НЕТОЛОГИЯ
групп

Блок

FEATURE ENGINEERING

Константин Гусев

Старший аналитик-моделист

Bi.zone (кибер-безопасность)

Ex-аналитик McKinsey & Co.

ks.gusev@physics.msu.ru

ksgusev

ЦЕЛИ ЗАНЯТИЯ

ЦЕЛИ ЗАНЯТИЯ

В КОНЦЕ ЗАНЯТИЯ ВЫ СМОЖЕТЕ

1

Уменьшать размерность пространство с помощью Lasso регрессии

2

Сжимать пространство признаков с помощью Ridge регрессии

3

Понимать почему происходит фильтрация признаков в Tree методах

4

Использовать sklearn для изменения размерности пространства признаков

ЧТО БУДЕМ ОБСУЖДАТЬ

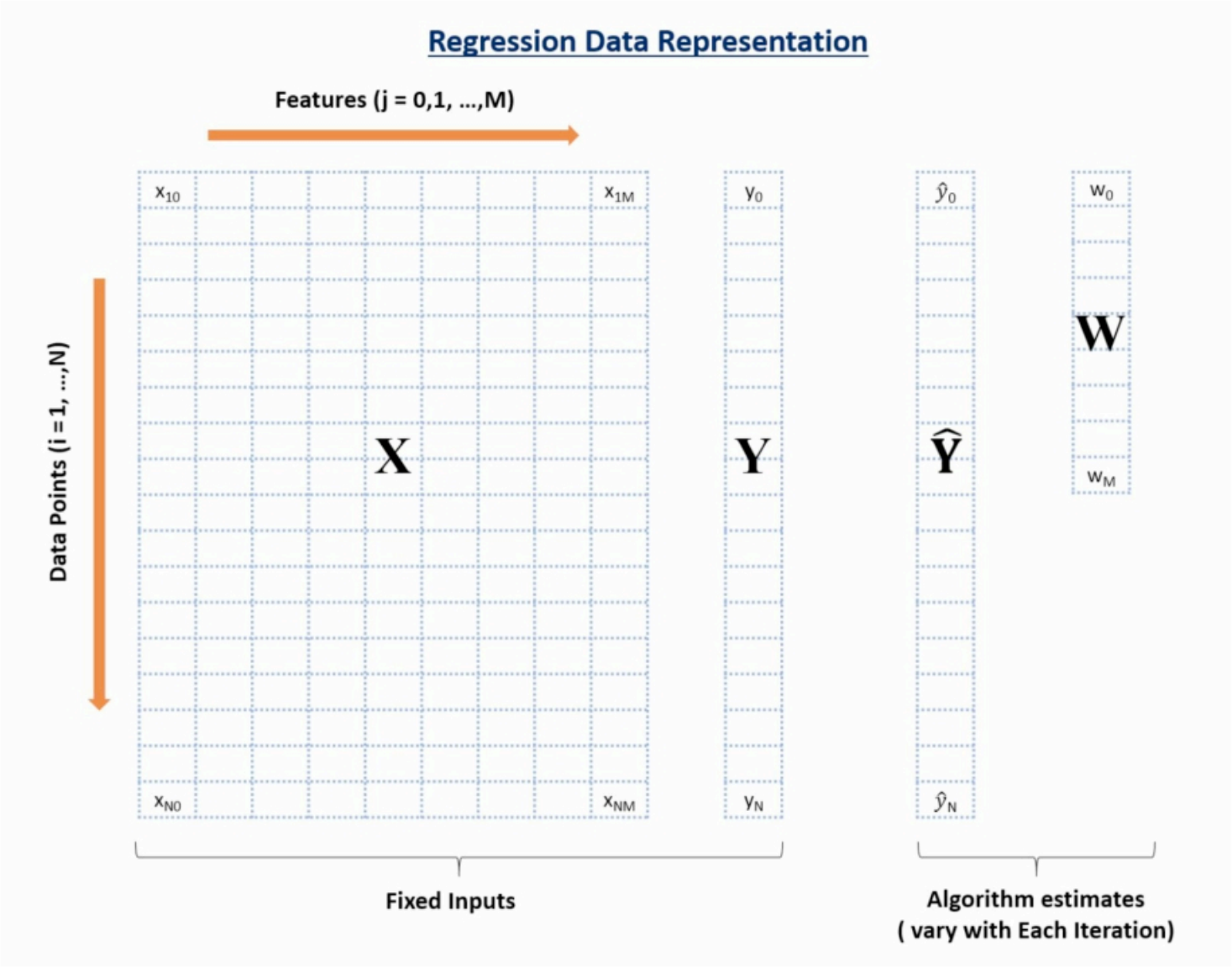
ПЛАН ЗАНЯТИЯ

1. Линейная регрессия
2. Ridge регрессия
3. Lasso регрессия
4. Tree методы
5. Обсуждение домашнего задания

Часть 1-3

Линейная регрессия
Ridge
Lasso

Линейная регрессия



Линейная регрессия

$$\hat{y}_i = \sum_{j=0}^{M} w_j * x_{ij}$$

Коэффициент bj * наблюдение i параметра j

$$Cost\ (W) = RSS(W) = \sum_{i=1}^{N} \{y_i - \hat{y}_i\}^2 = \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j\ x_{ij} \right\}^2$$

(Значение таргета в строке i - предсказанное значение)^2

Линейная регрессия

$$w_j^{t+1} = w_j^t + 2\eta \sum_{i=1}^{N} x_{ij} \left\{ y_i - \sum_{k=0}^{M} w_k \, x_{ik} \right\}$$

$$\frac{\partial}{\partial w_j} Cost\,(W) = -2 \sum_{i=1}^{N} x_{ij} \left\{ y_i - \sum_{k=0}^{M} w_k \, x_{ik} \right\}$$

Ridge регрессия - регуляризация Тихонова - l2

$$Cost\ (W) = RSS(W) + \ \lambda * (sum\ of\ squares\ of\ weights)$$

$$= \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j\, x_{ij} \right\}^2 + \lambda \sum_{j=0}^{M} w_j^2$$

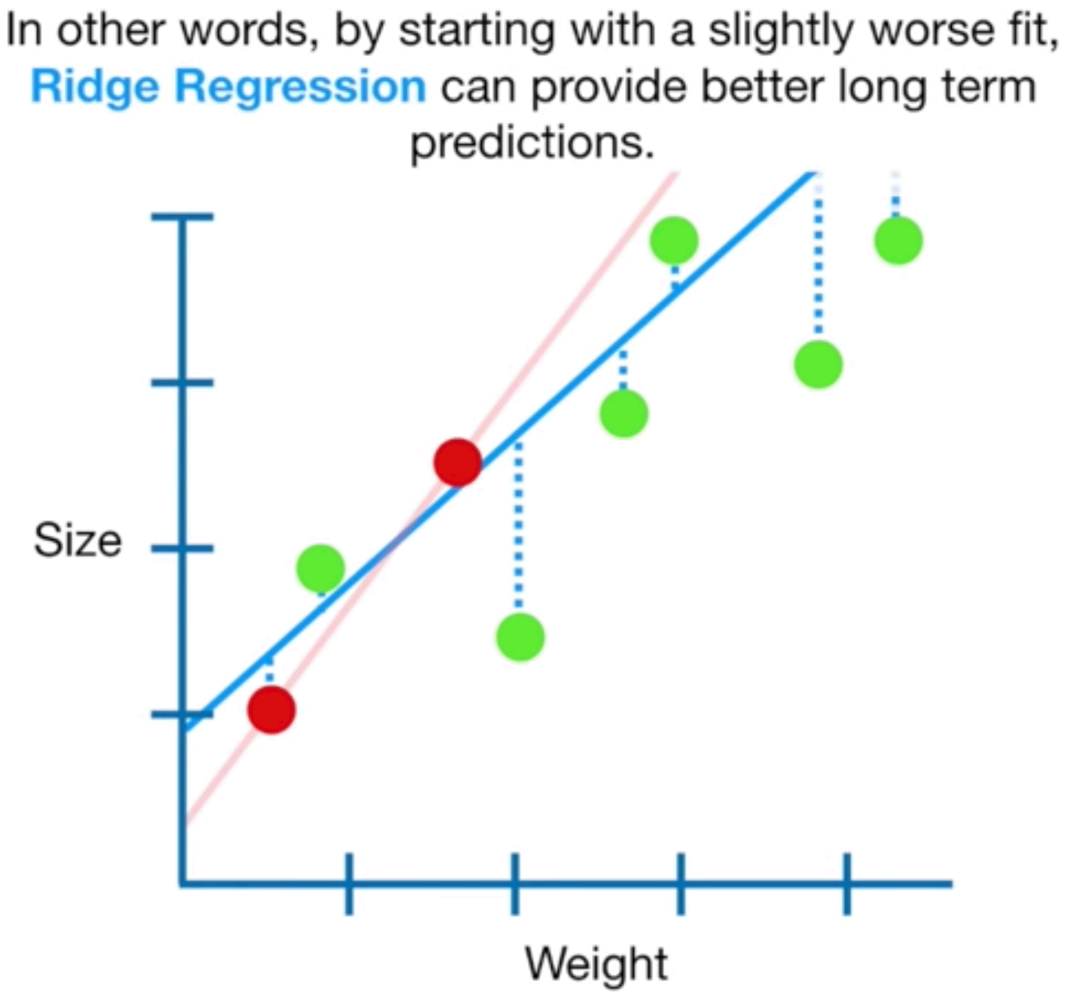
Ridge регрессия - регуляризация Тихонова - l2

Практический смысл

Красные точки - трейн сет

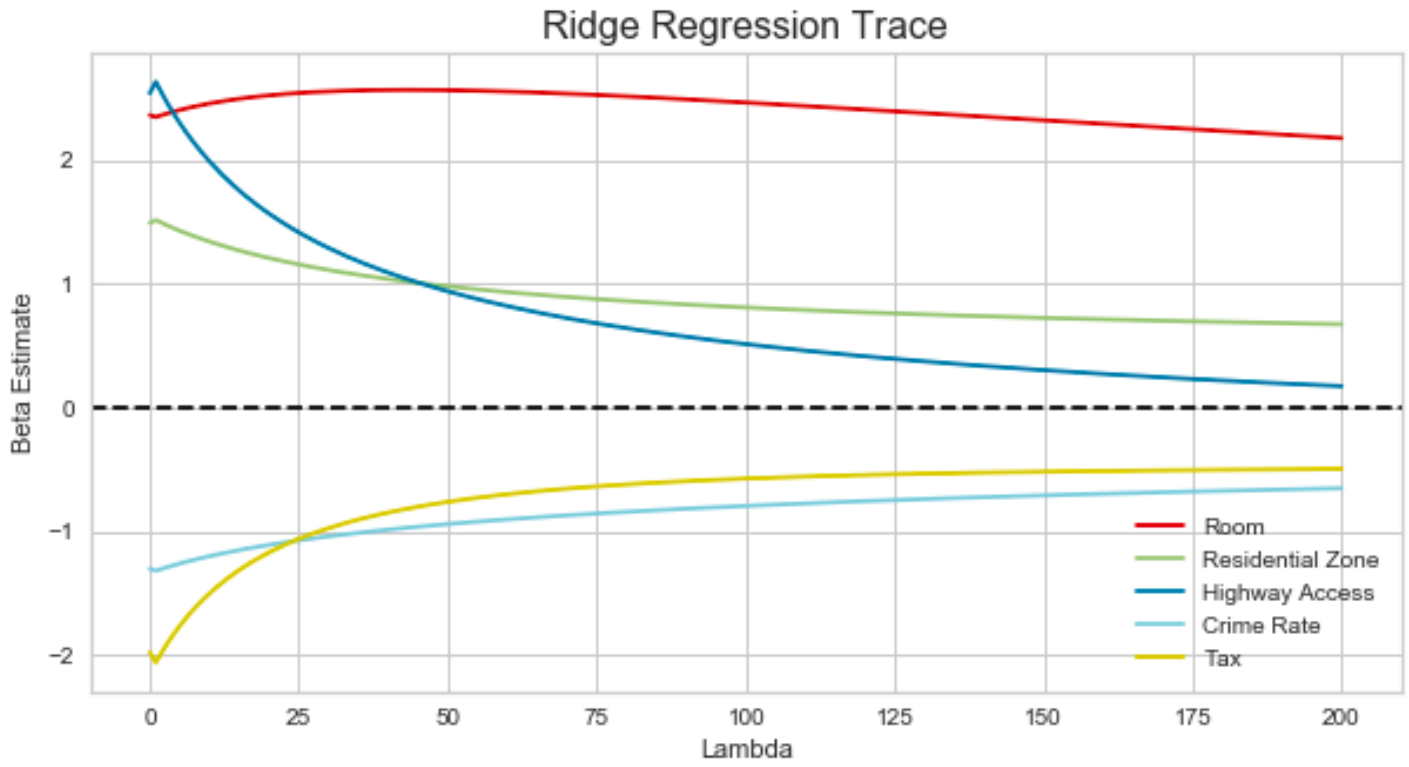
Зеленые точки - тестовый сет

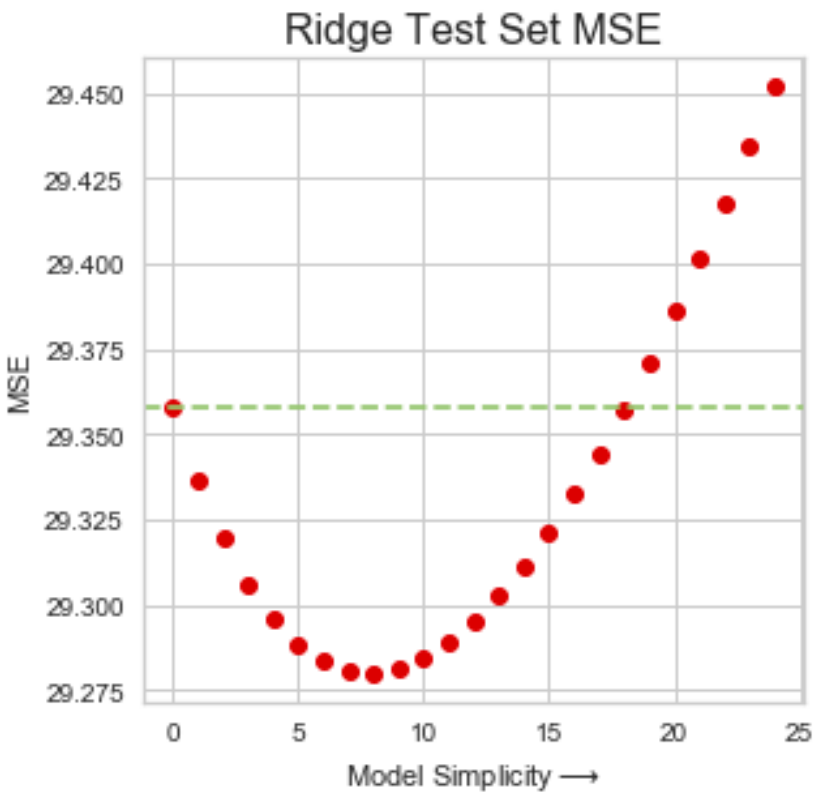
Чуть «менее точным» обучением на трейн сете, мы можем достичь лучших результатов в построение модели



Ridge регрессия - регуляризация Тихонова - l2

Уменьшение коэффициентов при увеличение регуляризационного коэффициента





Источник & code: https://towardsdatascience.com/ridge-regression-for-better-usage-2f19b3a202db

Ridge регрессия

$$w_j^{t+1} = w_j^t - \eta \left[-2 \sum_{i=1}^{N} x_{ij} \left\{ y_i - \sum_{k=0}^{M} w_k * x_{ik} \right\} + 2\lambda w_j \right]$$

$$w_j^{t+1} = (1 - 2\lambda\eta) w_j^t + 2\eta \sum_{i=1}^{N} x_{ij} \left\{ y_i - \sum_{k=0}^{M} w_k * x_{ik} \right\}$$

$$\frac{\partial}{\partial w_j} Cost\ (W) = -2 \sum_{i=1}^{N} x_{ij} \left\{ y_i - \sum_{k=0}^{M} w_k\, x_{ik} \right\} + 2\lambda w_j$$

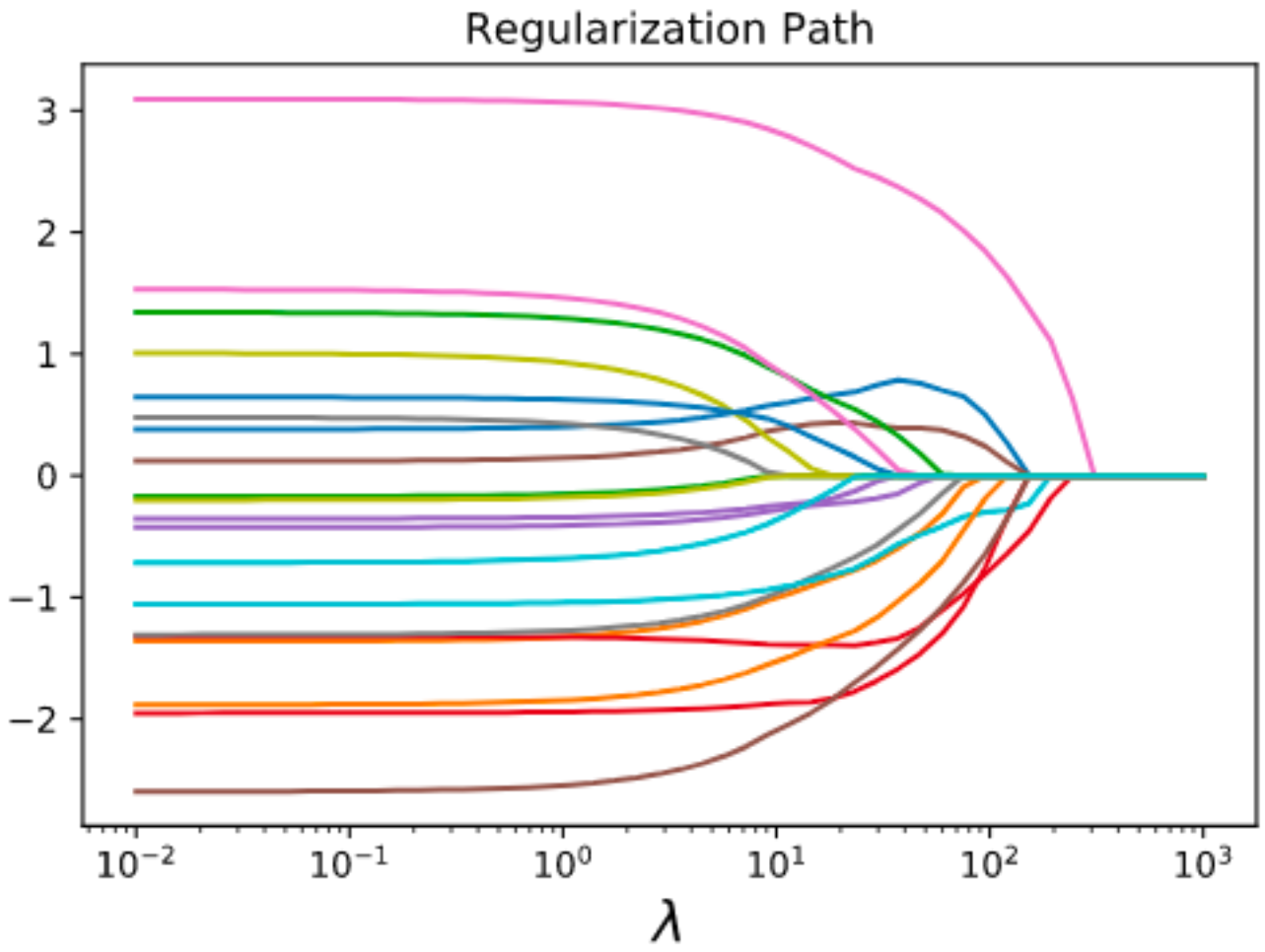
Lasso регрессия - l1

$$Cost\ (W) = RSS(W) + \lambda * (sum\ of\ absolute\ value\ of\ weights)$$

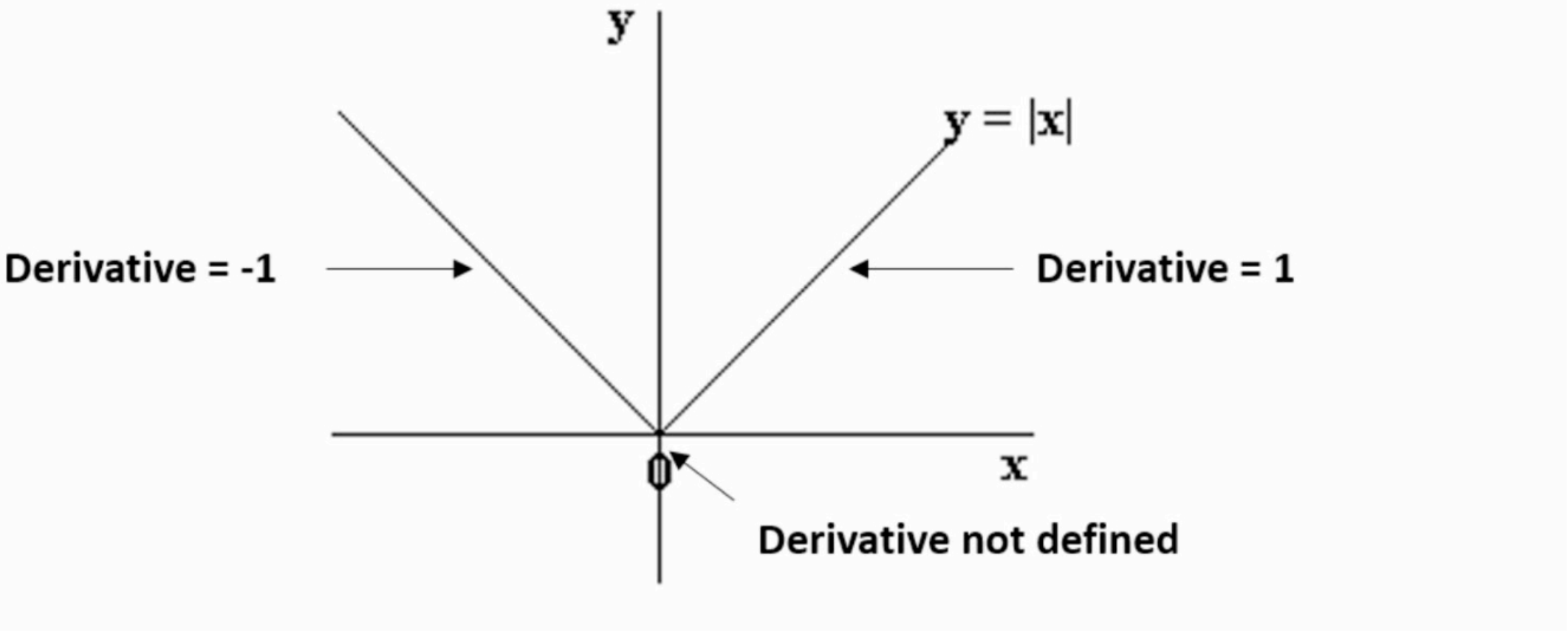
$$= \sum_{i=1}^{N} \left\{ y_i - \sum_{j=0}^{M} w_j\, x_{ij} \right\}^2 + \lambda \sum_{j=0}^{M} |w_j|$$

Lasso регрессия - l1

Затухание коэффициентов при увеличение регуляризационного коэффициента



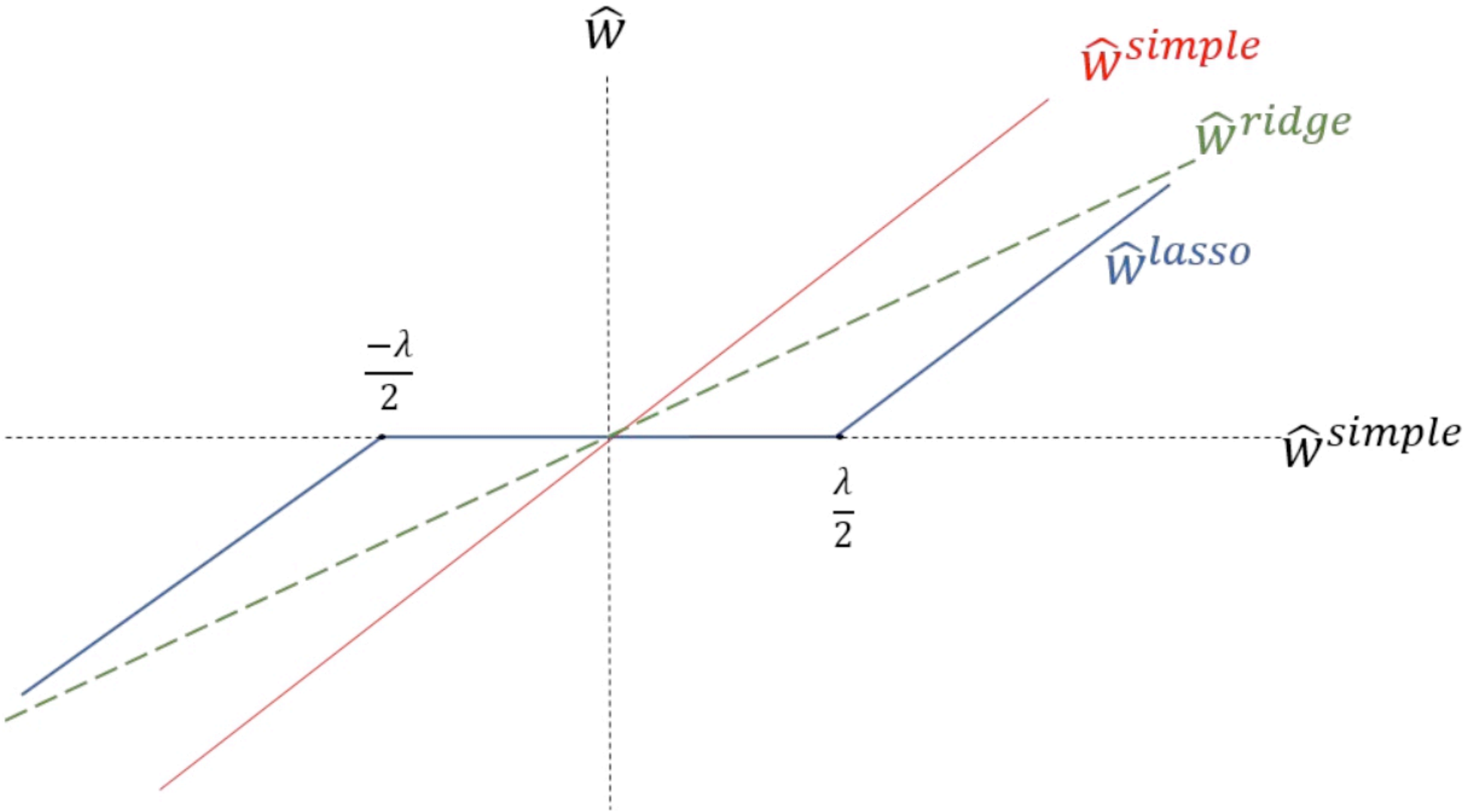
Lasso регрессия



Lasso регрессия

$$w_j = \begin{cases} g(w_{-j}) + \frac{\lambda}{2}, & if\ g(w_{-j}) < -\frac{\lambda}{2} \\ 0, & if\ -\frac{\lambda}{2} \leq g(w_{-j}) \leq \frac{\lambda}{2} \\ g(w_{-j}) - \frac{\lambda}{2}, & if\ g(w_{-j}) > \frac{\lambda}{2} \end{cases}$$

Сравнение коэффициентов при разных видах регрессии

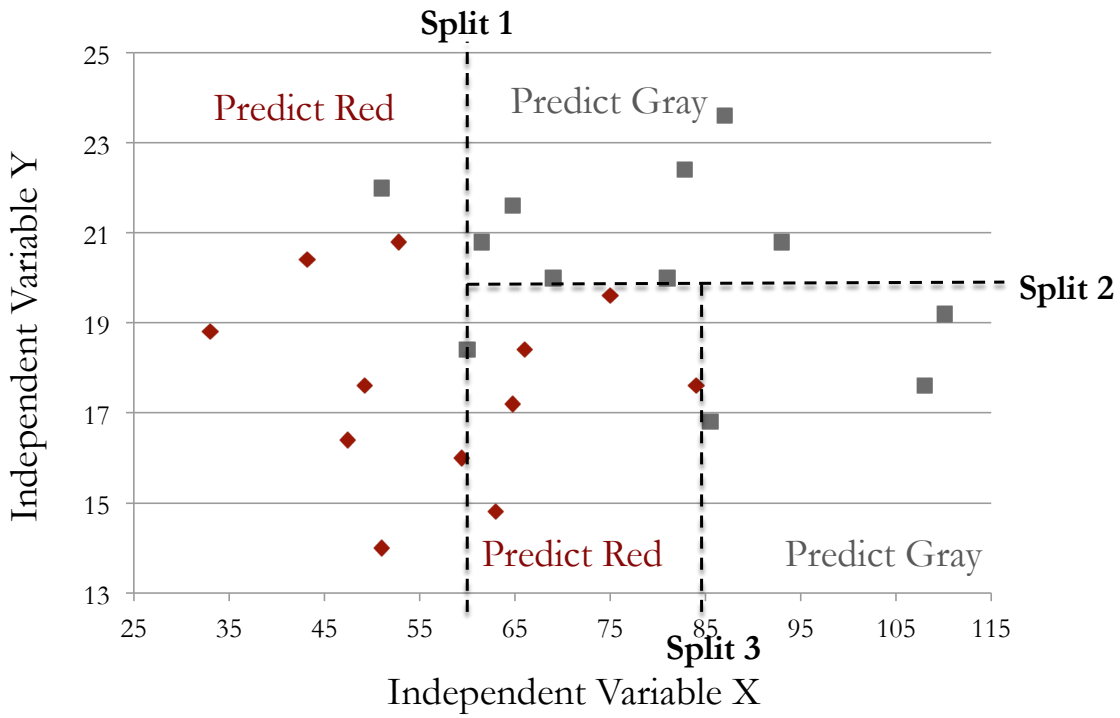


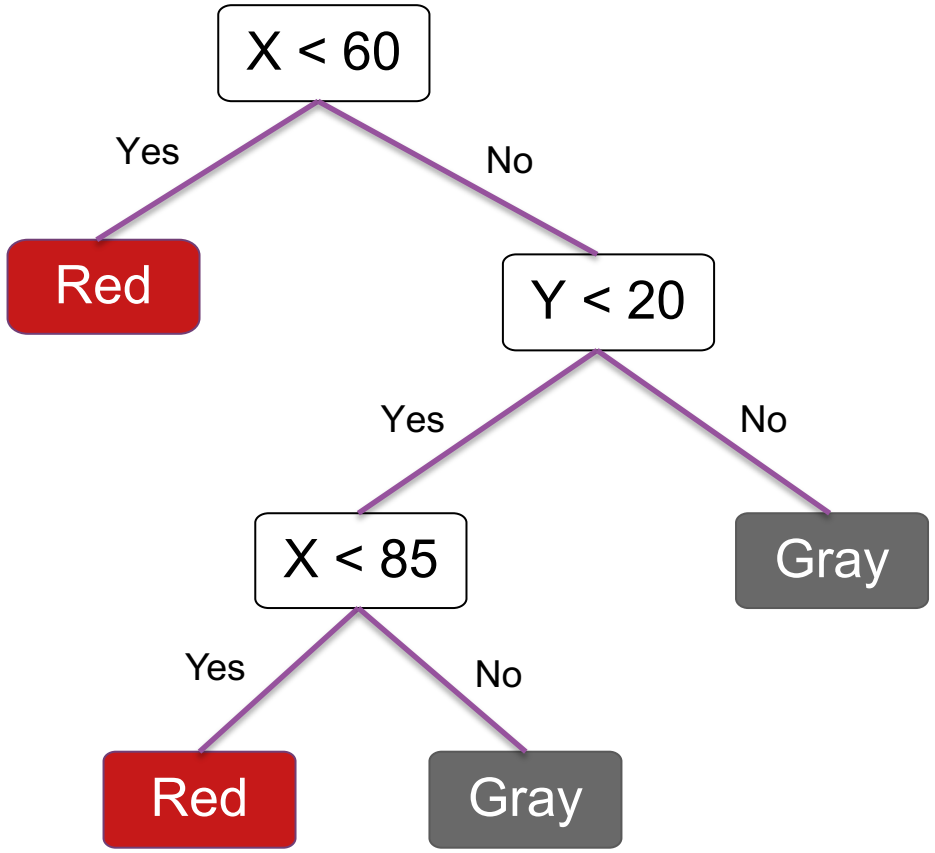
Практика

Как изменяются коэффициенты?

Часть 4
Tree методы

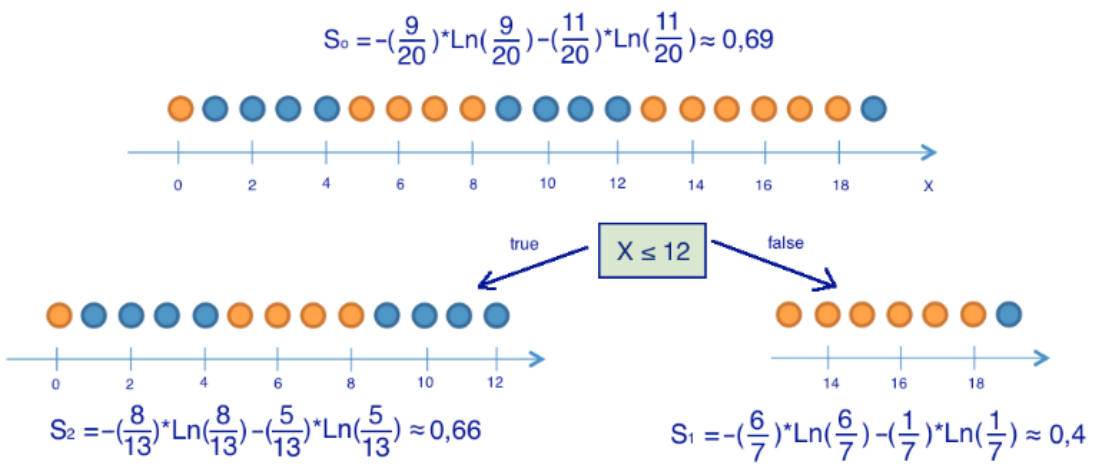
Решающее дерево (Decision tree)

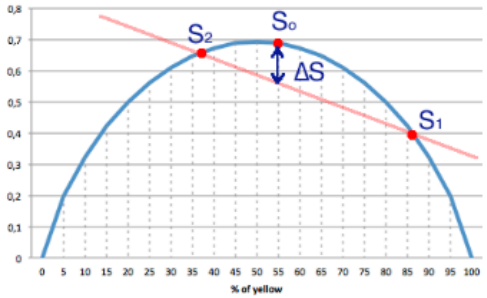


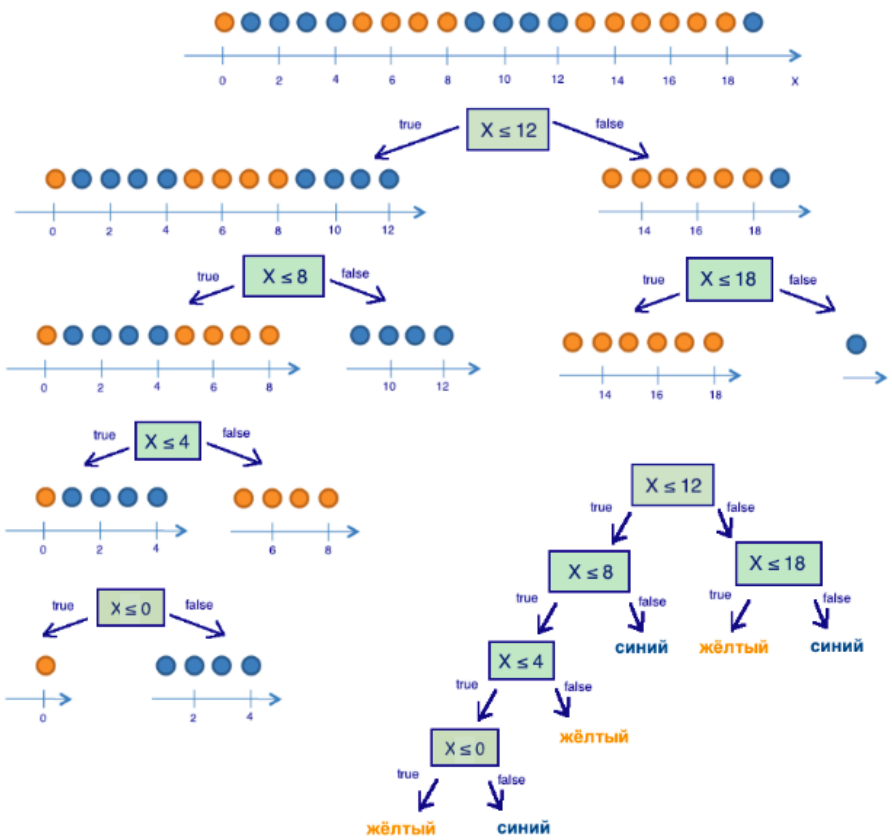


https://www.hse.ru/mirror/pubs/share/215285956

Решающее дерево (Decision tree)

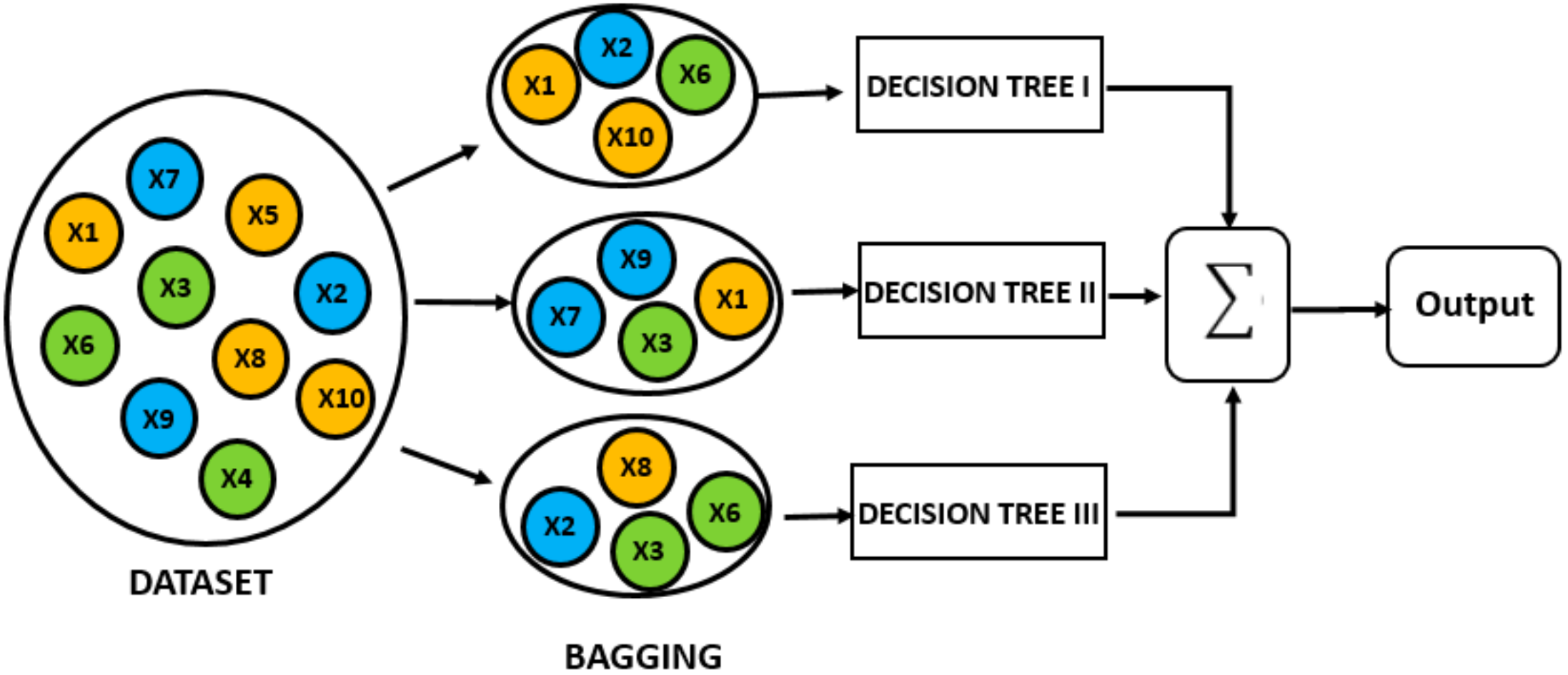






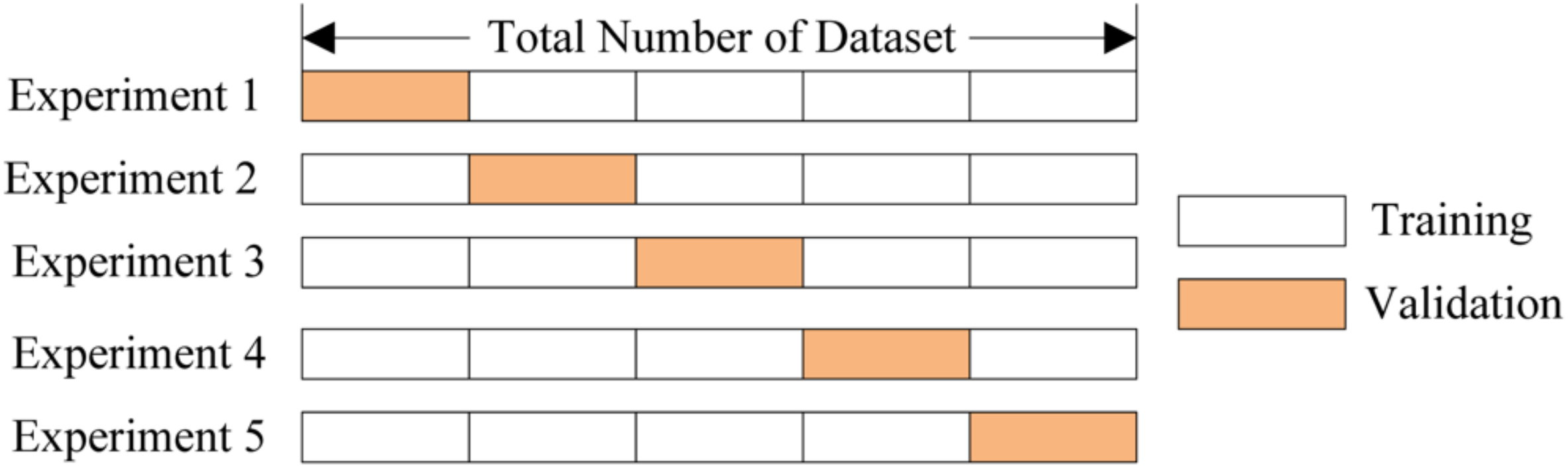
https://www.pvsm.ru/programmirovanie/29106

Random forest



Подробнее: https://habr.com/ru/company/ods/blog/324402/

K-fold кросс-валидация



ЧТО МЫ СЕГОДНЯ УЗНАЛИ

ИТОГИ

1. Как устроена линейная регрессия
2. Отличия Lasso и Ridge регрессий
3. Математика Tree методов
4. Как использовать отбор признаков в sklearn

СПАСИБО ЗА ВНИМАНИЕ