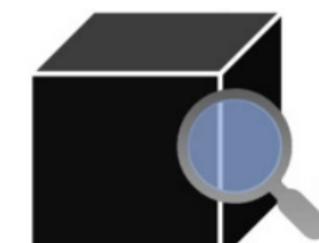


Интерпретируемое машинное обучение

Виктор Китов

v.v.kitov@yandex.ru

interpretation



black-box model

explanation



Содержание

- 1 Интерпретируемые модели
- 2 Важность признаков
- 3 Локальная интерпретация простой моделью
- 4 Графики влияния признаков
- 5 Прототипы и критики
- 6 Контрафактические объяснения

Интерпретируемые методы

- Книга по теме:
 - [Interpretable Machine Learning \(Christoph Molnar\)](#).
- Метод ближайших центроидов и K-NN обладают интерпретируемостью.
 - прогноз таков, потому что объект близок к ...
- ↑ интерпретируемость м-да близж. центроидов можно за счёт выбора реального объекта вместо центроида.
- Метод наивного Байеса интерпретируем за счёт независимого мультипликативного вклада кажд. пр-ка в прогноз:

$$p(y|x) \propto p(y) \prod_{d=1}^D p(x^d|y)$$

Линейная регрессия

$$\hat{y} = w_0 + w_1x^1 + \dots + w_Dx^D$$

- Предположения:
 - каждый признак вносит линейн. вклад в прогноз с весом β_i
 - характер влияния не зависит от значений др. признаков
- Применяется, когда важна не только точность, но и интерпретация.
 - насколько признак влияет
 - характер влияния (положительный / отрицательный)

Линейная регрессия

$$\hat{y} = w_0 + w_1x^1 + \dots + w_Dx^D$$

- Предположения:
 - каждый признак вносит линейн. вклад в прогноз с весом β_i
 - реально: нелинейный вклад
 - характер влияния не зависит от значений др. признаков
 - реально: признаки оказывают совместное влияние
- Применяется, когда важна не только точность, но и интерпретация.
 - насколько признак влияет
 - характер влияния (положительный / отрицательный)

Влияние признаков

- Влияние признаков:

- $x^i \in \mathbb{R}$: $\uparrow x^i$ на 1 \uparrow прогноз на w_i .
- $x^i \in \{0, 1\}$: активация признака \uparrow прогноз на w_i .
- $x^i \in \{1, 2, \dots, C\}$: j -ая категория \uparrow прогноз на β_{ij} по сравнению с референсной 1-й категорией при кодировке:

[0,0,...0,0]	1
[0,0,...0,1]	2
[0,0,...1,0]	3
...	...
[1,0,...0,0]	C

- Если хотим анализировать самые значимые признаки - Lasso регуляризация.

Значимость признаков

- Значимость признака x^i - значение t-теста^{1,2} с $H_0 : w_i = 0$:

$$t_{\beta_i} = \frac{\hat{w}_i}{SE(\hat{w}_i)}$$

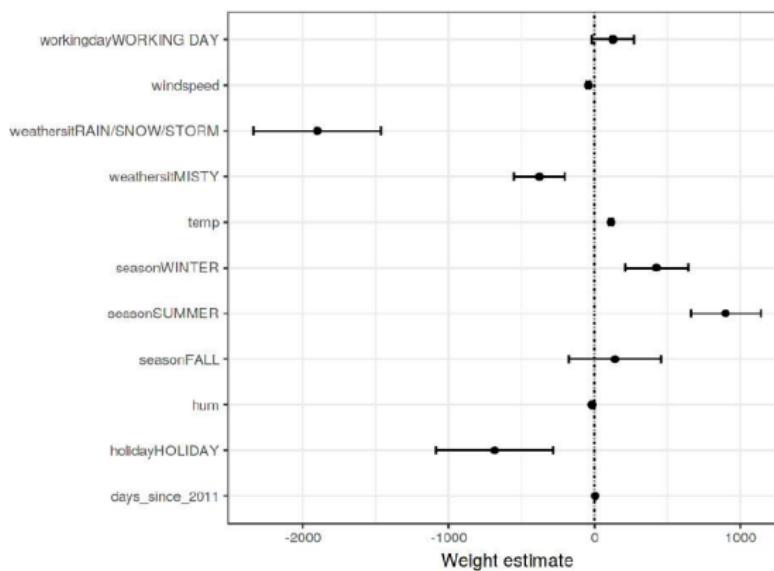
где $SE(\hat{w}_i)$ - станд. ошибка для w_i

$$SE(\hat{w}_i) = \left\{ \hat{\sigma}^2 (X^T X)^{-1} \right\}_{ii}$$

¹Тестирование гипотез для регрессии.

²Вывод для одного признака.

Пример анализа значимости признаков (BikeRental)



- Дождь/снег/шторм сильно влияют на #велосипедов.
- А индикатор workingday - незначимо
 - возможно, за счёт присутствия индикатора праздника

Пример анализа весов (BikeRental)

	Weight	SE	t
(Intercept)	2399.4	238.3	10.1
seasonSUMMER	899.3	122.3	7.4
seasonFALL	138.2	161.7	0.9
seasonWINTER	425.6	110.8	3.8
holidayHOLIDAY	-686.1	203.3	3.4
workingdayWORKING DAY	124.9	73.3	1.7
weathersitMISTY	-379.4	87.6	4.3
weathersitRAIN/SNOW/STORM	-1901.5	223.6	8.5
temp	110.7	7.0	15.7
hum	-17.4	3.2	5.5
windspeed	-42.5	6.9	6.2
days_since_2011	4.9	0.2	28.5

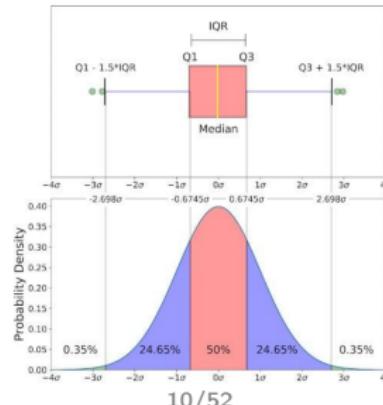
- ↑ температуры на 1 градус ↑#арендованных вел-дов на 110.
- Для референсной категории признака weathersit=good (weather situation)
 - weathersit=MISTY: -379 велосипедов
 - weathersit∈[RAIN,SNOW,STORM]: -1901 велосипед
- Для сравнения силы влияния разных признаков, их нужно привести к одной шкале.

Особенности

- Признаки принимают зависимые значения.
- Не имеет смысла $\uparrow x^i$ не изменяя x^j
 - например, #комнат в доме и жилая площадь
- Признаки могут влиять нелинейно
 - нужно учитывать их нелинейные версии $\phi(x^j)$

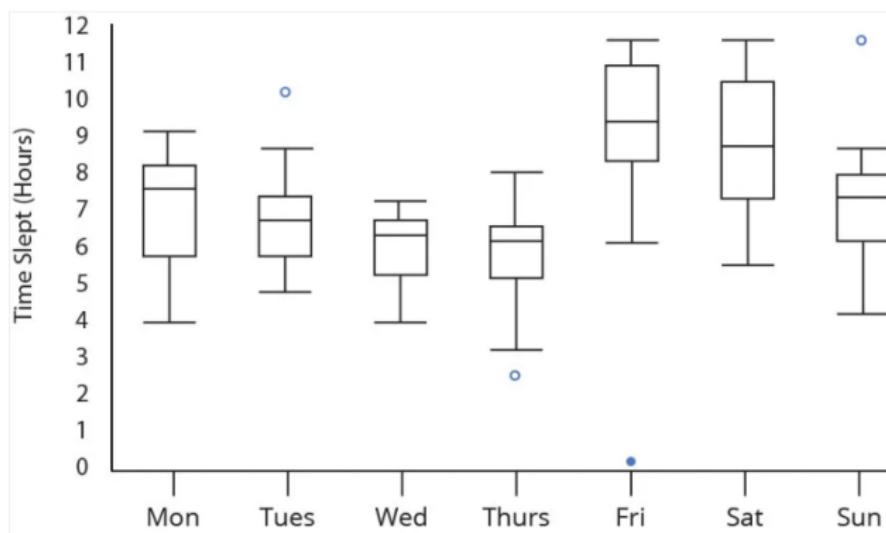
Box-plot для визуализации распределений

- Полоска посередине - медиана = $Q_2 = \text{quantile}_{0.5}$.
- Границы прям-ка: $Q_1 = \text{quantile}_{0.25}$ и $Q_3 = \text{quantile}_{0.75}$
- $IQR = \text{quantile}_{0.75} - \text{quantile}_{0.25}$
 - InterQuantile Range, межквартильный разброс
- Границы отрезка: мин. и макс. наблюдения
 $\in [Q_1 - 1.5 * IQR, Q_3 + 1.5 * IQR]$
 - для $\mathcal{N}(0, 1)$ только 0.7% точек лежат вне отрезка
- Точки вне отрезка (выбросы) рисуем отдельно.



Применение box-plots^{3,4}

Можем анализировать группы наблюдений по медианам, межквартильному разбросу и границам значений.

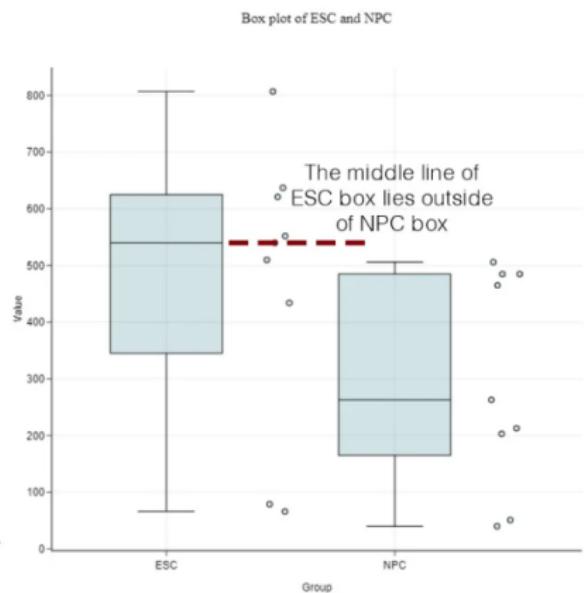
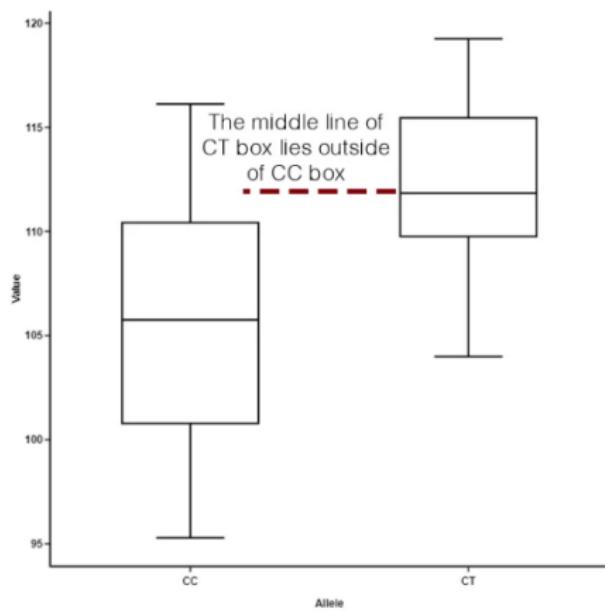


³<https://www.simplypsychology.org/boxplots.html>

⁴<https://builtin.com/data-science/boxplot>

Применение box-plots⁵

Можем сравнивать центры распределений (медианы). Если медиана вне $[Q1, Q3]$, то расхождение стат. значимо.

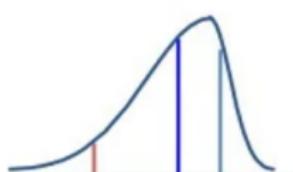


⁵ <https://www.simplypsychology.org/boxplots.html>

Применение box-plots⁶

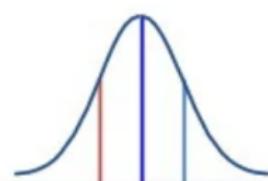
Можем оценивать симметричность распределения по положению Q2 внутри [Q1,Q3] и внутри отрезка.

Left-Skewed



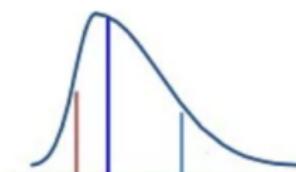
Q_1 Q_2 Q_3

Symmetric

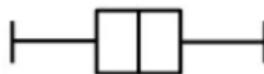
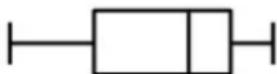


Q_1 Q_2 Q_3

Right-Skewed



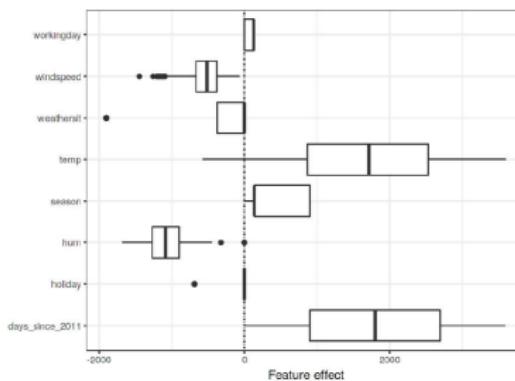
Q_1 Q_2 Q_3



⁶<https://www.simplypsychology.org/boxplots.html>

График эффектов

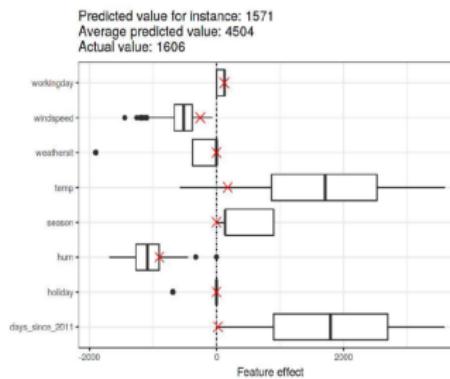
- Эффект признака x^i на итоговый прогноз = $w_i x^i$
 - не зависит от масштаба x^i , влияние для фикс. x
- График эффектов (effect plot) - распределение эффектов по всем объектам:



- Самый сильный эффект - температура и время с начала наблюдений (тренд).

График эффектов для одного объекта

- Отложим эффекты отдельных признаков на прогноз для одного объекта (красные крестики)



- Сравнительно малый прогноз объясняется низкой температурой и малым временем с начала наблюдений
- Центрирование вещественных признаков и выбор "типичной" референсной категории в one-hot кодировании
↑ интерпретируемость.

Логистическая регрессия

$$p(y = +1|x) = \frac{1}{1 + e^{-w^T x}}$$

$$1 + e^{-w^T x} = \frac{1}{p(y = +1|x)}$$

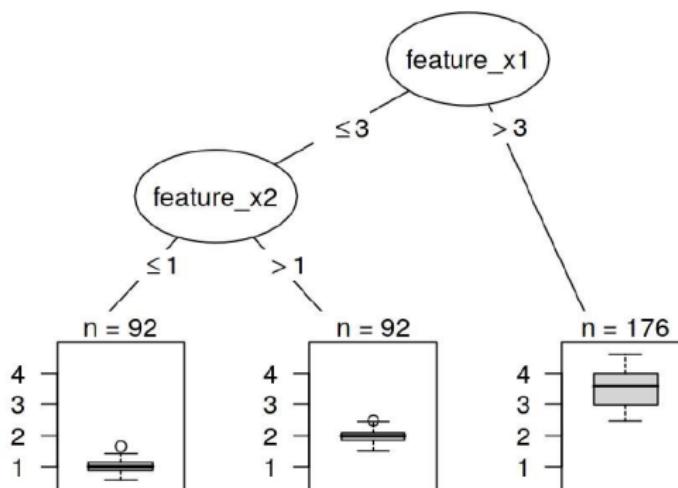
$$e^{-w^T x} = \frac{1}{p(y = +1|x)} - \frac{p(y = +1|x)}{p(y = +1|x)} = \frac{p(y = -1|x)}{p(y = +1|x)}$$

$$e^{w^T x} = \frac{p(y = +1|x)}{p(y = -1|x)}$$

$\uparrow x^i$ на 1 \Rightarrow odds ratio $= \frac{p(y=+1|x)}{p(y=-1|x)}$ в e^{w_i} раз

Деревья решений

- Деревья решений небольшой глубины - интерпретируемая модель:



Вклад признаков в прогноз

- В каждом узле дерева t считаем $\hat{y}(t)$, либо неопределенность $\phi(t)$.
- Декомпозиция прогноза по узлам

$$\hat{y}(x) = \sum_{t \in \text{path}(x) \setminus \{\text{root}\}} [\hat{y}(t) - \hat{y}(\text{Parent}(t))]$$

- Вклад признака в прогноз для x : сумма вкладов узлов, где признак участвовал.

Важность признаков: mean decrease in impurity

- Важность признаков по изменению критерия информативности (mean decrease in impurity, MDI).

Важность признаков: mean decrease in impurity

- Важность признаков по изменению критерия информативности (mean decrease in impurity, MDI).
 - рассмотрим признак f
 - пусть $T(f)$ -множество всех вершин, использующих f в функции ветвления
 - эффективность разбиения в t :

$$\Delta\phi(t) = \phi(t) - \sum_{c \in \text{children}(t)} \frac{N(c)}{N(t)} \phi(c)$$

- значимость f :

$$\sum_{t \in T(f)} N(t) \Delta\phi(t)$$
- Поощряет признаки с большим количеством уникальных значений.

Содержание

- 1 Интерпретируемые модели
- 2 Важность признаков
 - Перестановочная важность признаков
- 3 Локальная интерпретация простой моделью
- 4 Графики влияния признаков
- 5 Прототипы и критики
- 6 Контрафактические объяснения

Важность признаков

Перестановочная важность признаков

2 Важность признаков

- Перестановочная важность признаков

Важность признаков

Перестановочная важность признаков

Перестановочная важность признаков

- $L(f, X, Y)$ - потери модели на выборке X, Y .
 - напр., MSE, частота ошибок
- Перемешаем j -й столбец X , получим \tilde{X}_j
 - распределение j -го признака сохранится, но связь с y потерянна
- $L(f, \tilde{X}_j, Y)$ - потери модели на выборке \tilde{X}_j, Y .
- Перестановочная важность признака j (permutation feature importance):

$$\frac{L(f, \tilde{X}_j, Y)}{L(f, X, Y)} \text{ либо } L(f, \tilde{X}_j, Y) - L(f, X, Y)$$

- Не нужно перенастраивать модель, т.к. оцениваем все в рамках текущей модели.

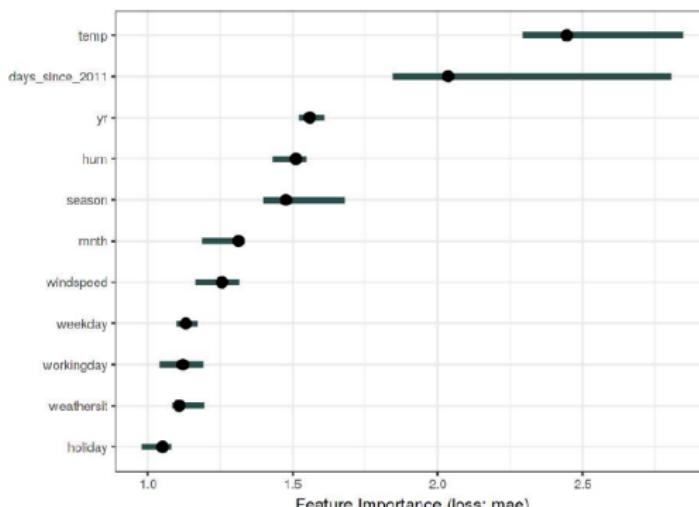
Перестановочная важность признаков

- На практике усредняют по всевозможным перестановкам.
 - чтобы не зависеть от частной перестановки
- Лучше использовать отношение, а не разность потерь
 - т.к. тогда можно сопоставлять важность признаков на разных задачах с разным уровнем L .
- Перестановочную важность признаков можно считать по
 - обучающей выборке: на что модель переобучилась?
 - тестовой выборке: какой признак полезнее для прогнозов?

Важность признаков

Перестановочная важность признаков

Пример: Bike Rental Dataset



Поскольку важность считается многократно - можем строить доверительные интервалы.

Анализ

- Глобальная и ёмкая мера важности признаков
- Мера привязана к определённой ф-ции потерь.
- Основывается на нереалистичных объектах.
 - за счёт случайности перестановок
- Если признаки скоррелированы - снижает важность каждого
 - модель восстанавливает информацию по похожему признаку

Содержание

- 1 Интерпретируемые модели
- 2 Важность признаков
- 3 Локальная интерпретация простой моделью
- 4 Графики влияния признаков
- 5 Прототипы и критики
- 6 Контрафактические объяснения

Метод LIME

Метод LIME (local interpretable model-agnostic explanations):

- ❶ Выбрать x для которого нужно объяснить прогноз
- ❷ Сгенерировать выборку вариаций в окрестности x

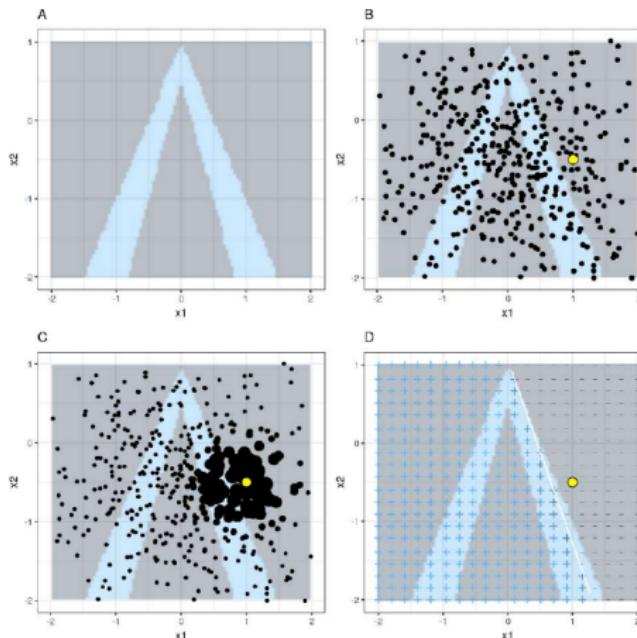
$$\{[\tilde{x}_1, f(\tilde{x}_1)]; [\tilde{x}_2, f(\tilde{x}_2)]; \dots [\tilde{x}_K, f(\tilde{x}_K)]\}$$

- ❸ Взвесить объекты по близости к x :

$$\{[w_1, \tilde{x}_1, f(\tilde{x}_1)]; [w_2, \tilde{x}_2, f(\tilde{x}_2)]; \dots [w_K, \tilde{x}_K, f(\tilde{x}_K)]\}$$

- ❹ Настроить интерпретируемую модель $g(x)$ по взвешенной выборке.
- ❺ Исследовать интерпретируемую модель.

Иллюстрация



LIME algorithm for tabular data. A) Random forest predictions given features x_1 and x_2 . Predicted classes: 1 (dark color) or 0 (light color). B) Instance of interest (yellow dot) and data sampled from a normal distribution (black dots). C) Assign higher weight to points near the instance of interest. D) Colors and signs of the grid show the classifications of the locally learned model from the weighted samples. The white line marks the decision boundary ($P(\text{class}=1) = 0.5$).

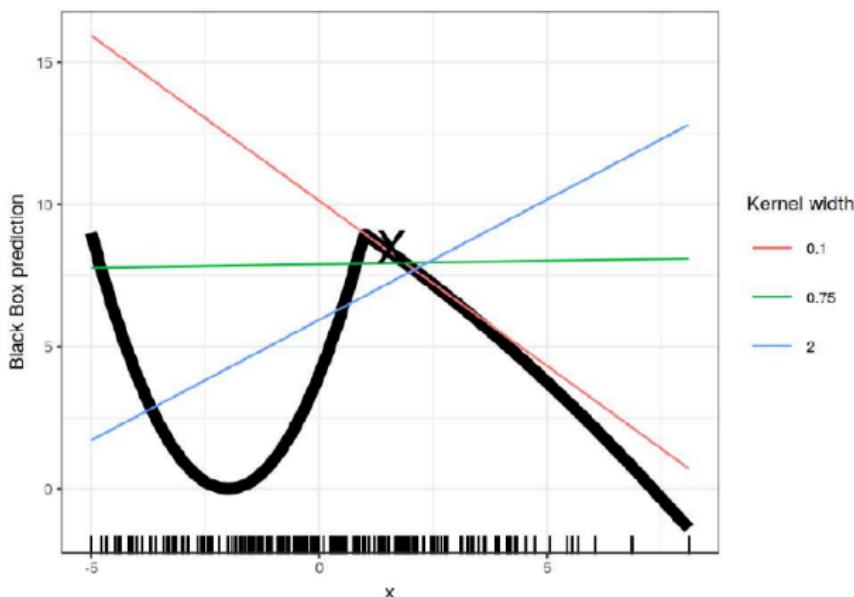
Обсуждение

- Интерпретируемая модель:
 - дерево небольшой глубины
 - линейная регрессия с малым $\#\text{признаков}$
 - отбор по L_1 регуляризации
 - forward-selection (последовательный выбор самых влияющих)
 - backward-selection (последовательное исключение самых незначимых)
- Важно контролировать ошибку аппроксимации $f(x)$ интерпретируемой $g(x)$
 - например, в сравнении с константным прогнозом:

$$\frac{\frac{1}{K} \sum_{k=1}^K (f(x_k) - g(x_k))^2}{\frac{1}{K} \sum_{k=1}^K \left(f(x_k) - \frac{1}{K} \sum_{k=1}^K f(x_k) \right)^2}$$

Выбор ширины окрестности x

В зависимости ширины окрестности x , в которой сэмплируется выборка, можем получать разные $g(x)$:



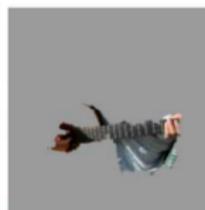
Генерирование выборки вокруг x

Генерирование выборки вокруг x :

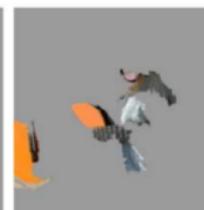
- x - вектор: вещественных чисел: добавить шум
- x - текст: включать/исключать слова
- x - изображение: включать/исключать суперпиксели



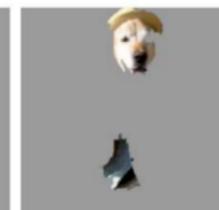
(a) Original Image



(b) Explaining Electric guitar



(c) Explaining Acoustic guitar



(d) Explaining Labrador

LIME explanations for the top 3 classes for image classification made by Google's Inception neural network. The example is taken from the LIME paper (Ribeiro et. al., 2016).

Интерпретирующая модель может использовать признаки, отличные от x (более интерпретируемые)

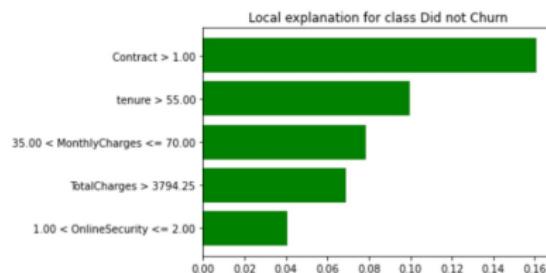
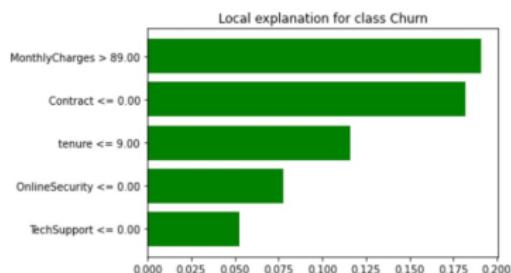
- например, суперпиксели изображения
- нужно исключать группы скоррелированных признаков

Пример для табличных данных

- Рассмотрим классификацию ухода клиентов от компании.
- Обоснуем прогнозы для 2x клиентов:

$$(x_1, f(x_1) = \text{churn}), (x_2, f(x_2) = \text{not churn})$$

- $g(x)$ - решающее дерево, вклад правил в его прогнозы⁷:



⁷ Ссылка на расчёт.

Содержание

- 1 Интерпретируемые модели
- 2 Важность признаков
- 3 Локальная интерпретация простой моделью
- 4 Графики влияния признаков
 - График частичной зависимости
 - График индивидуальных условных ожиданий
- 5 Прототипы и критики
- 6 Контрафактические объяснения

Графики влияния признаков

График частичной зависимости

4 Графики влияния признаков

- График частичной зависимости
- График индивидуальных условных ожиданий

График частичной зависимости

- График частичной зависимости (partial dependence plot, PDP) - функция группы признаков на прогноз.
 - $x = [u, v]$. u - интересующая группа признаков; v - остальные.

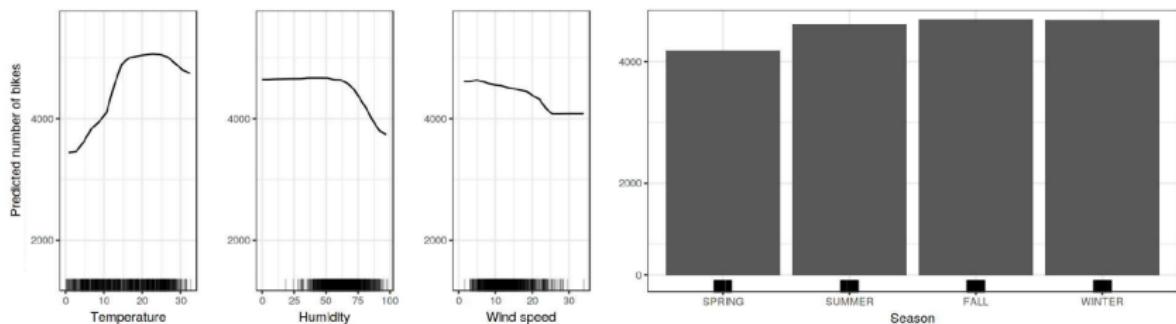
$$g_u(u) = \mathbb{E}_v \{f(u, v)\} = \int f(u, v) d\mathbb{P}(v)$$

- для интерпретируемости u - 1 или 2 признака.
 - Оценка:
- $$\hat{g}_u(v) = \frac{1}{N} \sum_{n=1}^N f(u, v_n)$$
- Предположение PDP - признаки u, v независимы.
 - если нарушено, $\{(u, v_n)\}_n$ будут включать малореальные объекты.
 - PDP даёт глобальную интерпретируемость модели.

Графики влияния признаков

График частичной зависимости

Пример: Bike Rental Dataset



- Зависимость от сезона частично покрыта признаком `#дней с начала измерений`.
- На оси отмечены наблюдаемые значения признака в выборке.

Анализ

- ⊕ : PDP интуитивен, легко реализовать.
- ⊖ : Показывает корреляцию, а не причинно-следственную связь.
- ⊖ : Интерпретируемо только для 1-2 признаков
- ⊖ : Имеет смысл для независимых u, v .
 - иначе $\{u, v_n\}_n$ могут соответствовать несуществующим объектам
 - пример: $x = [\text{рост}, \text{вес}]$ для пациента
- ⊖ : Слишком высокая агрегация
 - по всем точкам для каждого u (долго считать)
 - если для 50% связь положительна, а для 50% отрицательна, то можем увидеть отсутствие связи

Графики влияния признаков

График индивидуальных условных ожиданий

4 Графики влияния признаков

- График частичной зависимости
- График индивидуальных условных ожиданий

График индивидуальных условных ожиданий

- График частичной зависимости показывает усреднённый эффект признака на отклик по всем объектам.
 - если для 50% связь положительна, а для 50% отрицательна, то можем увидеть отсутствие связи
- График индивидуальных условных ожиданий (Individual Conditional Expectation, ICE) - зависимость отклика от признака для каждого объекта в отдельности.
 - разобъем x на интересующий признак u и все остальные v
 - график для каждого объекта центрируется, чтобы стартовать из одной точки

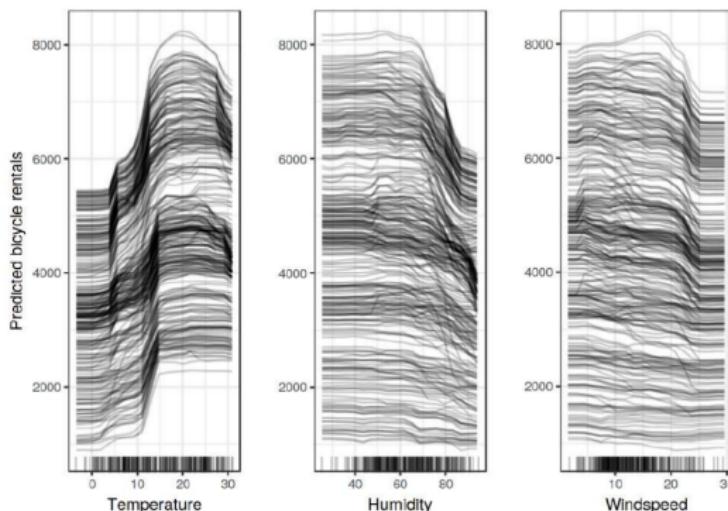
$$\{g_{un}(u) = f(u, v_n)\}_n \quad n = 1, 2, \dots N$$

- \oplus : Видны индивидуальные вариации по объектам.
- \ominus : Показывает отклики для возможно несуществующих объектов.
- \ominus : Перегружен информацией, если объектов много.

Графики влияния признаков

График индивидуальных условных ожиданий

Пример: bike rental dataset

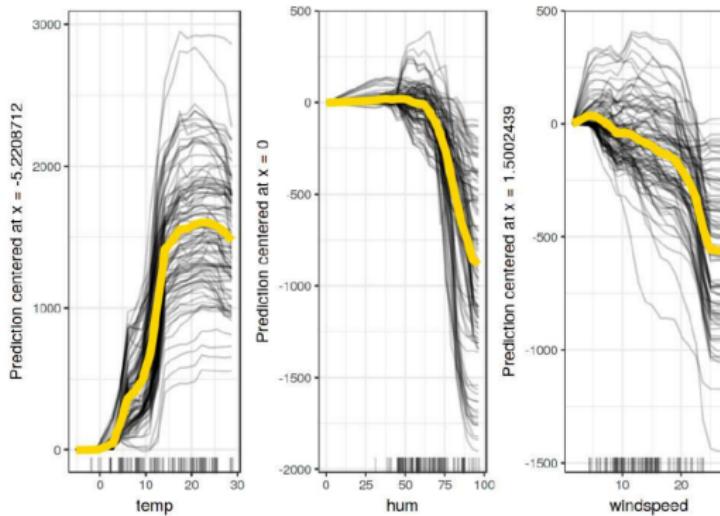


- Каждый график - для одного наблюдения.
- Можем видеть индивидуальные вариации по объектам.
 - сложно из-за наложения многочисленных графиков

Графики влияния признаков

График индивидуальных условных ожиданий

Пример с центрированием: bike rental dataset



- Centered ICE plot (c-ICE) - графики смещаются, чтобы стартовать из одной точки

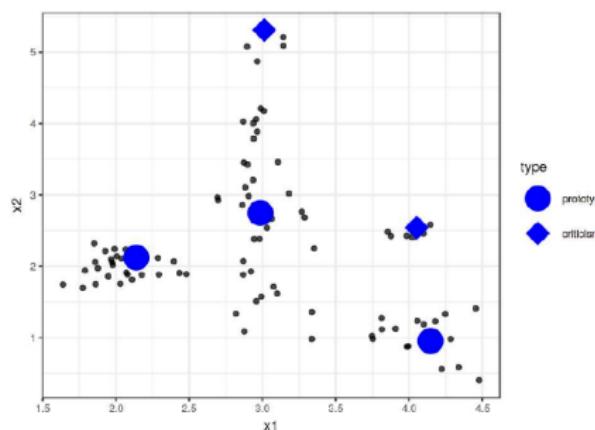
$$\{g_{un}(u) = f(u, v_n) - f(u_0, v_n)\}_n \quad n = 1, 2, \dots, N, \quad u_0 - \text{anchor}$$
- Лучше видны индивидуальные вариации по объектам.

Содержание

- 1 Интерпретируемые модели
- 2 Важность признаков
- 3 Локальная интерпретация простой моделью
- 4 Графики влияния признаков
- 5 Прототипы и критики
- 6 Контрафактические объяснения

Прототипы и критики

- Прототипы (prototypes) z_1, \dots, z_m - объекты, репрезентативные для x_1, \dots, x_n .
- Критики (criticisms) u_1, \dots, u_k - объекты, плохо описываемые прототипами.



- Например, K-medoids найдет прототипы (но не критиков).

Применения прототипов/kritиков

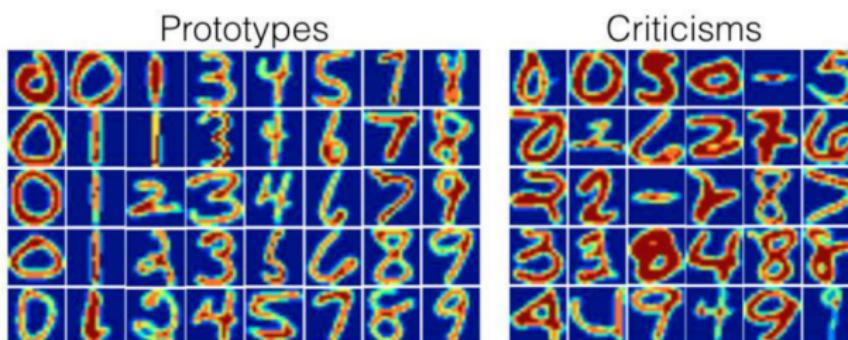
Применения прототипов/kritиков:

- Анализ сложно распределенных X .
- Объяснение работы модели $f(x)$ через её работу на прототипах:

$$\hat{y}(x) = f \left(\arg \min_n K(x, x_n) \right)$$

- Отладка работы модели $f(x)$
 - на типичных/нетипичных объектах

Примеры



Максимальное среднее расхождение (MMD)

- Рассмотрим 2 выборки: $X = \{x_i\}_{i=1}^n$ и $Z = \{z_j\}_{j=1}^m$.
 - Распределения X и Z совпадают? Проверка:
- 1 Преобразуем объекты $x \rightarrow \phi(x)$
 - 2 Стат. критерий - максимальное среднее расхождение (maximum mean discrepancy, MMD):

$$\begin{aligned} \|\mathbb{E}_x \phi(x) - \mathbb{E}_z \phi(z)\|^2 &\approx MMD^2[X, Z] = \left\| \frac{1}{n} \sum_{i=1}^n \phi(x_i) - \frac{1}{m} \sum_{j=1}^m \phi(z_j) \right\|^2 = \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{i'=1}^n \phi(x_i)^T \phi(x_{i'}) + \frac{1}{m^2} \sum_{j=1}^m \sum_{j'=1}^m \phi(z_j)^T \phi(z_{j'}) - \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m \phi(x_i)^T \phi(z_j) \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{i'=1}^n K(x_i, x_{i'}) + \frac{1}{m^2} \sum_{j=1}^m \sum_{j'=1}^m K(z_j, z_{j'}) - \frac{2}{nm} \sum_{i=1}^n \sum_{j=1}^m K(x_i, z_j) \end{aligned}$$

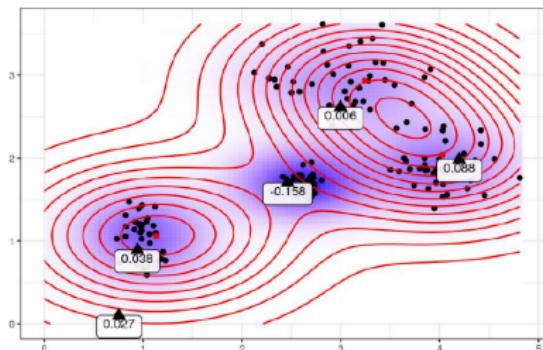
где ядро $K(x, y) = \phi(x)^T \phi(y)$, напр. $K(x, x') = e^{-\gamma \|x-x'\|^2}$.

Выбор прототипов

- Алгоритм выбора прототипов:
 - Стартуем из $Z = \{\}$
 - Пока $|Z| < t$:
 - $\hat{x} = \arg \min_{x \in X} MMD^2 [X, Z \cup \{x\}]$
 - $Z := Z \cup \{\hat{x}\}$
- Критики - объекты, где $witness(x) > t$:

$$witness(x) =$$

$$\frac{1}{n} \sum_{i=1}^n K(x, x_i) - \frac{1}{m} \sum_{i=1}^m K(x, x_i)$$



Содержание

- 1 Интерпретируемые модели
- 2 Важность признаков
- 3 Локальная интерпретация простой моделью
- 4 Графики влияния признаков
- 5 Прототипы и критики
- 6 Контрфактические объяснения

Контрфактические объяснения

- Контрфактическое объяснение (counterfactual explanation) для $(x, f(x))$ - максимально похожий на x объект x' с требуемым откликом y' .
- Примеры:
 - Модель для оценки стоимости аренды квартиры выдает 30К. Какие минимальные изменения в ней или условия аренды внести, чтобы сдавать за 40К?
 - Клиенту не оформили кредит в банке. Какие минимальные условия нужно выполнить, чтобы всё-таки одобрили?
 - либо вероятность одобрения стала выше порога?
- Задача^{8,9,10}:

$$(f(x') - y')^2 + \lambda \rho(x, x') \rightarrow \min_{x'} \iff \begin{cases} (f(x') - y')^2 \rightarrow \min_{x'} \\ \rho(x, x') \leq \varepsilon = F(\lambda) \end{cases}$$

⁸<https://arxiv.org/abs/1711.00399>

⁹Покажите эквивалентность задач из условий Каруша-Куна-Таккера.

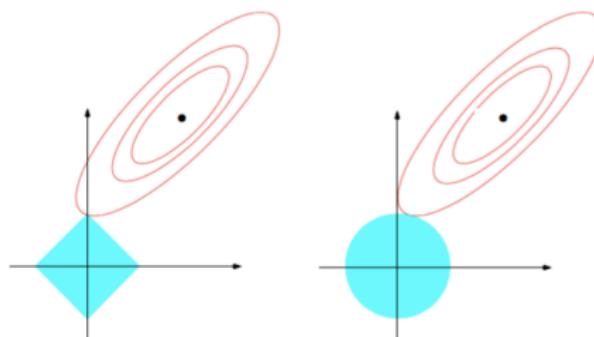
¹⁰ $F(\lambda)$ - возрастающая или убывающая функция?

Выбор расстояния

$$\rho(x, x') = \sum_{d=1}^D \frac{|x^d - x'^d|}{MAD(x^d)}$$

$$MAD(x^d) = \text{median}_{n \in \{1, \dots, N\}} |x_n^d - \text{median}_{n \in \{1, \dots, N\}} (x_n^d)|$$

- Модули отклонений в ρ обеспечивают разреженные изменения в x
 - отличаться в x и \tilde{x} будет малое подмножество признаков



Заключение

Исследовать модель на качеств. уровне можно используя интерпретируемые модели:

- метод близ. центроида, K-NN
- метод наивного Байеса
- лин. регрессия, лог. регрессия
- решающие деревья, решающие правила

Заключение

Исследование black-box моделей:

- вариация отдельного признака:
 - графики PDP, ICE
- числовая важность признаков
 - по дереву решений, перестановочная, значения Шепли
- Локальная интерпретация простой моделью
- Контрфактические объяснения
- Работа модели на прототипах и критиках