

Машинное обучение и майнинг данных

Лекция 9: Стекинг, вотинг. Градиентный бустинг.

Полина Полунина

Ансамбли моделей

- Ансамбли комбинируют решения нескольких моделей с целью улучшения качества обобщающих и предсказательных свойств
- Ансамбли можно собирать несколькими способами
 - Вотинг
 - Усреднение
 - Стекинг
 - Блендинг
 - Бэггинг
 - Бустинг

Max Voting

- Вотинг (голосование) – это метод, обычно используемый для задач классификации
- Предсказания каждой модели рассматриваются как «голос»
- То предсказание, за которое «проголосовало» большинство моделей используется в качестве финального решения

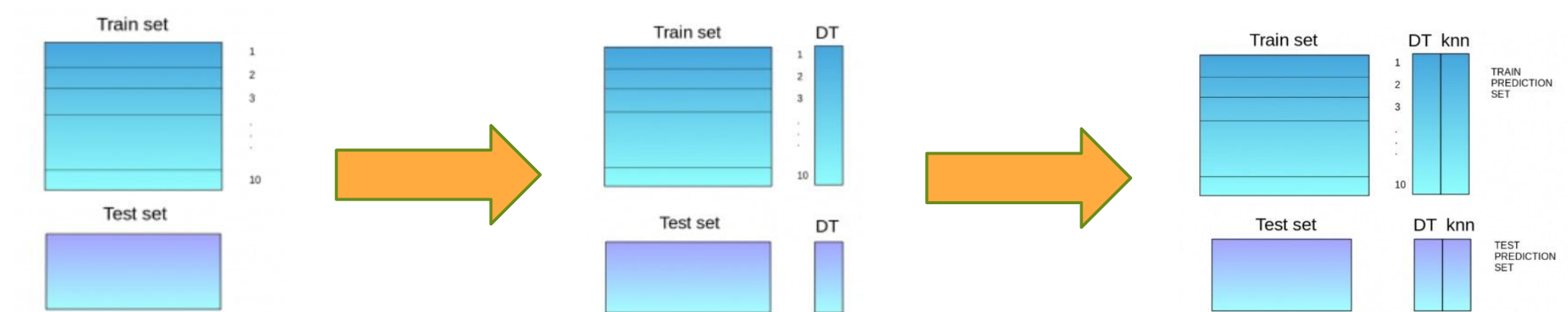
Averaging

- Усреднение – это метод, подходящий как для задач регрессии, так и для классификации
- Финальное решение – это среднее всех решений
- Можно усреднять с некими весами

Stacking

- Обучаем несколько разных моделей разными методами
- Делаем предикт на трейне
- Используем этот предикт как признак в изначальном датасете
- Обучаемся

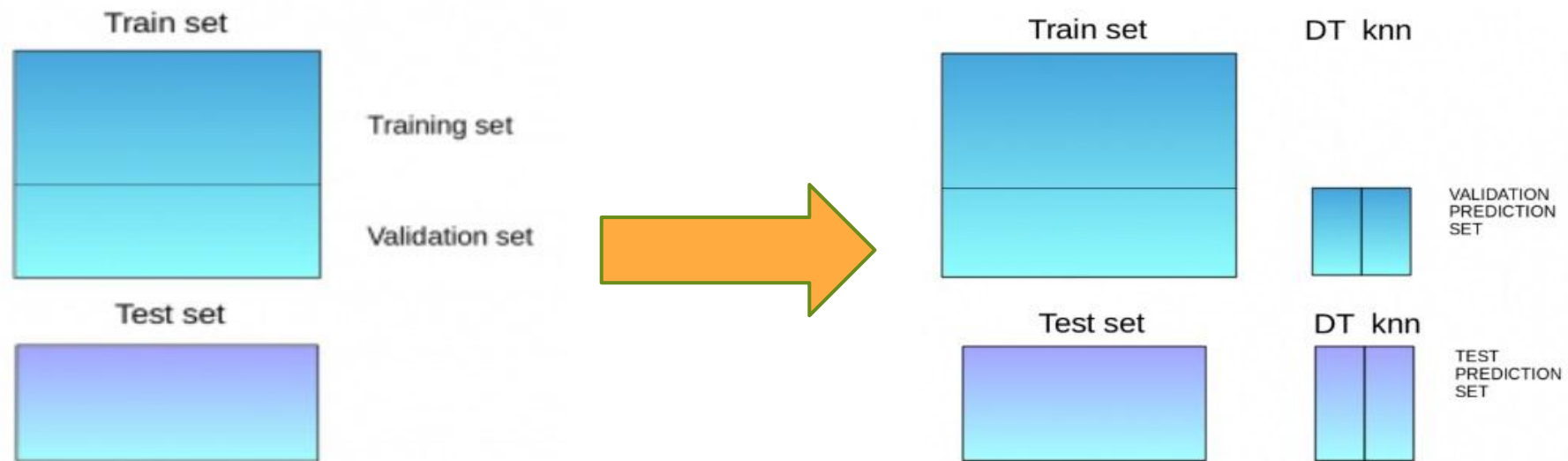
Stacking



Blending

- Делим тренировочный датасет на две части
- Обучаемся разными моделями на первой части, предсказания делаем на второй
- Используем этот предикт как признак в изначальном датасете
- Обучаемся на второй части тренировочного датасета + новые признаки

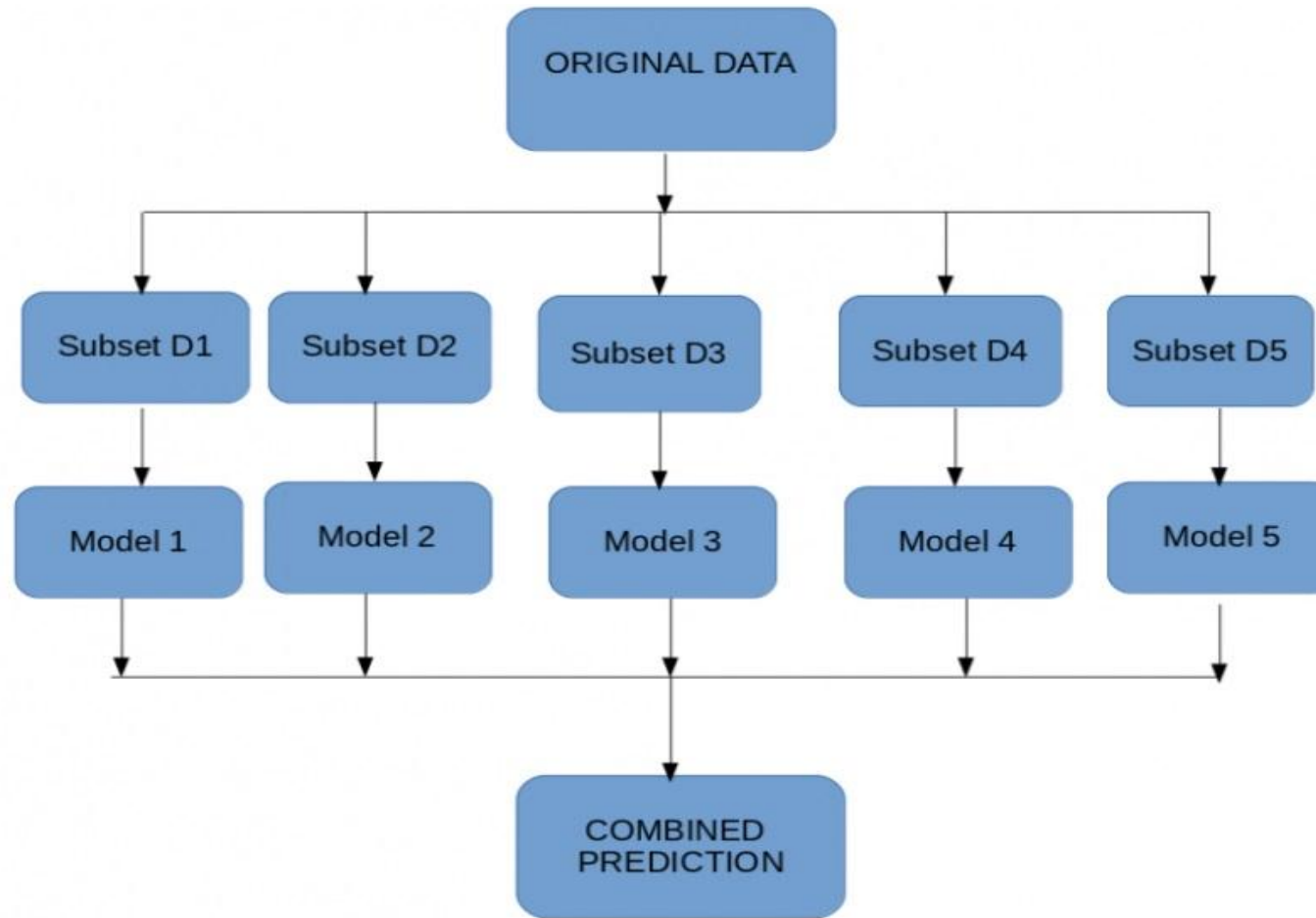
Blending



Bagging

- Семплируем подвыборки из основного датасета с возвращением (например, бутстрепом)
- Обучаемся разными моделями на всех подвыборках параллельно
- Финальные предсказания формируются как комбинация всех моделей

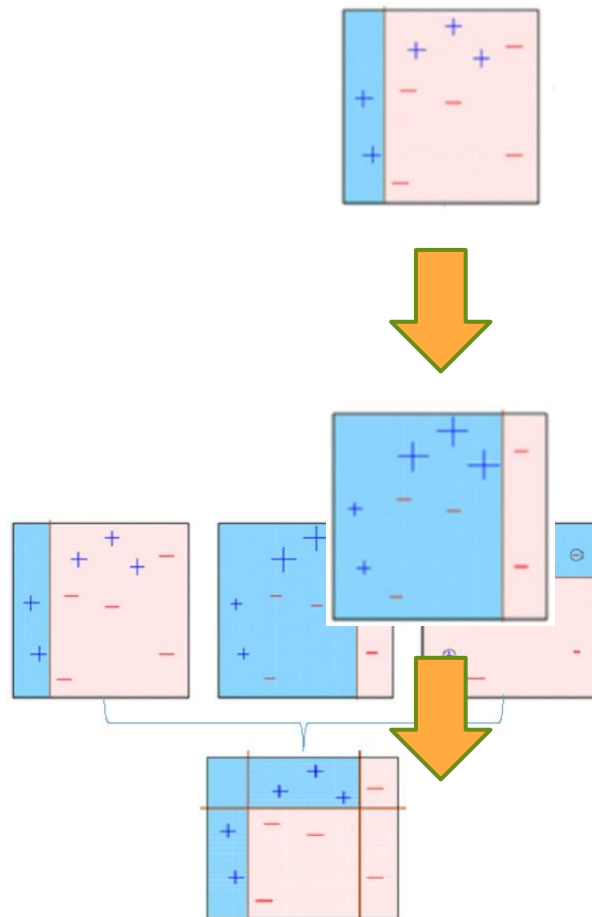
Bagging



Boosting

- Бустинг – это итеративный алгоритм
- Каждая последующая модель пытается исправить ошибки предыдущей модели
- Строится модель на подвыборке данных
- Делаем предсказания для всего датасета
- Считаем ошибки
- Наблюдениям с большей ошибкой присваивается больший вес
- Строим новую модель
- ...
- Финальная модель – это взвешенное среднее всех предыдущих моделей

Boosting



Градиентный бустинг

- Градиентный бустинг использует градиентный спуск для минимизации ошибки предсказания
- Другими словами, на каждой итерации мы обновляем значения предсказаний таким образом, чтобы минимизировать функцию потерь

Градиентный бустинг

