SHAP, LIME

Интерпретируемость

Мир → Данные → Модель → Интерпр. → Profit!

1) Модель интерпретируема
   а) Линрег → Коэф.

$$\hat{\beta}_j \;,\; \hat{\beta}_j x_{ij}$$

$$y_i = \sum_j \beta_j x_{ij} + \beta_0$$

   б) Логрег $\frac{\partial P}{\partial x_j}$, отношение шансов
   в) Деревья, разбиения

Model-agnostic methods

1) Global 2) Local

Global models

PDP, ALE, Global surrogate

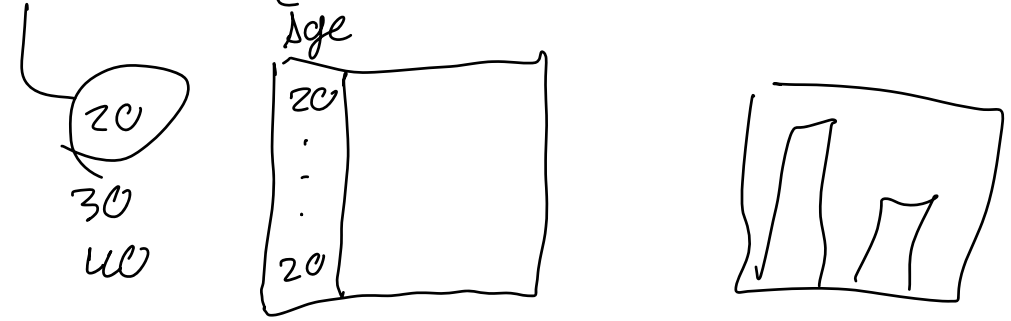
PDP - partial dependence plot

$\hat{f}$ – наша модель

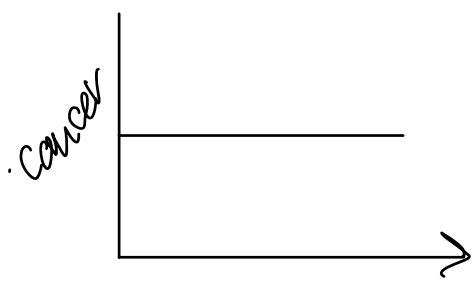
S – мн-во фичей, которые мы рассматрив.

C – мн-во остальных фичей

$$\hat{f}_S(x_S) = E_{X_C}[\hat{f}(x_S, X_C)] = \int \hat{f}(x_S, x_C)\, dP(X_C)$$



$$\hat{f}_S(x_S) \approx \frac{1}{n}\sum_{i=1}^{n} \hat{f}(x_S, x_C^{(i)})$$



$$I(x_S) = \sqrt{\frac{1}{K-1}\sum_{k=1}^{K}\left(\hat{f}_S(x_S^k) - \frac{1}{K}\sum_{k=1}^{K}\hat{f}_S(x_S^{(k)})\right)^2}$$

кол-во уник значений x_S

$$I(x_S) = \max_k(f_S(x_S^{(k)})) - \min_k(f_S(x_S^{(k)}))$$ /.n_cats

ALE -
Accumulated Local Effects

PDP:

$$\hat{f}_{s,PDP}(x_s) = E_{x_c}[\hat{f}(x_s, X_c)]$$

M-plot

$$\hat{f}_{s,M}(x_s) = E_{X_c|X_s}[f(X_s, X_c) | X_s = x_s]$$

ALE:

$$\hat{f}_{s,ALE}(x_s) = \int_{z_{0,s}}^{x_s} E_{X_c|X_s=z_s}[\hat{f}^s(X_s, X_c) | X_s = z_s] dz_s$$

+ Несмещённость
+ Быстрее считать
+ Интерпретация
+ Декомпозиция

$f = f_1 + f_2 \ldots$

- Неустойчивость
- Не декомпозируется (...)

- сложнее интерпретировать

Global surrogate

$f(x)$ – целевая фо-я (black box)

$\hat{f}(x) = g(x)$

1) Получить прогнозы на всей выборке $\hat{f}(x)$
2) Выберите интерпр. модель $g(x)$
3) Оценим $\hat{f}(x) = g(x)$
4) Оценить качество $g(x)$
5) Интерпретировать $g(x)$

$$R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum_{i=1}^{n} (\hat{y}_g^{(i)} - \hat{y}_f^{(i)})^2}{\sum_{i=1}^{n} (\hat{y}_f^{(i)} - \bar{\hat{y}}_f^{(i)})^2}$$

+ Гибкость
+ Интуитивность
+ R^2 просто интерпретировать
- Нечем cut-off для R^2
- Все минусы модели g

Local

ICE-plots