Локальные методы интерпретации (одно наблюдение)

LIME

Local interpreeable Model-agnostic explanations

1) Выбрать наблюдение x
2) Сгенерировать наблюдения из окрестности x
3) Взвешиваем наблюдения согласно близости к x.
4) На получившемся датасете строим интерпретируемую модель
5) Интерпретация локальной модели

Плюсы

1) Независимость от модели
2) Универсальность (таблицы, тексты, картинки)
3) Простота

Минусы:

1) Непонятно как семплить и в какой области
2) Семпл из нормального
3) Нестабильность
4) Легко манипулировать

Shapley Values

$$\hat{f}(x) = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p$$

$$\phi_j(f) = \beta_j x_j - E(\beta_j X_j) = \beta_j x_j - \beta_j E(X_j)$$

$$\sum_{j=1}^{p} \phi_j(\hat{f}) = \sum_{j=1}^{p} (\beta_j x_j - E(\beta_j X_j)) =$$

$$= (\beta_0 + \sum \beta_j x_j) - (\beta_0 + \sum E(\beta_j X_j)) =$$

$$= \hat{f}(x) - E(\hat{f}(X))$$

S - мн-во интересующих нас фичей

$$\phi_j(val) = \sum_{S \subseteq \{1,\dots,p\} \setminus j} \frac{|S|!\,(p-|S|-1)!}{p!} (val(S \cup \{j\}) - val(S))$$

$$Val_x(S) = \int \hat{f}(x_1, \dots, x_p)\, d P_{x \notin S} - E_X(\hat{f}(X))$$

и фичи $S = \{x_1, x_3\}$

$$Val_x(S) = Val_x(\{1,3\}) =$$

$$= \int_R \int_R \hat{f}(x_1, X_2, x_3, X_4)\, d\, P_{X_2, X_4} - E_X(\hat{f}(X))$$

$$Val(S \cup \{j\}) = Val(S) \Rightarrow \phi_j = 0$$

$$\phi_j = \phi_j^1 + \phi_j^2$$

Вычисление.

M – кол-во итераций

x – наблюдение, X – выборка, f – модель

for m in range(M):

1) Выбрать z из X рандомно
2) Выбирается случайная перестановка фичей

$$x_0 = (x_{(1)}, \ldots, x_{(j)}, \ldots, x_{(p)})$$

$$z_0 = (z_{(1)}, \ldots, z_{(j)}, \ldots, z_{(p)})$$

3) Создаём два новых наблюдения

1) $x_{+j} = (x_{(1)}, \dots, x_{(j-1)}, x_{(j)}, z_{(j+1)}, \dots, z_{(p)})$

2) $x_{-j} = (x_{(1)}, \dots, x_{(j-1)}, z_{(j)}, z_{(j+1)}, \dots, z_{(p)})$

$\phi_j^m = \hat{f}(x_{+j}) - \hat{f}(x_{-j})$

$\phi_j(x) = \frac{1}{M} \sum_{m=1}^{M} \phi_j^m$

порядок восстановлен.

Плюсы:

1) Самый обоснов. метод

Минусы:

1) Долго → Всегда использует все фичи
2) Выдаёт одно значение
3) Нужен доступ к данным
4) Нереалистичные данные

SHAP

Shapley Additive Explanations

$z' \in \{0,1\}^M$ M – max coalition size

$$g(z') = \phi_0 + \sum_{j=1}^{M} \phi_j z'_j$$

x – все эл-ты $z' = 1$

Kernel SHAP

1) Сэмплируем коалиции $z'_k \in \{0,1\}^M$, $k \in \{1, \dots, K\}$

2) Переведем коалиции в строки $h_x(z'_k)$

3) Считаем веса наблюд. через SHAP-kernel

4) Фитим линрег.

5) Коэфф-ты это наши ϕ