков, для которых график частичной зависимости определяется
- x_C - все другие признаки, используемые в модели f; $x = x_S || x_C$ (конкатенация)
- Частичная зависимость работает путем маргинализации выходных данных модели f по распределению признаков x_C, так что оставшаяся функция показывает связь между x_S и прогнозом

График частичной зависимости

- Частичная зависимость по данным из датасета:

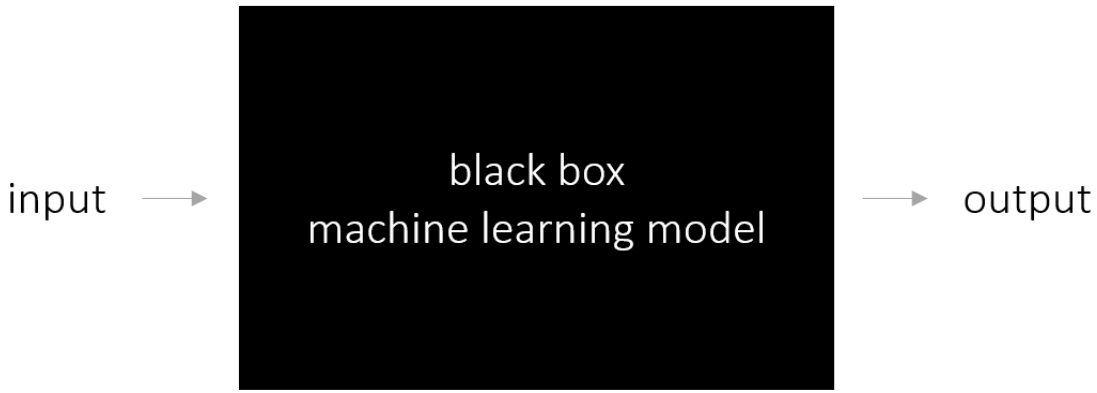
$$f_{x_S}(x_S) = \frac{1}{n}\sum_{i=1}^{n} f(x_S, x_{C_i})$$

- x_{C_i} - фактические значения признаков из датасета, в которых мы не заинтересованы
- Используемое предположение: признаки x_S не коррелируют с признаками x_C

Локальная интерпретация - Метод LIME (1)

- **Local Interpretable Model-agnostic Explanations** (Ribeiro, Singh, Guestrin, 2016)
- **Агностицизм**: LIME не делает никаких предположений относительно модели, прогноз которой объясняется, для него модель как «черный ящик»
- **Интерпретируемость**: LIME использует представление данных (называемое интерпретируемым представлением), которое отличается от исходного пространства признаков
- **Локальность**: LIME дает объяснение в окрестности примера, который хотим объяснить.

Метод LIME (2)



Метод LIME (3)

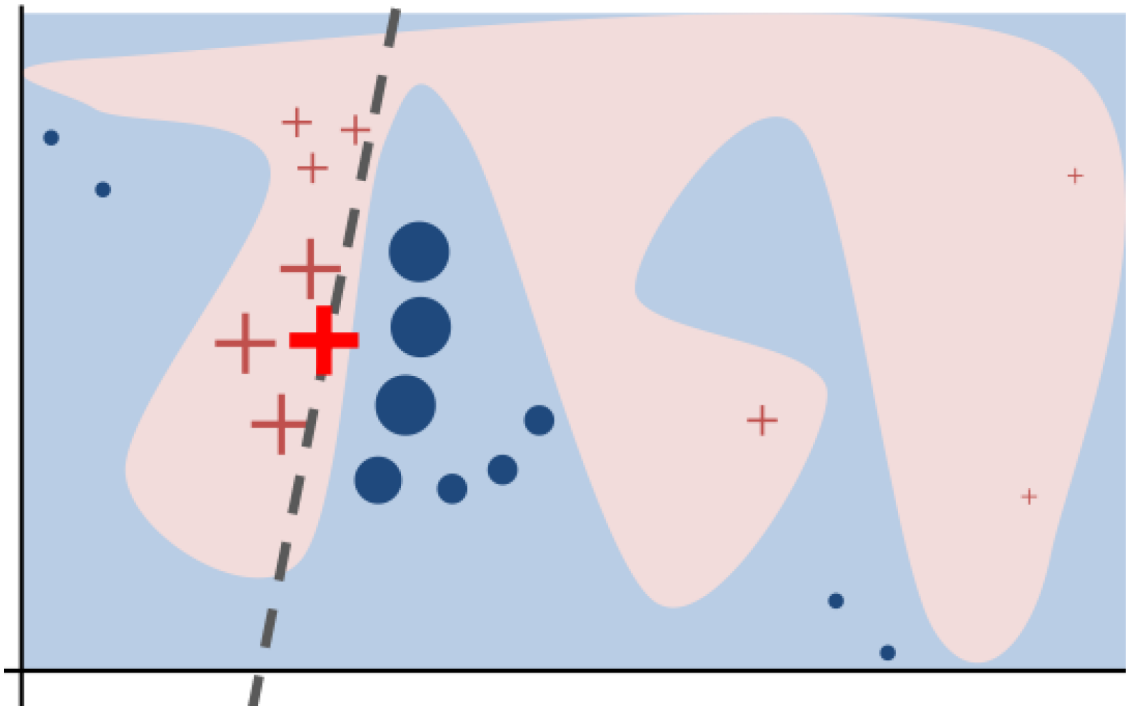
LIME минимизирует функцию

$$\xi = \arg\min_{g \in G} L(f, g, \pi_X) + \Omega(g)$$

g - объяснительная модель для оригинальной модели f;
π_X - веса в виде ядер

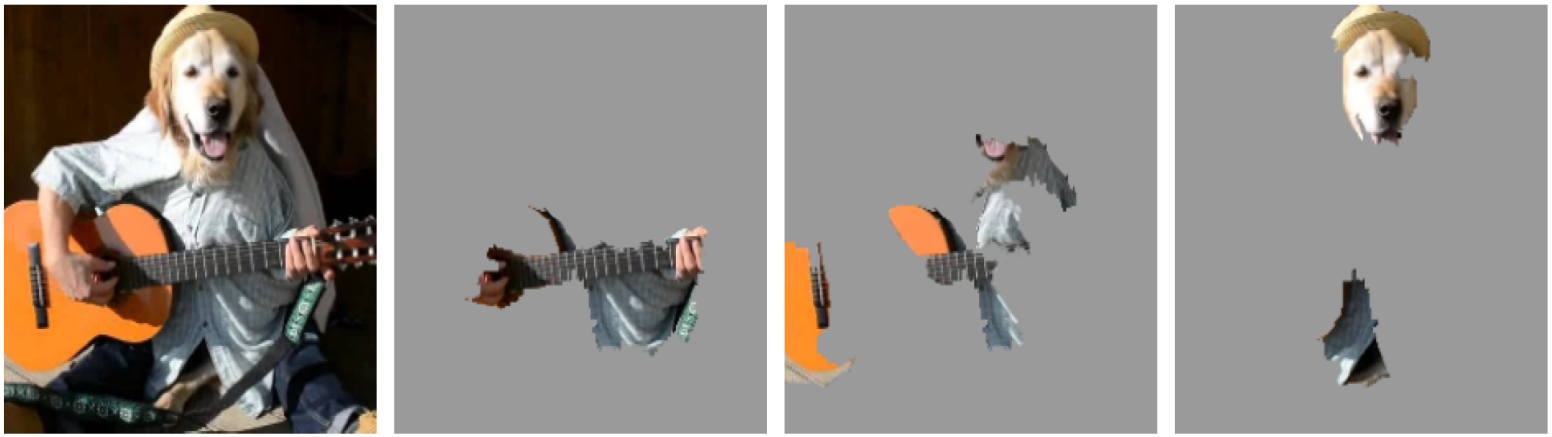
$$g(z) = \phi_0 + \sum_{i=1}^{M} \phi_i z_i$$

Метод LIME (4)



M.T. Ribeiro, S. Singh, C. Guestrin. "Why should I trust you?: Explaining the predictions of any classifier". In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016

Метод LIME (5)



(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Объяснение вариантов классификации. Три основных прогнозируемых класса: “Electric Guitar” (p = 0.32), “Acoustic guitar” (p = 0.24) и “Labrador” (p = 0.21)

Shapley Values (1)

- IME (Interactions-based Method for Explanation) (Strumbelj and Kononenko, 2010)
- Числа Шепли - это вклад каждого игрока, усредненный по каждой возможной последовательности, в которой игроки могли быть добавлены в группу

$$\phi_i = \sum_{S \subseteq F \setminus \{i\}} \frac{|S|!(|F| - |S| - 1)!}{|F|!} \left[f_{S \cup \{i\}} \left(x_{S \cup \{i\}} \right) - f_S \left(x_S \right) \right]$$

- $|F|$ - размер полной коалиции; S - подмножество коалиции, которое не включает игрока i, а $|S|$ - размер S, $S!$ - число перестановок множества S
- В квадратных скобках: «насколько больше выигрыш, когда мы добавляем игрока i к подмножеству S»

Shapley Values (2)

- А как теперь с признаками?
- Вклад i-го признака:

$$\phi_i = \sum_{z' \subseteq x'} \frac{|z'|!(M - |z'| - 1)!}{M!} \left[f_x\left(z'\right) - f_x\left(z' \backslash i\right) \right]$$

- M - общее число признаков; z' - подмножество признаков, которое является объяснением
- Оцениваем значение модели с и без i-го признака ($f_x\left(z'\right)$ и $f_x\left(z' \backslash i\right)$)

Shapley Values (3)

- Интерпретация числа Шепли X: Значение признака A сделало вклад X в прогнозируемое значение конкретного примера по сравнению со средним прогнозируемым значением для набора обучающих данных.

Это вклад значения признака в разность между фактическим прогнозируемым значением и средним прогнозируемым значением.

Shapley Values - интерпретация (4)

- Интерпретация числа Шепли x: Значение признака A сделало вклад x в прогнозируемое значение конкретного примера по сравнению со средним прогнозируемым значением для набора обучающих данных.
- Это вклад значения признака в разность между фактическим прогнозируемым значением и средним прогнозируемым значением.

Вопросы

?