Интерпретируемость и объяснимость моделей машинного обучения

Воронцов Константин Вячеславович

д.ф.-м.н., профессор РАН • профессор ВМК МГУ,
руководитель лаборатории машинного обучения
и семантического анализа Института ИИ МГУ,
г.н.с. ФИЦ ИУ РАН, профессор МФТИ

заседание секции №4 Конгресса ИИ «Научная проблематика
в области искусственного интеллекта»
17 августа 2023

Содержание

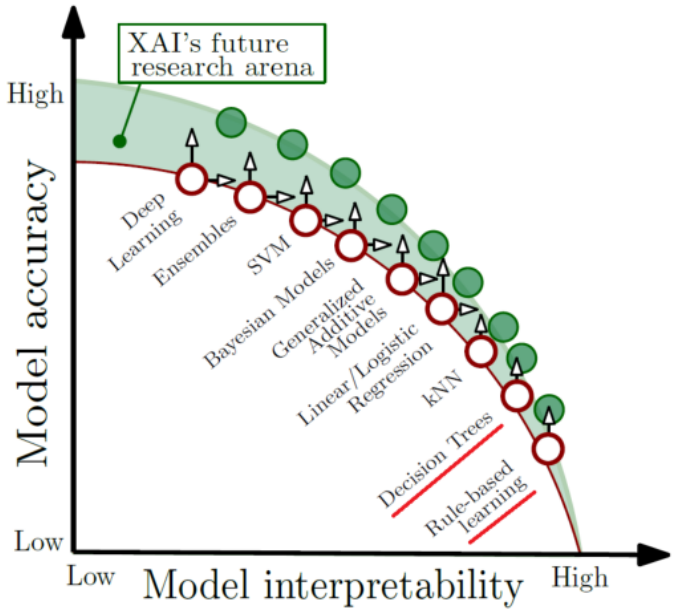
1. **Интерпретируемость и объяснимость**
   - Цели, задачи, основные понятия
   - Интерпретируемые модели (линейные модели)
   - Важность признаков (произвольные модели)
2. **Интерпретация в пространстве признаков**
   - Вектор Шепли
   - Метод LIME
   - Метод SHAP
3. **Интерпретация в пространстве объектов**
   - Вектор Шепли для объектов
   - Метод Gradient Shapley
   - Контрфактическое объяснение

Объяснимость (XAI, eXplainable Artificial Intelligence)

Interpretability — пассивная интерпретируемость устройства модели или предсказания на объекте

Explainability — активная генерация объяснений как дополнительных выходных данных для объекта

Comprehensibility — возможность представить выученные закономерности в виде понятного людям знания



Understandability, **Transparency** — понятность строения модели, её составных частей и промежуточных результатов

"Do you want an interpretable model, or the one that works?"
[Yann LeCun, NIPS'17]

V.Belle, I.Papantonis. Principles and practice of explainable machine learning. 2020

Объяснимость — для кого и зачем

- **Кто:** эксперты предметной области
  Зачем: доверие к моделям, получение знаний из данных
- **Кто:** конечные пользователи
  Зачем: понимание причин принимаемых решений
- **Кто:** регуляторы
  Зачем: аудит соответствия моделей стандартам и нормам
- **Кто:** исследователи, разработчики
  Зачем: понимание свойств моделей, продуктов и сервисов
- **Кто:** бенефициары, менеджеры
  Зачем: понимание влияния моделей на бизнес-процессы

A.B.Arrieta et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. 2019.

Неочевидные проблемы, решаемые с помощью объяснимости

Детекция разладок или сдвигов в данных (data shift)
— наличие дисбалансов в распределениях признаков
— изменение корреляций от выборки к выборке

Нерепрезентативные примеры (out-of-distribution, OOD)
— объекты, которые никогда не встречались при обучении
— намеренно сконструированные атаки на модель

Выявление утечек (data leakage, target leakage)
— KDD-Cup 2008 breast cancer prediction competition:
паразитная корреляция ID пациента с диагнозом на train и test

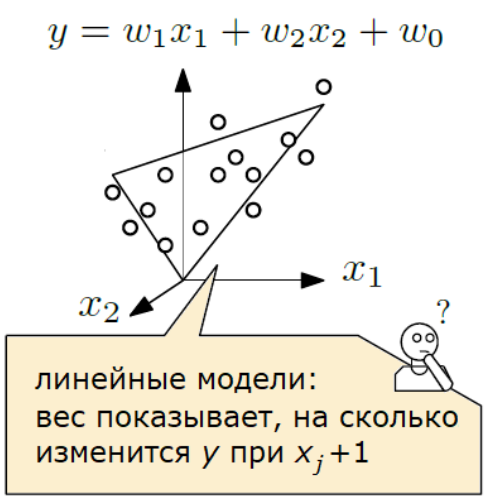
---

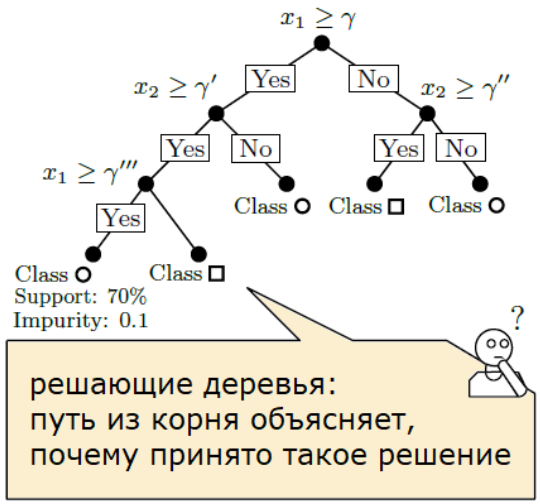
Jingkang Yang, Kaiyang Zhou, Yixuan Li, Ziwei Liu. Generalized Out-of-Distribution Detection: A Survey. 2021

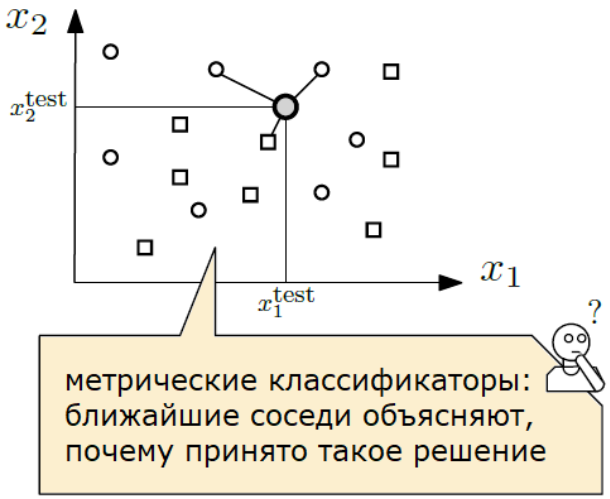
Zheyan Shen, Jiashuo Liu, Yue He, Xingxuan Zhang, Renzhe Xu, Han Yu, Peng Cui. Towards Out-Of-Distribution Generalization: A Survey. 2021

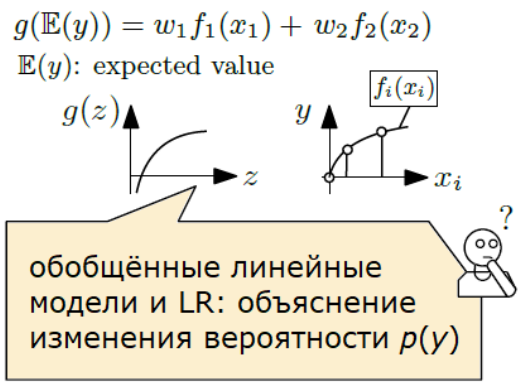
S.Kaufman, S.Rosset, C.Perlich. Leakage in Data Mining: Formulation, Detection, and Avoidance. 2011

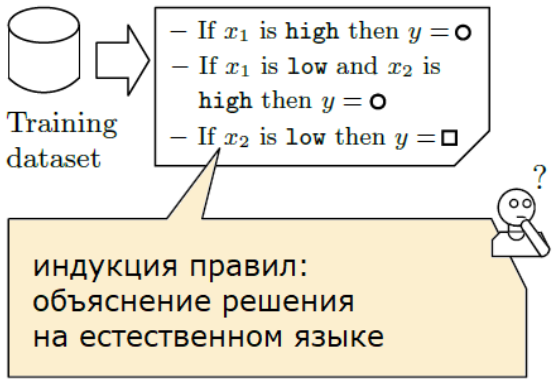
Интерпретируемые модели машинного обучения

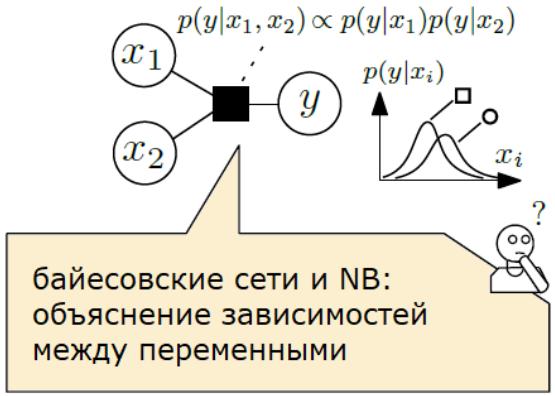












A.B.Arrieta et al. Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. 2019.

Пример. Многомерная линейная регрессия

Модель линейной регрессии на n признаках $f_1(x), \ldots, f_n(x)$:

$$f(x, \alpha) = \sum_{j=1}^{n} \alpha_j f_j(x), \qquad \alpha \in \mathbb{R}^n$$

Метод наименьших квадратов, обучение по выборке $(x_i, y_i)_{i=1}^{\ell}$:

$$Q(\alpha) = \sum_{i=1}^{\ell} \big(f(x_i, \alpha) - y_i\big)^2 = \|F\alpha - y\|^2 \to \min_{\alpha}$$

$\alpha^* = (F^\mathsf{T} F)^{-1} F^\mathsf{T} y$ — решение задачи НК, $F = \big(f_j(x_i)\big)_{\ell \times n}$

Коэффициент детерминации $R^2 \in [0, 1]$, чем выше, тем лучше:

$$R^2 = 1 - \frac{\min_\alpha \|F\alpha - y\|^2}{\min_c \|c - y\|^2} = 1 - \frac{\|F\alpha^* - y\|^2}{\|\bar{y} - y\|^2} = \frac{y^\mathsf{T} F\alpha^* - n\bar{y}^2}{y^\mathsf{T} y - n\bar{y}^2}$$

Оценки значимости признаков в линейной регрессии

- Коэффициент α_j^* равен изменению f при увеличении f_j на 1

— не учитывается масштаб, сдвиг, дисперсия, корреляции, мультиколлинеарность признаков (источник переобучения)

- *t-статистика значимости признака* (feature importance)

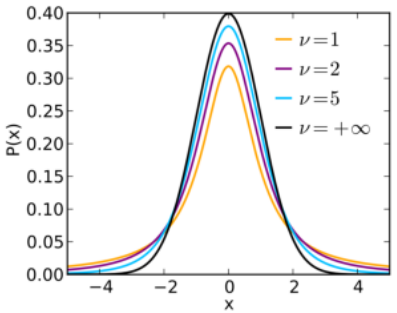
— учитывает дисперсию оценки α_j^*:

$$T_j = \frac{\alpha_j^*}{\hat{\sigma}\sqrt{(F^\top F)^{-1}_{jj}}} \sim t_{\ell-n}, \quad \hat{\sigma}^2 = \frac{Q(\alpha^*)}{\ell - n}$$

— позволяет проверять гипотезу $\alpha_j^* = 0$,

— вычислять p-value для этой гипотезы,

— доверительные интервалы для α_j^*.



t-распределение Стьюдента с $\nu = \ell - n$ степенями свободы

- *Чистый эффект* (net effect) NEF_j признака в разложении R^2:

$$R^2 = y^\top F \alpha^* = \sum_{j=1}^{n} \alpha_j^* (f_j^\top y) = \sum_{j=1}^{n} \mathrm{NEF}_j \quad (\text{при } y^\top y = 1,\ \bar{y} = 0)$$

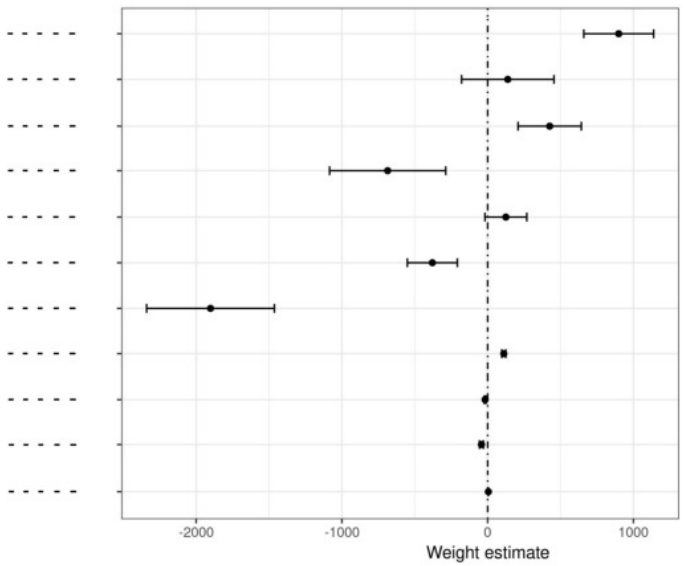
Пример. Задача прогнозирования аренды велосипедов

x_i — дата, y_i — число арендованных велосипедов

Weight $= \alpha_j^*$; Standard Error SE $= \hat{\sigma}\sqrt{(F^\top F)^{-1}_{jj}}$; t $= |T_j|$

Intercept — свободный член, коэффициент при признаке $f_1 = 1$

	Weight	SE	t
(Intercept)	2399.4	238.3	10.1
season SUMMER	899.3	122.3	7.4
season FALL	138.2	161.7	0.9
season WINTER	425.6	110.8	3.8
holiday	-686.1	203.3	3.4
workingday	124.9	73.3	1.7
weathersit MISTY	-379.4	87.6	4.3
weathersit RAIN/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



UCI ML Repo: `http://archive.ics.uci.edu/ml/datasets/Bike+Sharing+Dataset`

Christoph Molnar. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2019

Перестановочные оценки значимости признаков

Перестановочная оценка PFI (permutational feature importance)

$$\text{PFI}_j = Q^j/Q \text{ или } Q^j - Q$$

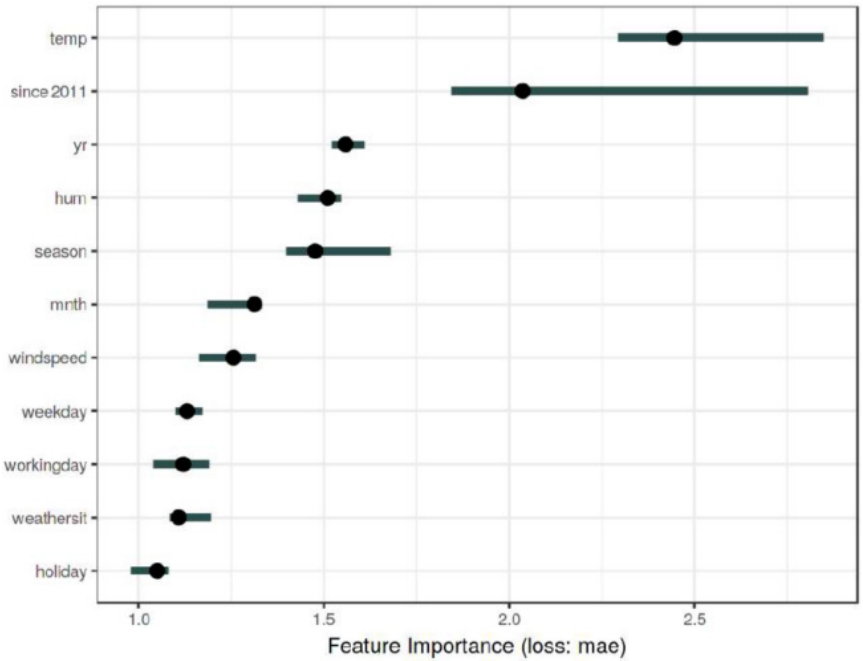
потери на исходной выборке:

$$Q = \sum_i \mathscr{L}\big(f(x_i), y_i\big)$$

потери после перемешивания:

$$Q^j = \sum_i \mathscr{L}\big(f(\tilde{x}_i^j), y_i\big)$$

где $f(x)$ — обученная модель,
$\mathscr{L}(f, y)$ — функция потерь,
$\tilde{x}_i^j = \text{замена}\big(f_j(x_i) \to f_j(x_{\text{rand}})\big)$.



⊕ любая модель ⊕ однократное обучение ⊕ учёт корреляций
⊖ перемешивание может порождать нереалистичные объекты

Christoph Molnar. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2019

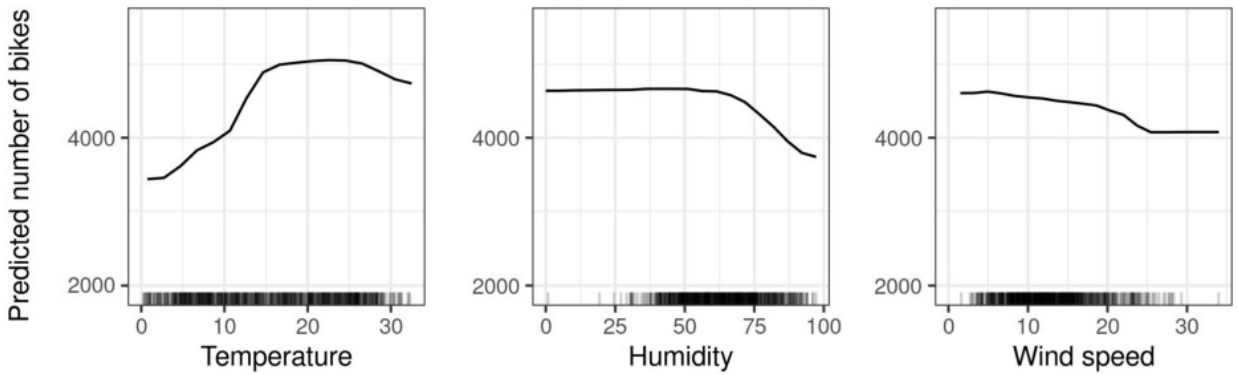
Графики частичной зависимости (Partial Dependence Plot, PDP)

Как модель $f(x)$ зависит от части признаков $S \subseteq \{f_1, \ldots, f_n\}$?
$x = (x_S, \tilde{x})$, x_S — признаки из S, $\tilde{x}$ — остальные признаки.

Оценивание математического ожидания методом Монте-Карло:

$$g(x_S) = \mathsf{E}_{\tilde{x}} f(x_S, \tilde{x}) = \int f(x_S, \tilde{x}) dP(\tilde{x}|x_S)$$

$$\hat{g}(x_S) = \frac{1}{\ell} \sum_{i=1}^{\ell} f(x_S, \tilde{x}_i) \quad \text{или} \quad \hat{g}(x_S) = \frac{\sum_i K(x_S, x_{Si}) f(x_S, \tilde{x}_i)}{\sum_i K(x_S, x_{Si})}$$



Christoph Molnar. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2019

Вектор Шепли (из теории кооперативных игр)

Признаки $F = \{f_1, \ldots, f_n\}$ играют в кооперативную игру
$V(S)$ — совместный выигрыш коалиции $S \subseteq F$, $V(\varnothing) = 0$

Игроки вступают в S по очереди, задаваемой перестановкой π
$\Delta(j, S) = V(S \cup j) - V(S)$ — полезность игрока f_j в коалиции S
$S_{\pi j} \subset F$ — множество игроков, идущих перед f_j в перестановке π

Вектор Шепли ϕ — справедливый делёж общего выигрыша:

$$\phi_j = \frac{1}{n!} \sum_{\pi} \Delta(j, S_{\pi j}) = \sum_{S} \frac{|S|!\,(n - |S| - 1)!}{n!} \Delta(j, S)$$

$|S|!$ — число способов образовать коалицию S
$(n - |S| - 1)!$ — число способов продолжить образование коалиции после присоединения f_j к S
$n!$ — число перестановок π множества n игроков

Lloyd Stowell Shapley. A value for n-person games. 1952

Свойства вектора Шепли

Теорема

Это единственный способ делёжа, удовлетворяющий аксиомам:

1. эффективность:
   $\sum_{j=1}^{n} \phi_j = V(F)$
2. симметричность (анонимность игроков):
   $\forall S, j, k \ \Delta(j, S) = \Delta(k, S) \quad \Rightarrow \quad \phi_j = \phi_k$
3. невозможность халявы для «болвана»:
   $\forall S, j \ \Delta(j, S) = 0 \quad \Rightarrow \quad \phi_j = 0$
4. состоятельность:
   $\forall S, j \ \Delta_1(j, S) \leqslant \Delta_2(j, S) \quad \Rightarrow \quad \phi_{1j} \leqslant \phi_{2j}$
5. аддитивность:
   $\forall S \ V(S) = \alpha_1 V_1(S) + \alpha_2 V_2(S) \ \Rightarrow \ \forall j \ \phi_j = \alpha_1 \phi_{1j} + \alpha_2 \phi_{2j}$

Lloyd Stowell Shapley. A value for n-person games. 1952

Оценивание вектора Шепли методом Монте-Карло

Π — случайное подмножество перестановок; для каждой $\pi \in \Pi$ в модель инкрементно добавляются признаки $\pi(j)$, $j = 1, \ldots, n$:

$$\hat{\phi}_j = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \Delta(j, S_{\pi j})$$

Варианты определения «выигрыша» $V(S)$, $S \subseteq \{f_1, \ldots, f_n\}$:

- *Коэффициент детерминации* $V(S) = R^2_S$ линейной модели, модель дообучается при добавлении каждого признака
- *Shapley regression value* $V(S) = f_S(x)$ на объекте x, где модель f_S обучена только на признаках из S
- *Shapley sampling value* $V(S) = \mathrm{E}_{\tilde{x}} f(x_S, \tilde{x})$, где $x = (x_S, \tilde{x})$ $\mathrm{E}_{\tilde{x}}$ — среднее по объектам выборки $x_i = (x_{Si}, \tilde{x}_i)$: $x_{Si} \approx x_S$
- *Shapley loss* $V(S) = -\mathrm{E}_{\tilde{x}} \mathscr{L}\big(f(x_S, \tilde{x}), y\big)$ — функция потерь

E.Štrumbelj, I.Kononenko. Explaining prediction models and individual predictions with feature contributions. 2014

Суррогатное моделирование в окрестности объекта x

$(x_i, y_i)_{i=1}^{\ell}$ — обучающая выборка, $\mathscr{L}(f, y)$ — функция потерь
$f(x, \alpha)$ — неинтерпретируемая модель, обученная по выборке:

$$\sum_{i=1}^{\ell} \mathscr{L}\big(f(x_i, \alpha), y_i\big) \to \min_{\alpha}$$

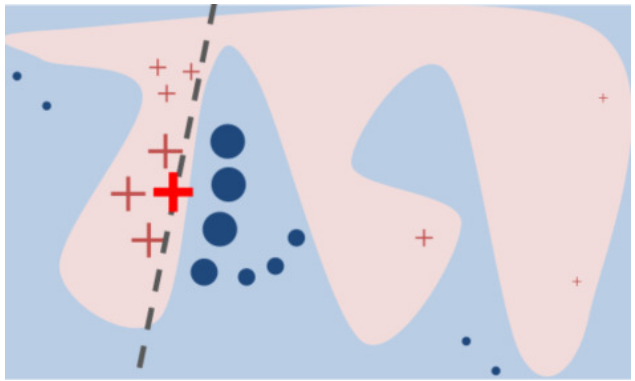
$g_x(z, \beta)$ — интерпретируемая *суррогатная модель* для аппроксимации f в окрестности объясняемого объекта x:

$$\sum_{i=1}^{k} w_{xi}\, \mathscr{L}\big(g_x(z_i, \beta), f(z_i, \alpha)\big) + \Omega(\beta) \to \min_{\beta}$$

$(z_i)_{i=1}^{k} \sim \pi\big(K_h(z, x)\big)$ — *суррогатные объекты*, сэмплируемые из радиального распределения с центром в x и радиусом h
$w_{xi} = K_h(z, x)$ — веса объектов в h-окрестности объекта x
$K_h(z, x)$ — функция близости (kernel) радиуса h
$\Omega(\beta)$ — регуляризатор, штраф за сложность модели $g_x(z, \beta)$

Метод LIME (Local Interpretable Model-agnostic Explanations)

$g_x(z,\beta) = \sum_{j=1}^{m} \beta_j b_j(z)$ — локальная линейная аппроксимация



1. фиксируется объект x, для которого требуется объяснение
2. синтезируются *суррогатные объекты* z_i в его окрестности
3. на них вычисляются значения основной модели $f(z_i, \alpha)$
4. строится локальная аппроксимация *суррогатной моделью*
5. для объекта x строится объяснение и его визуализация

M.Ribeiro, S.Singh, C.Guestrin. "Why should I trust you?" Explaining the predictions of any classifier. 2016

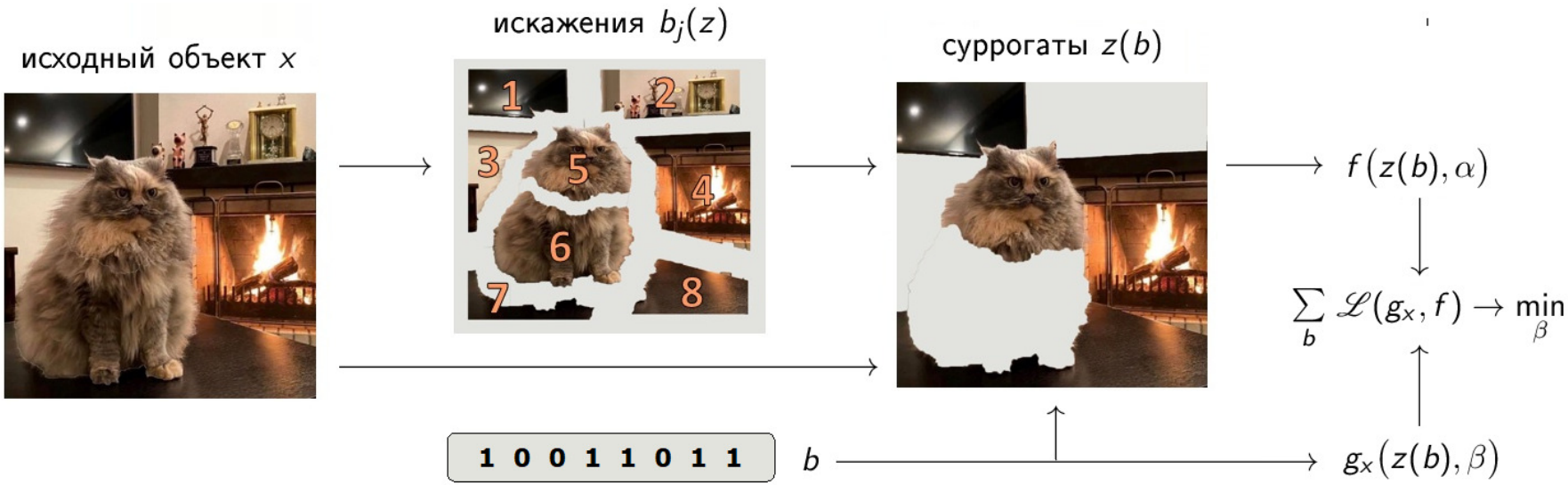
Метод LIME: синтез суррогатных объектов

$g_x(z,\beta) = \sum_{j=1}^{m} \beta_j b_j(z)$ — локальная линейная аппроксимация

Признаки $b_j(z) = [j\text{-го искажения объекта } x \text{ в суррогате } z \text{ нет}]$

Синтез суррогата $z(b) =$ применить к x все искажения j: $b_j = 0$

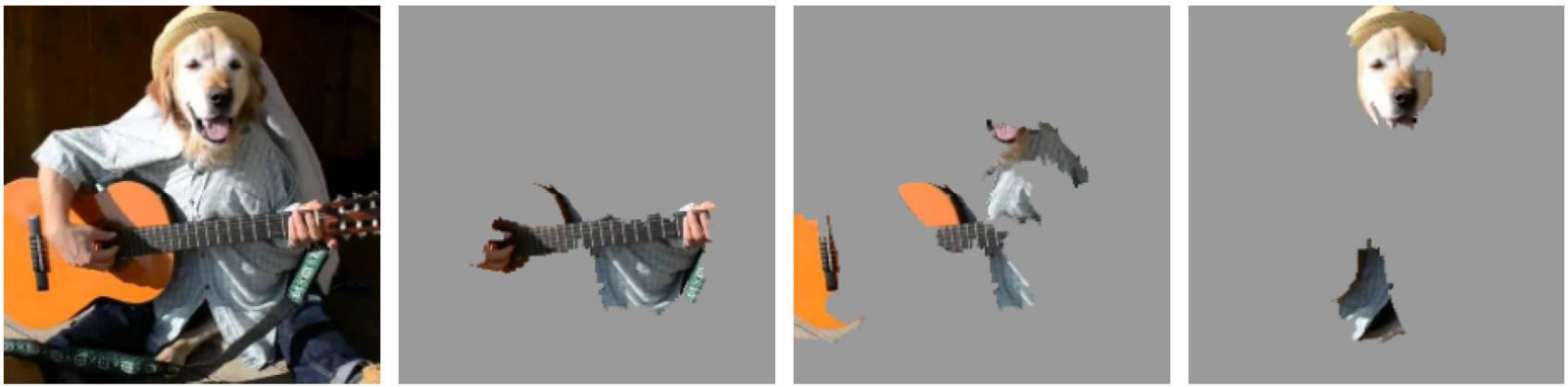
Синтез выборки суррогатов $(z_i)_{i=1}^{k}$ — по случайным $b_j \in \{0,1\}$



Олег Седухин. Интерпретация моделей и диагностика сдвига данных: LIME, SHAP и Shapley Flow. 2022-01-13. https://habr.com/ru/companies/ods/articles/599573

Пример LIME. Задача классификации изображений

Признаки $b_j(z)_{j=1}^m$ — сегменты (super-pixel) из изображения x
$m = 10$, признаки конструируются под объясняемый объект x



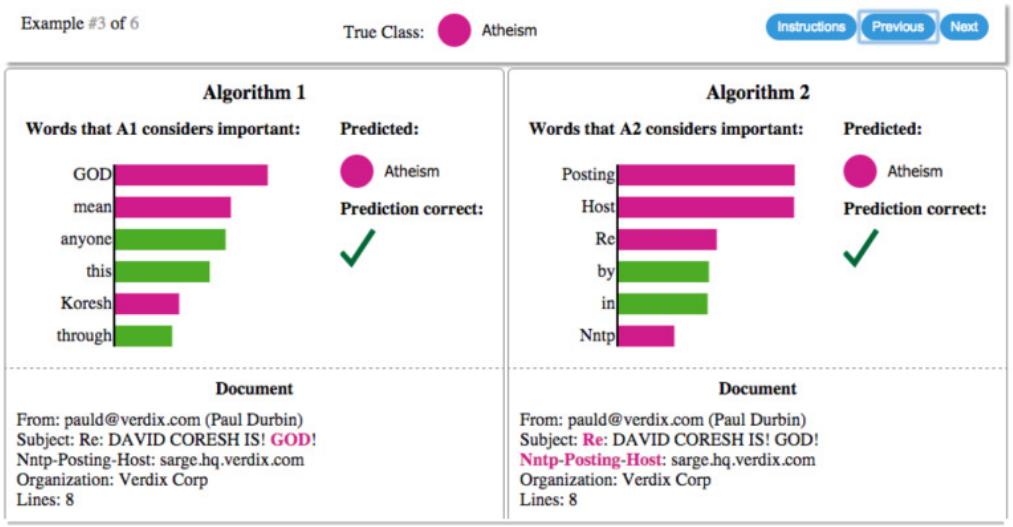
(a) Original Image (b) Explaining *Electric guitar* (c) Explaining *Acoustic guitar* (d) Explaining *Labrador*

Модель классификации — глубокая нейросеть Google Inception
Три наиболее вероятных класса: «electric guitar» ($p = 0.32$), «acoustic guitar» ($p = 0.24$), «labrador» ($p = 0.21$)

Ясно, почему модель перепутала «acoustic» с «electric»
— из-за грифа, см. рис. (b)

Пример LIME. Задача классификации текстов (20NewsGroups)

Признаки $b_j(z) = [\text{наличие слова } j \text{ из текста } x \text{ в тексте } z]$
Гистограмма весов β_j: важности слов j для исходного текста x



Модель классификации SVM-RBF имеет точность 94% на тесте, но при различении классов «christianity» и «atheism» считает важными мусорные слова «Posting», «Host», «Re».

Ясно, в чём проблема, и как её исправлять (фильтровать слова)

Метод SHAP (SHapley Additive exPlanations)

$g_x(z,\beta) = \sum_{j=1}^{m} \beta_j b_j(z)$ — локальная линейная аппроксимация

Признаки $b_j(z) = [j\text{-го искажения объекта } x \text{ в суррогате } z \text{ нет}]$

Синтез суррогата $z(b) =$ применить к x все искажения j: $b_j = 0$

Приращение $f(z,\alpha)$, если в суррогате $z(b)$ убрать искажение j:
$\Delta(j,b) = V(b|_{b_j=1}) - V(b|_{b_j=0})$, «выигрыш» $V(b) = f\big(z(b),\alpha\big)$

Три желательных свойства локальной модели $g_x(z,\beta)$

1. локальная согласованность аппроксимации в точке x:
   $\forall j\ b_j(z) = 1 \quad \Rightarrow \quad g_x(z,\beta) = f(x,\alpha)$
2. бесполезность болвана — признака b_j, пропущенного в x:
   $\forall j\ b_j(x) = 0 \quad \Rightarrow \quad \beta_j = 0$
3. состоятельность: с ростом приращения $\Delta(j,b)$ растёт β_j,
   $\forall b,j\ \Delta_1(j,b) \leqslant \Delta_2(j,b) \quad \Rightarrow \quad \beta_{1j} \leqslant \beta_{2j}$

Метод SHAP: теоретическое обоснование

Теорема 1

Единственным распределением весов β_j, удовлетворяющим свойствам 1 2 3 является вектор Шепли:

$$\beta_j = \sum_{b \in \{0,1\}^m} \frac{|b|!\,(m - |b| - 1)!}{m!}\,\Delta(j, b)$$

где $|b| = \{j\colon b_j = 1\}$ — число единиц в векторе b.

Теорема 2 (метод Shapley Kernel)

Вектор Шепли (β_j) является решением задачи НК с весами:

$$\sum_{b \in \{0,1\}^m} w_b \big(g_x(z(b), \beta) - f(z(b), \alpha)\big)^2 \to \min_{\beta}$$

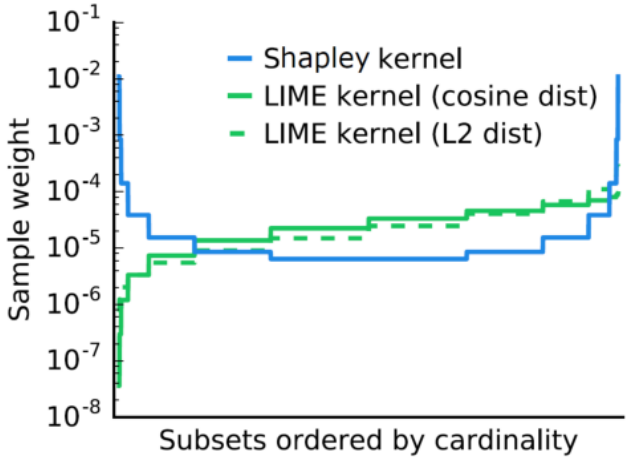
где w_b — веса 2^m суррогатов, $w_b = \frac{|b|!\,(m-|b|-1)!}{m!}\,\frac{m-1}{|b|} = \frac{1}{m C_{m-2}^{|b|-1}}$

Метод Shapley Kernel: вариант реализации SHAP

⊕ Вектор Шепли (β_j) вычисляется взвешенной линейной регрессией

⊕ LIME решает ту же задачу, но веса суррогатов w_b задаются эвристически, неоптимально

⊕ При больших 2^m векторы b можно сэмплировать из распределения w_b



⊕ SHAP лучше LIME в экспериментах, где они сравнивались с тем, как эксперты объясняют решения моделей

⊖ SHAP и LIME оба оценивают значимость признаков по нереалистичным (out-of-distribution) суррогатным объектам

Scott Lundberg, Su-In Lee. A unified approach to interpreting model predictions. 2017

E.Kumar, S.Venkatasubramanian, C.Scheidegger, S.A.Friedler. Problems with Shapley-value-based explanations as feature importance measures. 2020

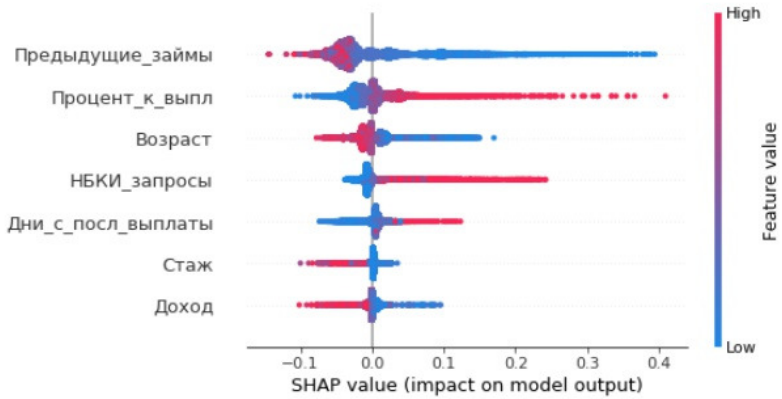
Метод SHAP: пример визуализации

Модель вероятности дефолта $f(x)$, градиентный бустинг

Индивидуальное объяснение для x: $f(x) = 19\%$ при $\bar{y} = 6\%$

Значения Шепли показываются цветом: $\beta_j(x) < 0$, $\beta_j(x) > 0$

Агрегированные объяснения по всей выборке $\{\beta_j(x_i)\}$:



ось X: $\beta_j(x_i)$

ось Y: признаки j

цвет точки: значение признака $f_j(x_i)$

ширина линии $\propto$ число точек

---

https://rb.ru/opinion/uzhe-ne-black-box

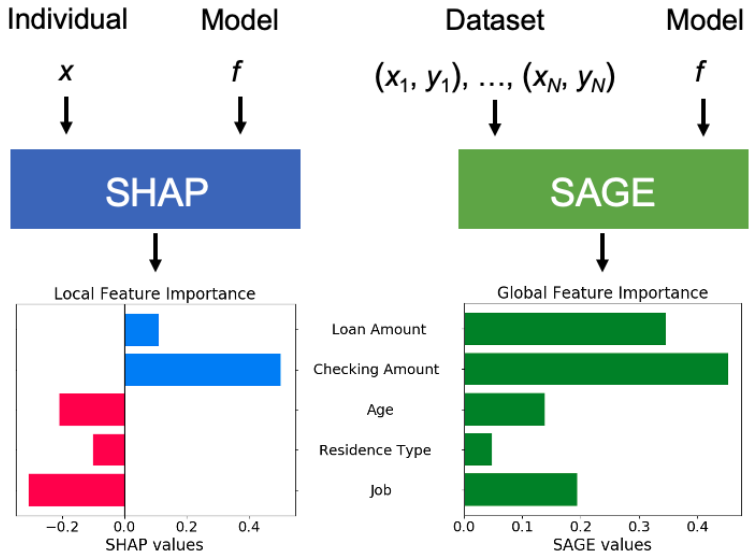
Метод SAGE (Shapley Additive Global importancE)

SHAP:

каковы вклады признаков f_j
в предсказание $f(x)$

SAGE:

как качество модели в целом
зависит от признаков f_j



Модификация SHAP → SAGE:

$V(S) = -\mathscr{L}\left(\mathrm{E}_{\tilde{x}} f(x_S, \tilde{x})\right)$ — раскладываются потери (LossSHAP)

$\phi_j = \frac{1}{\ell}\sum_{i=1}^{\ell} \phi_j(x_i)$ — значения Шепли усредняются по выборке

$\phi_j = \frac{1}{|X|}\sum_{x_i \in X} \phi_j(x_i)$ — или по случайной подвыборке, если долго

Ian C. Covert, Scott Lundberg, Su-In Lee. Understanding Global Feature Contributions With Additive Importance Measures. 2020

Вектор Шепли для объектов: инкрементное обучение

Теперь обучающие объекты играют в кооперативную игру:
$f_s(x)$ — модель, обученная на подвыборке $S \subseteq \{x_1, \ldots, x_\ell\}$
$V(S) = -\sum_x \mathscr{L}(f_s(x))$ на тестовых объектах x (hold-out)
$\Delta(i, S) = V(S \cup i) - V(S)$ — полезность обучающего объекта x_i
$\phi_i = \frac{1}{|\Pi|} \sum_{\pi \in \Pi} \Delta(i, S_{\pi i})$ — несмещённая оценка Монте-Карло

для всех перестановок $\pi_t \in \Pi$, $t = 1, \ldots, |\Pi|$:
    $S := \varnothing$; $v_0 := V(\varnothing)$;
    для всех $i = \pi_t(1), \ldots, \pi_t(\ell)$:
        $S := S \cup \{x_i\}$;
        обновить модель $f_s(x)$, дообучив её на объекте x_i;
        оценить модель $v_i := V(S)$;
        $\phi_i := \frac{t-1}{t}\phi_i + \frac{1}{t}(v_i - v_{i-1})$;

A.Ghorbani, J.Zou. Data Shapley: equitable valuation of data for machine learning. 2019

Встраивание оценок Шепли в онлайновый градиентный спуск

Градиентная минимизация аддитивного критерия:

$$\sum_{i=1}^{\ell} \mathscr{L}\big(f(x_i, \alpha), y_i\big) \to \max_{\alpha}$$

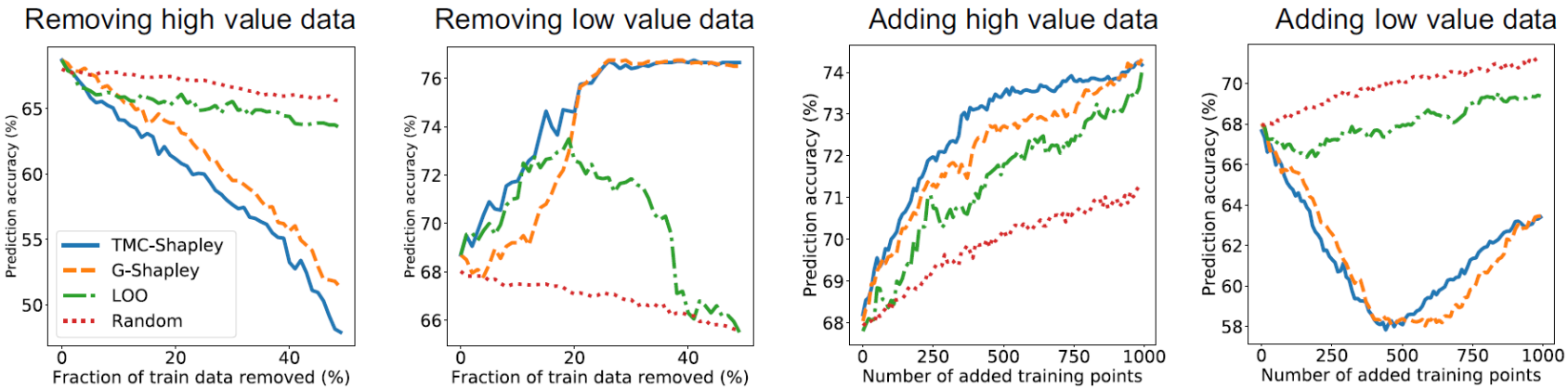
Алгоритм инкрементного обучения Online Gradient Descent:

для всех перестановок $\pi_t \in \Pi$, $t = 1, \ldots, |\Pi|$:
  $S := \varnothing$; $v_0 := V(\varnothing)$; инициализировать α_0;
  для всех $i = \pi_t(1), \ldots, \pi_t(\ell)$:
    $S := S \cup \{x_i\}$;
    обновить модель $\alpha_i := \alpha_{i-1} - \eta_i \nabla_\alpha \mathscr{L}\big(f(x_i, \alpha_{i-1}), y_i\big)$;
    оценить модель $v_i := V(S)$;
    $\phi_i := \frac{t-1}{t}\phi_i + \frac{1}{t}(v_i - v_{i-1})$;

A.Ghorbani, J.Zou. Data Shapley: equitable valuation of data for machine learning. 2019
M.Zinkevich. Online convex programming and generalized infinitesimal gradient ascent. 2003.

Интерпретация объектов с помощью значений Шепли

- низкое ϕ_i — выбросы, такие x_i можно удалять из выборки
- высокое ϕ_i — опорные, пограничные, таких x_i не хватает
- более устойчивая оценка по сравнению с leave-one-out



Задача UCI:BreastCancer

(1) изъятие из обучения лучших объектов, по убыванию ϕ_i
(2) изъятие из обучения худших объектов, по возрастанию ϕ_i
(3) добавление объектов, похожих на лучшие, по убыванию ϕ_i
(4) добавление объектов, похожих на худшие, по возрастанию ϕ_i

---

A.Ghorbani, J.Zou. Data Shapley: equitable valuation of data for machine learning. 2019

Задача поиска контрфактов

Контрфакт x' — объект, схожий с x, но существенно отличающийся предсказанием модели $f(x', \alpha^*)$.

- Модель кредитного скоринга выдала отказ.
  Какие изменения признаков поменяют решение модели?
  (закрыть другие кредиты? переехать в другой город?
  сменить работу? изменить структуру расходов?)
- Модель оценила для собственника стоимость аренды.
  Какие факторы способны увеличить оценку стоимости?
  (улучшить ремонт? разрешить домашних животных?)

Важно: находить реализуемые изменения признаков:

- минимально изменять минимальное число признаков
- выбирать из множества разнообразных контрфактов

Riccardo Guidotti. Counterfactual explanations and how to find them: literature review and benchmarking. 2022

Метод поиска контрфактов (Counterfactual explanations)

Контрфакт x' — объект, схожий с x, но существенно отличающийся предсказанием модели $f(x', \alpha)$.

Оптимизационная задача поиска контрфакта x' при $y' \neq y(x)$:

$$\mathscr{L}\big(f(x', \alpha), y'\big) + \lambda \|x - x'\|_1 \to \min_{x'} \min_{\lambda}$$

L_1-регуляризатор обеспечивает разреженность решения — чем больше λ, тем больше совпадений признаков $x_j = x'_j$:

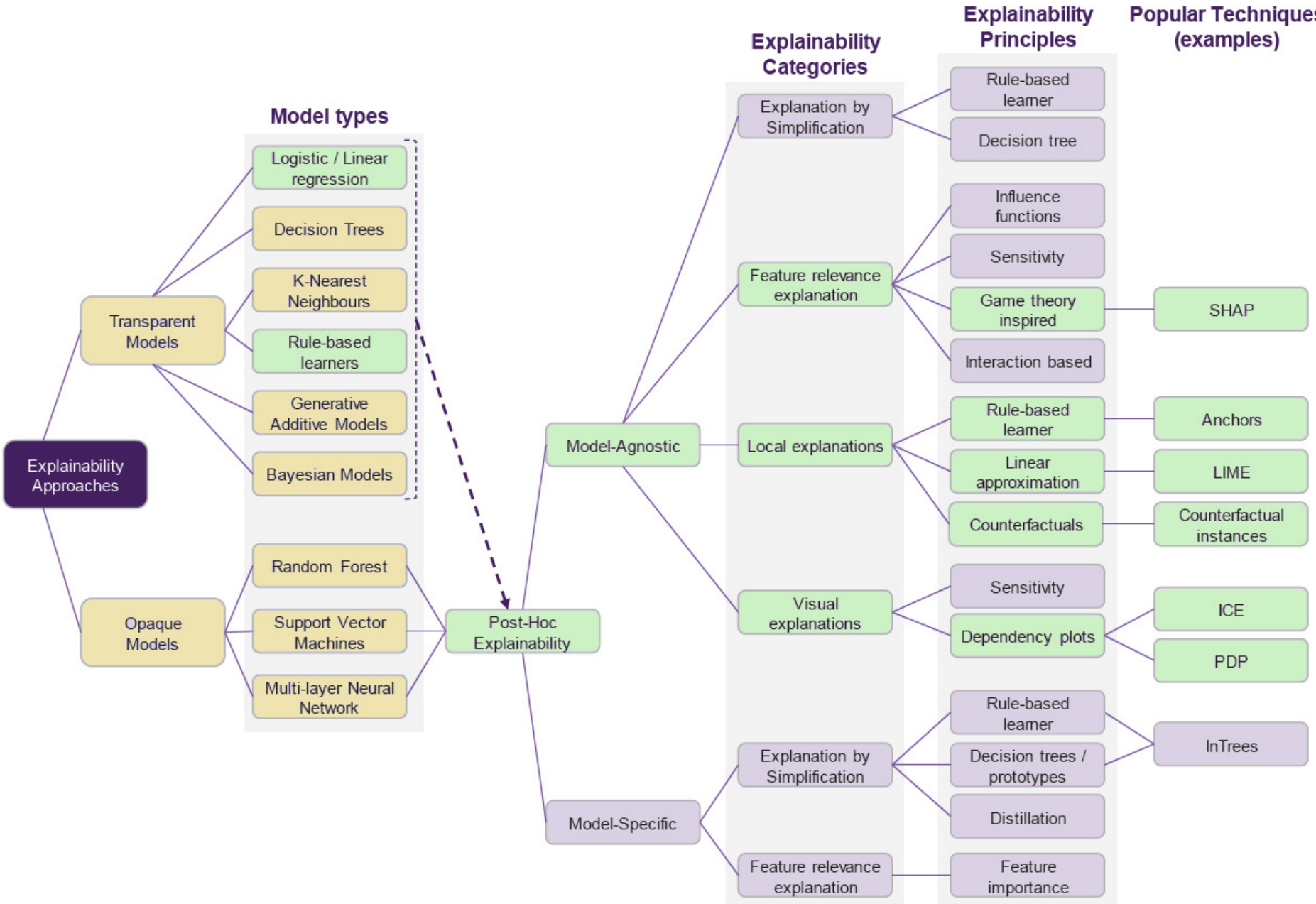
$$\|x - x'\|_1 = \sum_{j=1}^{n} \frac{|x_j - x'_j|}{\mathrm{MAD}_j},$$

Median Absolute Deviation $\mathrm{MAD}_j = \underset{i}{\mathrm{med}}\, |x_{ij} - M_j|, \quad M_j = \underset{i}{\mathrm{med}}\, x_{ij}$

Постепенно уменьшая λ, подгоняем $f(x', \alpha) \to y'$.

S.Wachter, B.Mittelstadt, C.Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. 2017

Подходы к объяснимости моделей машинного обучения



V.Belle, I.Papantonis. Principles and practice of explainable machine learning. 2020

Резюме

- *Интерпретируемость* — прозрачность строения модели, либо понятность её результата на объекте
- Интерпретируемых моделей не много:
  линейные (MVLR, LR, GAM, GLM), логические (DT, RI), метрические (kNN, PW, RBF), байесовские (NB, BN)
- *Объяснимость* решения на объекте — как правило, с помощью интерпретируемой *суррогатной модели*
- *Вектор Шепли* оценивает индивидуальные значимости
  — игроков-признаков по успешности их коалиций
  — игроков-объектов по успешности обучающих выборок
- SHAP, SAGE — наиболее продвинутые методы объяснения

К.В.Воронцов. Машинное обучение (курс лекций, К.В.Воронцов).
http://www.machinelearning.ru

P.Linardatos, V.Papastefanopoulos, S.Kotsiantis. Explainable AI: A Review of Machine Learning Interpretability Methods. 2021

Zachary C. Lipton. The Mythos of Model Interpretability. 2018

Christoph Molnar. Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. 2019