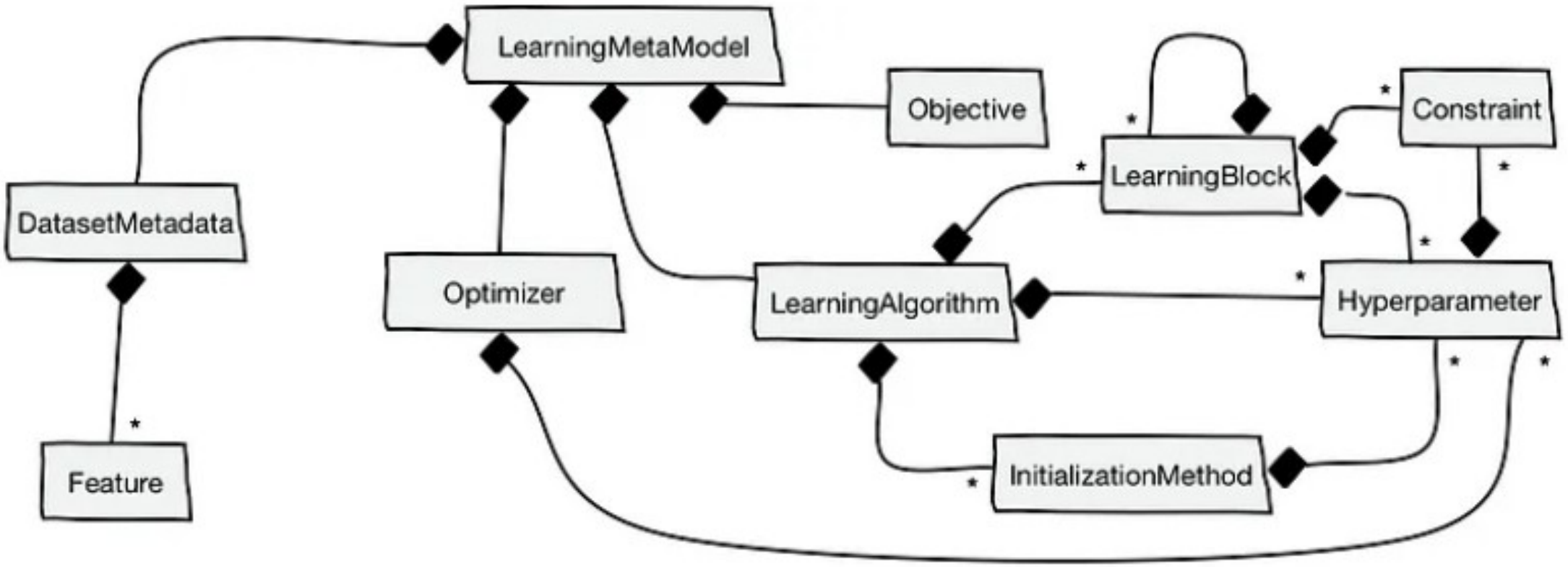
Resampling в SVM

Золин Иван гр. 5030102/00201

Проблематика

- Необходимо оценивать различные алгоритмы обучения и настраивать многочисленные параметры в соответствии с конкретными проблемами. Это является затратной задачей.
- Метамодель – метод автоматизации этой проблемы путём подбора функции регрессии на основе целевой функции.
- В общем случае, метамодель состоит из целевой функции (objective), алгоритма обучения, оптимизатора, метаданных.
- Важно: мета-модель лишь приблизительно описывает исходную задачу, что может вносить искажения, ошибки в полученный оптимум.

Метамодель в ML



Методы повторной выборки

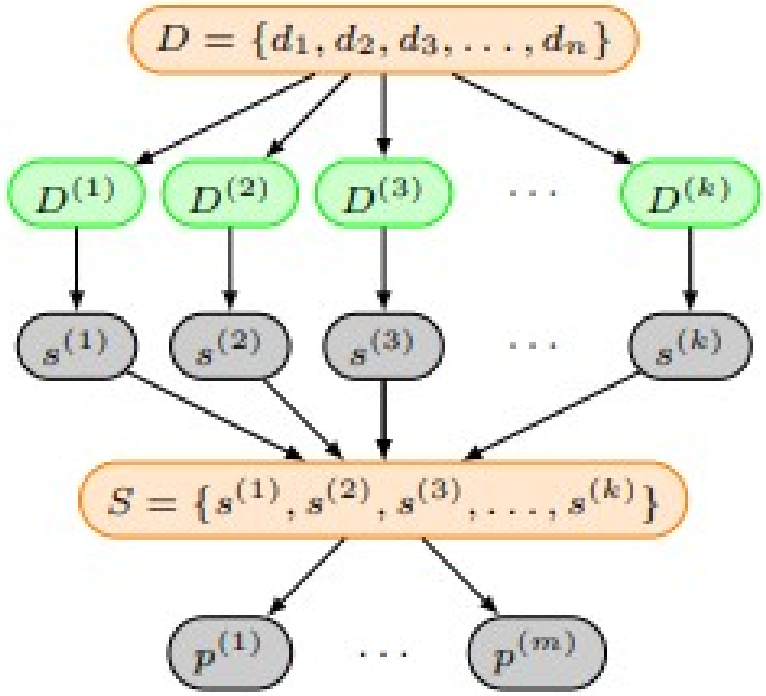
- Оценка точности мета-модели
- Выбор модели
  - Часто несколько классов моделей являются подходящими для желаемой метамодели
  - Рекомендация: выбор менее сложной модели на выборках небольшого размера
- Настройка гиперпараметров
  - Оценка модели, обновлённой настройкой гиперпараметров
- Замечание
  - Метамодель должна использоваться вместе с целевой функцией, чтобы обеспечить приближение к оптимуму.

Постановка задачи

- Целевая функция f , набор данных , где - ковариантный вектор:
- Цель: найти регрессионную функцию к нашим данным найти метамодель, которая аппроксимирует , используя информацию в
- Функции потерь : MSE, MAE
- Разбиваем датасет : на обучающую и тестовые выборки
- Однако …

Идея

- Проблемы
  - Требуется большой объём данных D
  - Требуется достаточное количество выборок в тестовом наборе
  - Также невозможно обнаружить дисперсию и нестабильность модели из-за изменений в обучающем наборе
- Методы повторной выборки
  - Многократно генерируют обучающие и тестовые наборы из имеющегося набора
  - Подгоняют модель к каждому обучающему набору и оценивают качество по тестовому



Общая схема повторной выборки

– обучающая выборка

– множество состояний функции потерь

Cross-validation

- Идея
  - Разделить исходный набор на k блоков одинакового размера
  - Использование k-1 блоков для подгонки и проверки на оставшемся
  - Это производится для всех возможных комбинаций k-1 из k блоков
- k=10, 10-кратная кроссвалидация (k=5,10,n)
- k=n – LOOCV – leave-one-out
  - Каждое наблюдение – набор проверки, остальные n-1 – обучающий набор

Algorithm 3: Subsets for k-fold CV.

input : A dataset D of n observations d_1 to d_n and the number of subsets k to generate.

output: k subsets of D named $D^{(1)}$ to $D^{(k)}$.

$D \leftarrow$ **Shuffle** (D)
for $i \leftarrow 1$ **to** k **do**
  $D^{(i)} \leftarrow D$
for $j \leftarrow 1$ **to** n **do**
  $i \leftarrow (j \bmod k) + 1$
  $D^{(i)} \leftarrow D^{(i)} \setminus \{d_j\}$

Bootstrap

- Идея
  - Генерация k подмножеств с заменой
  - Каждый обучающий набор используется для подгонки модели, остальные – тестовый набор
- Преимущество и недостаток
  - Размер обучающего набора равен исходному набору данных => обеспечивается надёжность оценки
  - Некоторые наблюдения в обучающем наборе могут присутствовать несколько раз
  - Решение: добавление случайного шума
- Обычно k = 100…1000 (верхнего предела нет)

Algorithm 4: Subsets for bootstrap.

input : A dataset D of n observations d_1 to d_n and the number of subsets k to generate.

output: k subsets of D named $D^{(1)}$ to $D^{(k)}$.

```
for i ← 1 to k do
    D^(i) ← ∅
    for j ← 1 to n do
        d ← RandomElement (D)
        D^(i) ← D^(i) ∪ {d}
```

Subsampling

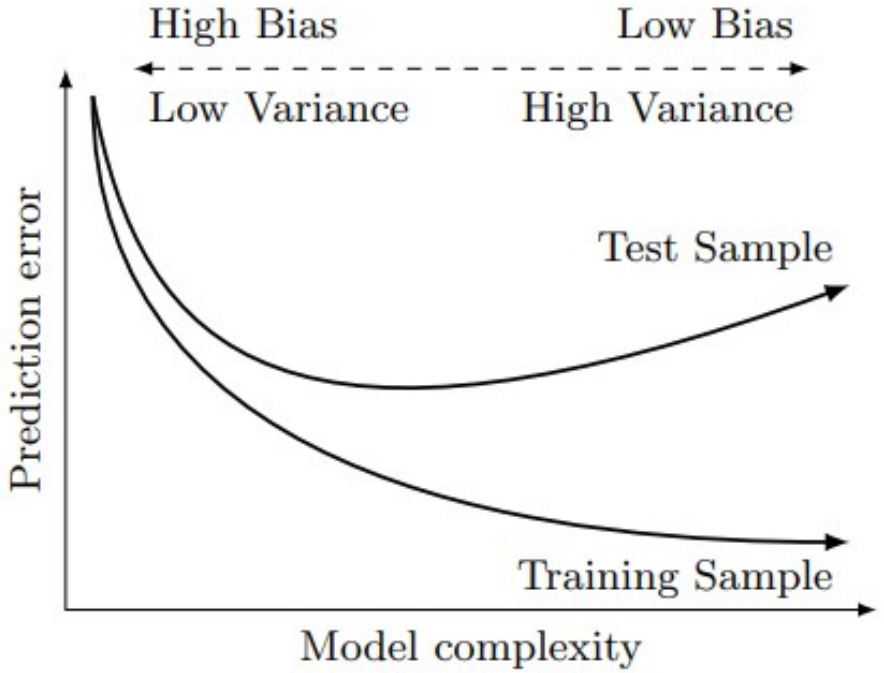
- Идея схожа с Bootstrap
  - Отличие: наблюдения из D берутся без возвращения. Т.е. обучающий набор должен быть меньше D
  - k должен быть выбран пользователем априори
  - Варианты выбора такие же: k=100..1000

input : A dataset D of n observations d_1 to d_n, the number of subsets k to generate and the subsampling rate r.
output: k subsets of D named $D^{(1)}$ to $D^{(k)}$.

```
1 m ← ⌊r|D|⌋
2 for i ← 1 to k do
3 |   D' ← D
4 |   D^(i) ← ∅
5 |   for j ← 1 to m do
6 |   |   d ← RandomElement (D')
7 |   |   D^(i) ← D^(i) ∪ {d}
8 |   |   D' ← D' \ {d}
```

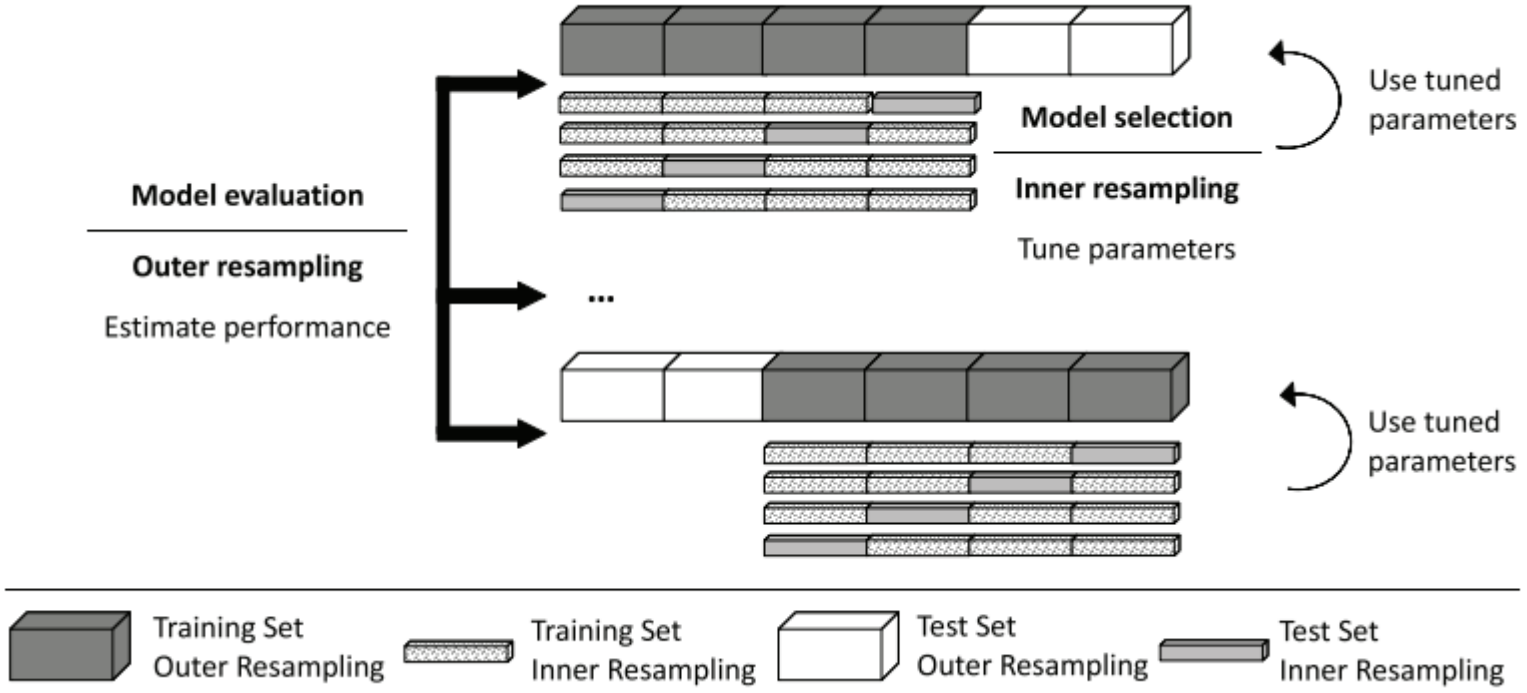
Вложенная повторная выборка (1/2)

- Часто получение модели > подгонка регрессионной модели
  - Необходимость выбора гиперпараметров
- Оптимальные гиперпараметры зависят от данных
  - НО: нельзя выполнять этапы выбора модели на одних и тех же наборах выборки, которые используются для оценки. Может привести к переобучению! - «обучение на тестовом наборе»



Вложенная повторная выборка (2/2)

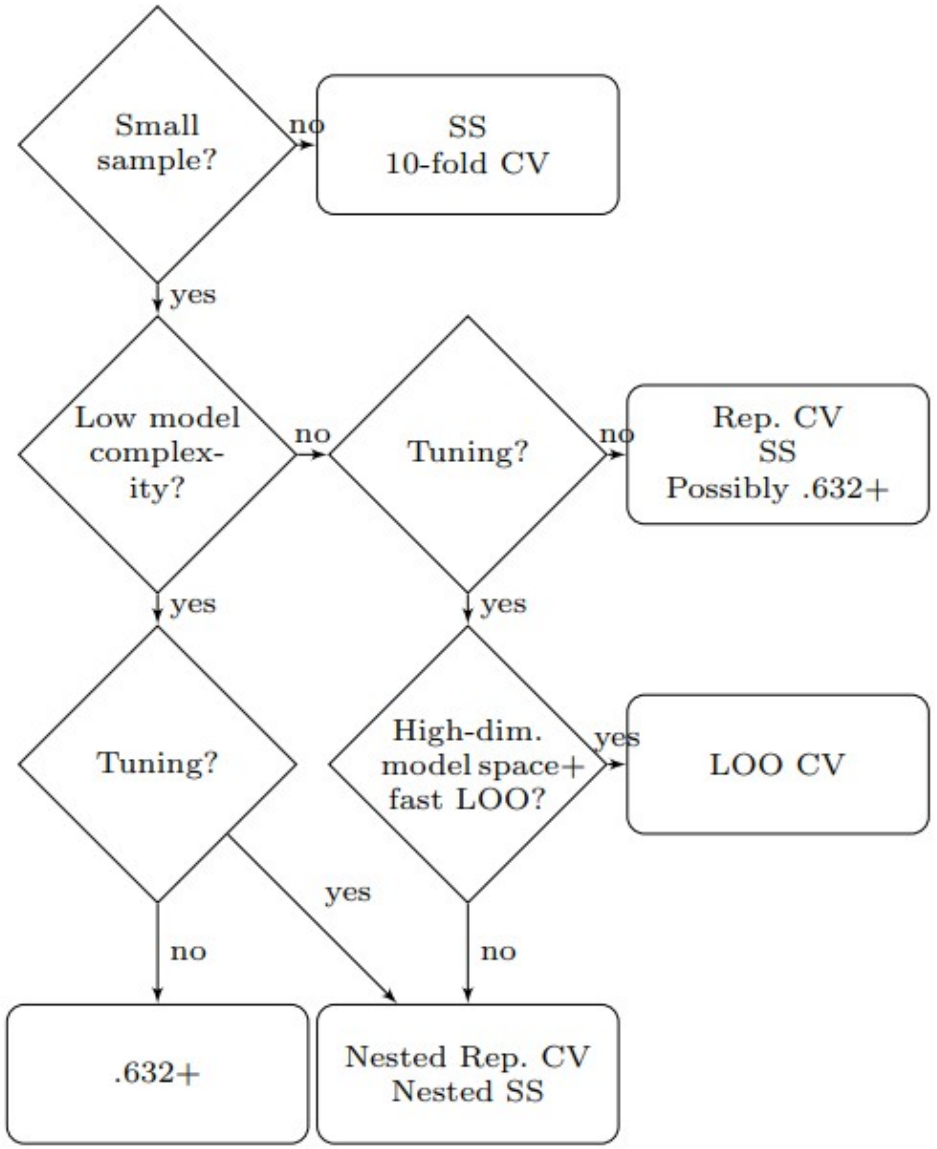
- Выбор модели - часть подгонки, требующая повторения для каждого набора.
- Вложенная повторная выборка включает в себя оценку модели во внешнем цикле и повторный выбор внутри каждого обучающего набора.



- Например, используем subsampling с k=100 для оценки и 5-кратную кроссвалидацию для выбора гиперпараметра.
- Требуется много вычислительных затрат, но обеспечивает объективные результаты.

Стратегия выбора метода повторной выборки

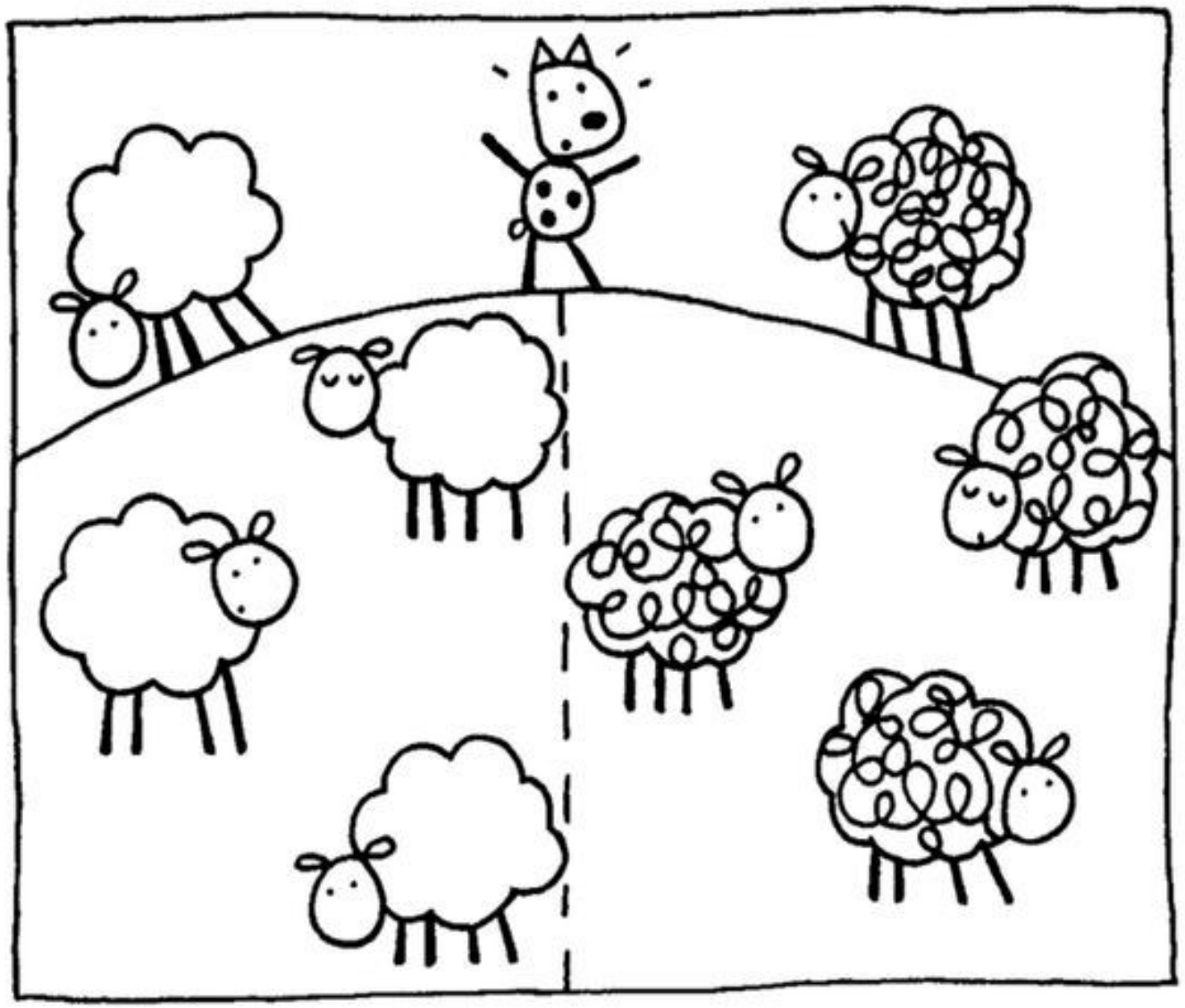
- Cross-validation – для быстрой настройки модели
- .632+ (bootstrap) – для небольших размеров выборок с моделями низкой сложности и когда настройка не требуется
- Во всех остальных случаях: Subsampling, REPCV



Выводы

- Правильная валидация модели имеет решающее значение в обучении
- Мета-модель, которая плохо аппроксимирует целевую функцию, не может приводить к надёжным результатам оптимизации.
- Настройка гиперпараметров, выбор конечного количества различных моделей, выбор соответствующих функций – важные шаги оптимизации

Спасибо
за
внимание!



Список литературы

- Workshop on Experimental Methods for the Assessment of Computational Systems (WEMACS 2010)
- Meta-Modelling Meta-Learning (Thomas Hartmann)