Семинар 2

Евгений Елтышев

План

- Валидация
- KNN

Функционал качества

- В общем виде: $Q(y, \tilde{y})$
- MSE = $\frac{1}{|X|} \sum_{(x,y) \in X} (y - f(x))^2$
- Accuracy = $\frac{1}{|X|} \sum_{(x,y) \in X} I(y = a(x))$

На каких данных проверять?

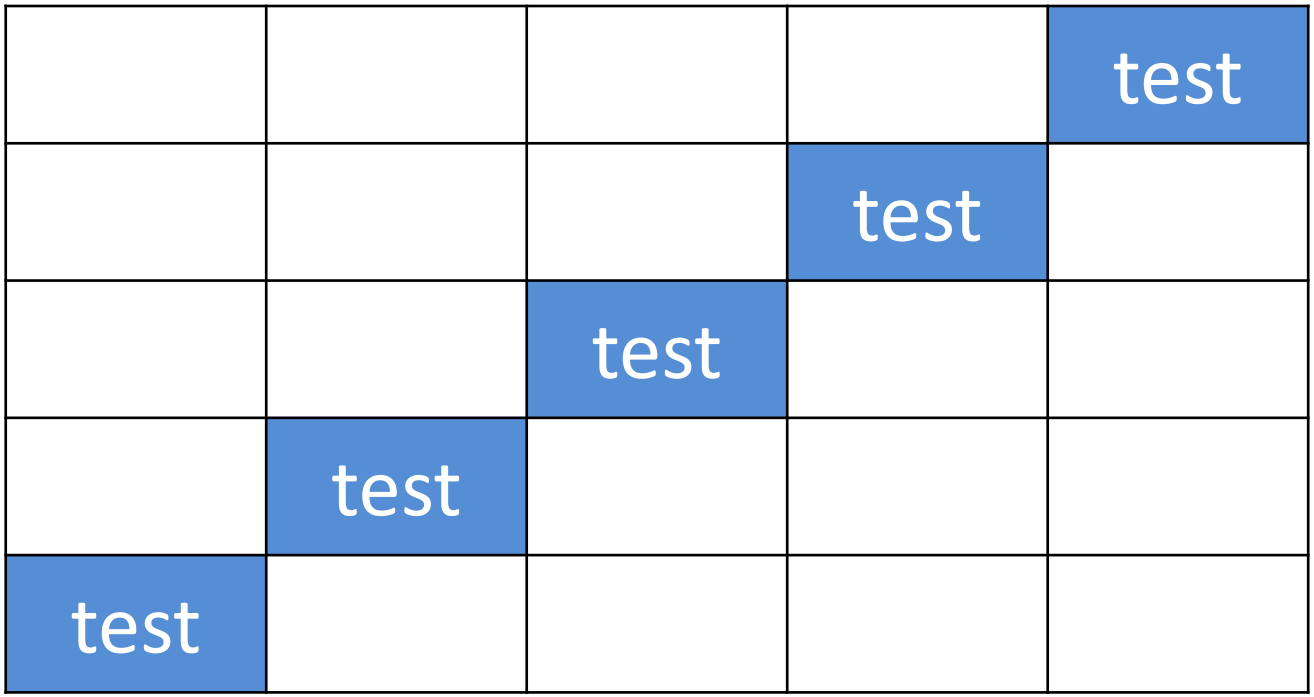
Dataset

Классический подход

Классический подход

+ Быстро
+ Меньше переобучения
- Меньше обучающая выборка
- Все равно можно переобучиться на Validation

K-fold



K-fold

+ Меньше дисперсия оценки качества
+ Не переобучаемся под Validation
- Долго

История из жизни

- Модерация изображений в альбомах
- Данные: ручные оценки ассессоров
- Признаки: количество вердиктов «насилие», «экстремизм», «нормальное»
- В трейне оценки собирали 2 недели
- Но в будущем планируют собирать только 3 дня

Тонкости

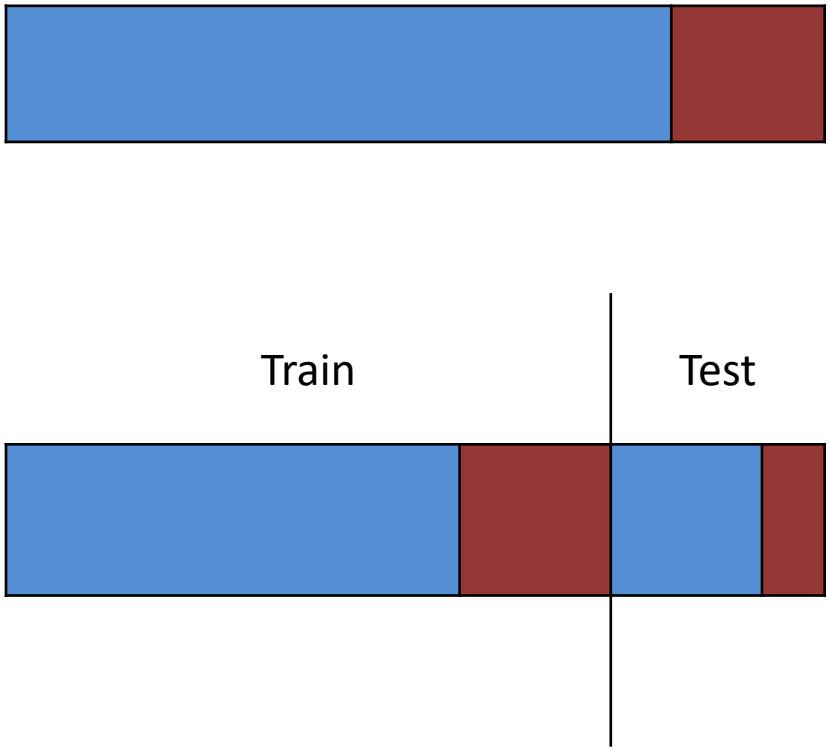
- На валидации должно быть то же распределение, что и в проде

Баланс классов

- Пусть точность на классе 0: 90%
  на классе 1: 30%
- Пропорции «в проде»: 10% - 90%

- На несбалансированной валидации: Acc = 60%
- Потом в проде: 0.1 * 90% + 0.9 * 30% = 36%

Stratified Split



Группировка объектов

- Caterpillar Tube Pricing

Тип	Количество
Труба круглая №1	1
Труба круглая №1	10
Труба круглая №1	50
Труба круглая №1	100
Труба круглая №2	1
Труба круглая №2	25
Труба круглая №2	50
Труба круглая №2	100

Группировка объектов

Тип	Количество	Выборка
Труба круглая №1	1	Train
Труба круглая №1	10	Train
Труба круглая №1	50	Train
Труба круглая №1	100	Train
Труба квадратная №2	1	Test
Труба квадратная №2	25	Test
Труба квадратная №2	50	Test
Труба квадратная №2	100	Test

Временная зависимость и KFold

- Предсказание продаж на следующую неделю

Временная зависимость и KFold