

Занятие № 3

Оценка качества моделей и работа с признаками

План занятия



1. Оценка качества моделей
2. Извлечение признаков
3. Преобразование признаков
4. Работа с пропущенными данными
5. Отбор признаков
6. Мастер-класс по работе с гео-данными



Оценка качества моделей

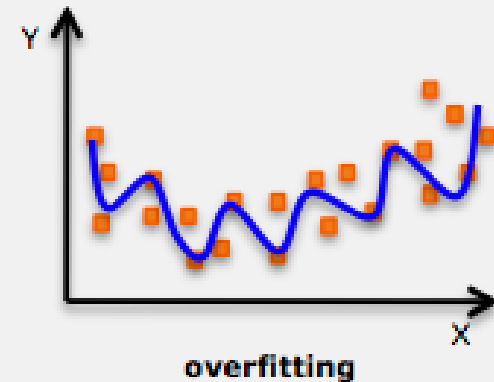
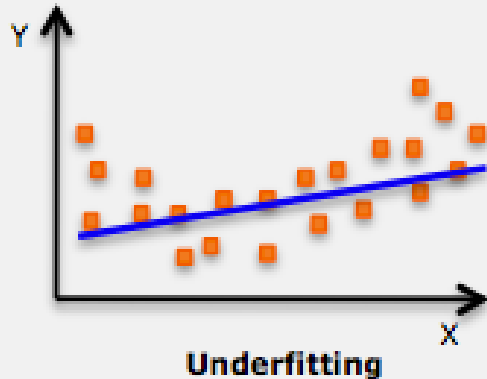
Как выбрать лучшую модель для решения конкретной задачи?

Проблема переобучения



Способы обучиться на наборе данных:

1. Найти общие закономерности в предоставленном наборе данных
2. Запомнить правильные ответы



Для оценки качества модели нельзя использовать те же данные, что и для построения модели.

Train, Test, Validation



На **Train** обучаем модели-кандидаты

На **Test** оцениваем модели-кандидаты и выбираем лучшую

На **Validation** проверяем, что все работает как ожидалось

- **Validation** никак не используем при построении модели!
- На гиперпараметрах модели тоже можно переобучиться!

Shuffle & Split



Перемешиваем сэмплы, делим датасет на две части (**Train** и **Test**) в некоторой пропорции. На **Train** обучаем модель, на **Test** оцениваем качество.

Особенности:

- Простая реализация
- Разумно использовать когда данных "много"

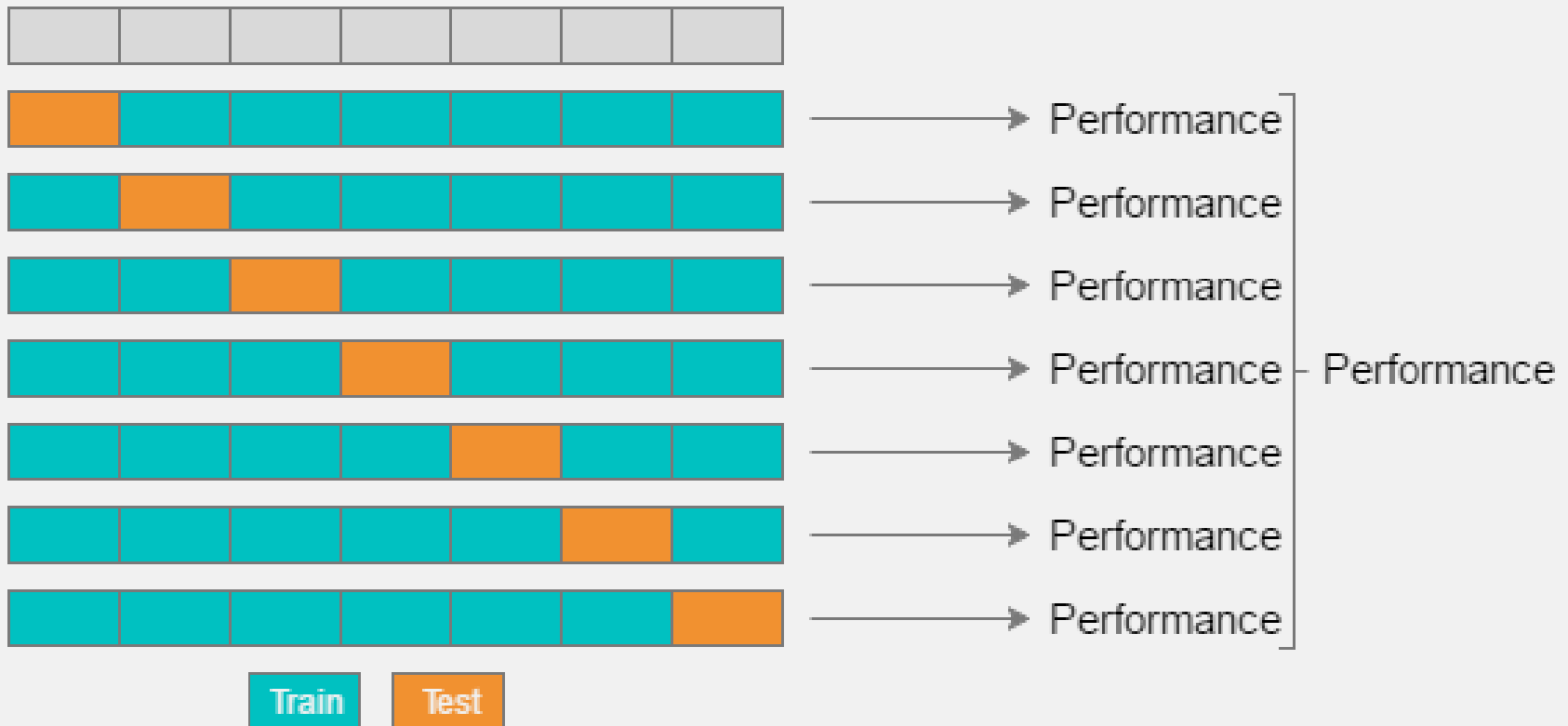


Разбиваем датасет на **K** равных частей, затем строим **K** моделей где в качестве **Test** берем одну из частей, а все остальные используем как **Train**. На **Train** обучаем модель, на **Test** оцениваем качество.

Особенности:

- Используем все данные как для построения моделей, так и для оценки качества
- Один из наиболее популярных методов оценки качества моделей

K-fold



Leave One Out



Экстремальный случай **K-fold**, когда **K** равно числу сэмплов в наборе данных.

Особенности:

- Модель на датасете без одного сэмпла практически идентична модели на полном датасете
- Может быть эффективно посчитан для некоторых видов моделей

Повторные разбиения



K-fold или **Shuffle & Split**, повторенный **N** раз с различными разбиениями.

Особенности:

- В N раз выше вычислительная сложность
- Эффективны когда данных "мало" или данные "шумные"

Стратифицированные разбиения



Стратифицированные разбиения - это такие разбиения, которые сохраняют определенные свойства исходной выборки.

Свойствами могут быть:

- Распределение целевой переменной
- Распределения некоторых признаков

Стратифицированные разбиения



Плюсы:

- Могут обеспечить большую точность, чем простой случайный выбор
- Позволяют избежать "непредставительной" выборки (например отсутствие какого-либо класса объектов в разбиении)

Минусы:

- Сложнее в реализации



Иногда в наборе данных сэмплы разбиты на **группы**.

Например:

- Фотографии гистологических срезов одной и той же опухоли **группу**
- Продукция из одной партии составляет **группу**

Игнорирование **групп** приводит к некорректной оценке качества модели!



Групповые K-fold или **Shuffle & Split** - это такие разбиения при которых все сэмплы одной группы попадают в один и то же фолд или сплит.

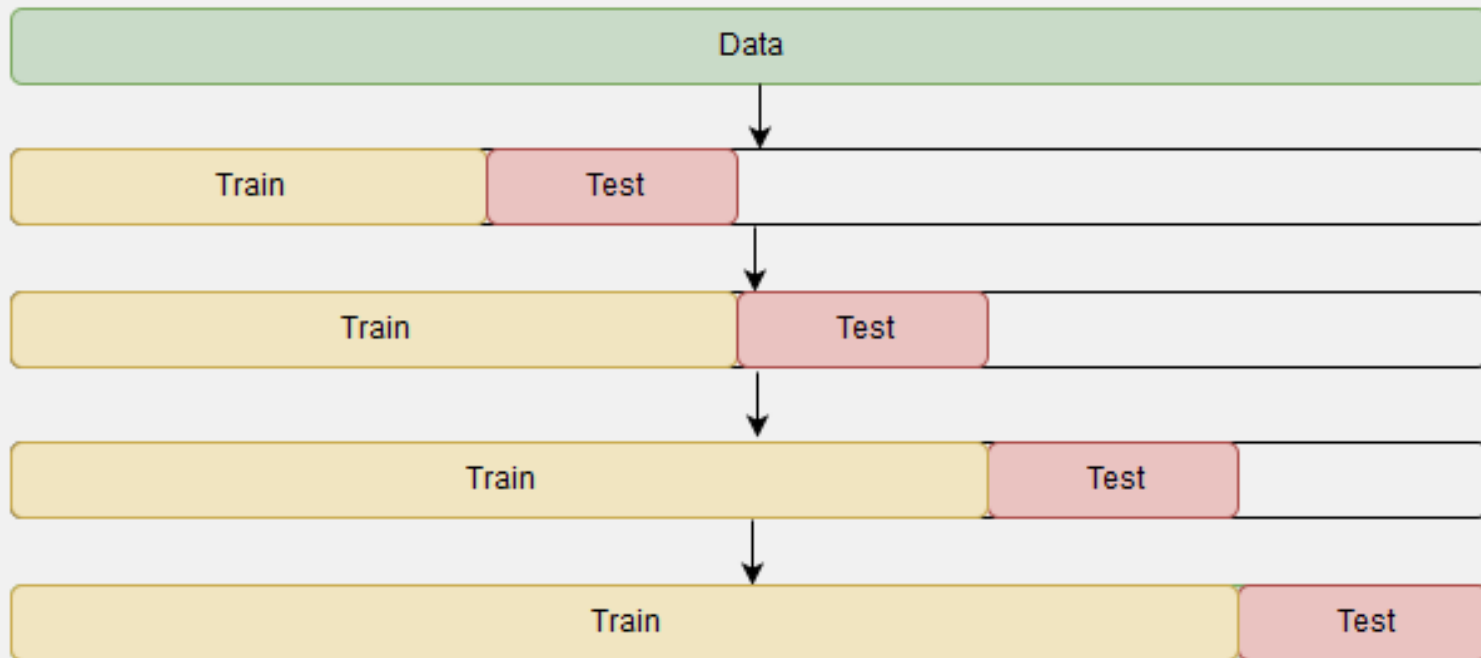
Особенности:

- Позволяют корректно работать с данными в которых есть группы

Разбиение временных рядов



Разбиение временных рядов следует проводить по времени события.





Выбор схемы оценки качества – это баланс вычислительной сложности и точности оценки

Схемы оценки моделей:

1. Shuffle & Split
2. K-fold $K=2-4$
3. K-fold $K=5-10$
4. Repeated K-fold $K=5-10$

Как правило 3x5-fold работает лучше чем 1x15-fold



Извлечение признаков

Описываем реальный мир на языке, понятном модели.

Извлечение признаков



Сампл (пример) - это вектор чисел.

Объект - представлен набором данных.

Извлечение признаков - представление реального или цифрового объекта в виде вектора чисел.



Признаки бывают:

- **Количественные**
- **Порядковые**
- **Категориальные**
- **Бинарные**





Данные бывают:

1. Числовые
2. Дата и время
3. Геоданные (latitude, longitude)
4. Временные ряды
5. Текстовые данные
6. Графические изображения
7. Звук
8. Видео
9. И др.



Что можно извлечь:

1. Что находится по заданной координате
2. Расстояния до особых объектов



Что можно извлечь:

1. Абсолютное время
2. Периодичность (час, день, месяц ...)
3. Временной интервал до особого события

Временные ряды



Что можно извлечь:

1. Среднее значение за период
2. Стандартное отклонение за период
3. Тренд за период
4. Количество пиков за период

Извлечение признаков



Год выпуска	2011
Пробег	98 000 км
Кузов	Внедорожник 5 дв.
Цвет	Белый
Двигатель	6.2 л / 409 л.с. / бензин
Коробка	Автоматическая
Привод	Полный
Руль	Левый
Состояние	Не требует ремонта
Владельцы	3 владельца
ПТС	Оригинал
Владение	9 месяцев
Таможня	Растаможен
VIN	XWFS47EF*C0***62
Автокод ?	Без ограничений

Характеристики модели в каталоге





Преобразование признаков

Адаптируем признаковое представление под конкретный тип модели.

Зачем преобразовывать признаки?



1. Чтобы конкретный алгоритм машинного обучения их правильно интерпретировал
2. Чтобы конкретный алгоритм машинного обучения эффективно находил взаимосвязи
3. Чтобы внести априорные знания о наборе данных или свойствах признаков





Для каждого признака в наборе вычитаем среднее и делим на стандартное отклонение.

Применяется к **количественным, порядковым и бинарным** признакам

Актуально для:

Линейные модели

Метод ближайших соседей



Значения каждого признака в наборе приводят к диапазону $[0,1]$.

Применяется к **количественным, порядковым и бинарным** признакам

Актуально для:

Линейные модели

Метод ближайших соседей

Монотонные преобразования



Применение монотонного преобразования к признаку (например: логарифмирование, возведение в степень)

Применяется к **количественным** и **порядковым** признакам

Актуально для:

Линейные модели

Метод ближайших соседей

Линеаризация (регрессия)



Применяем нелинейное преобразование к одному или более признакам чтобы получить новый признак, линейно зависящий от целевой переменной (например: физические законы).

Применяется к **количественным** признакам

Актуально для:

Линейные модели



Область значений **количественного** или **порядкового** признака делим на N участков и представляем в виде N бинарных признаков.

Применяется к **количественным** и **порядковым** признакам

Актуально для:

Линейные модели

Полиномиальные признаки



Заменяем исходный набор признаков полиномом от исходных признаков.

$$(x_1, x_2) \rightarrow (x_1, x_2, x_1^2, x_2^2, x_1x_2)$$

Применяется к **количественным, порядковым и бинарным** признакам

Актуально для:

Линейные модели

Метод ближайших соседей

Решающие деревья

One hot encoding



Представление **категориального признака** с **N** категорий как **N** бинарных признаков.

Применяется к **категориальным** и **порядковым** признакам

Актуально для:

Линейные модели

Метод ближайших соседей

Некоторых типов решающих деревьев

Задача



Алгоритм: k ближайших соседей с евклидовым расстоянием

Признаки:

1. категория кинотеатра [1..43]
2. день недели [0..6]
3. час суток [0..23]
4. цена билета [100..1000]

Целевая переменная: заполненность зала в %

Что делать?



Работа с пропущенными данными

Что делать, если для части объектов данные не полные?



Простые решения:

- Некоторые алгоритмы поддерживают работу с пропущенными данными "из коробки".
- Закодировать пропущенные данные особым значением (0, -999 и т.п.).
- Закодировать пропущенные данные типичным значением (среднее, медиана, наиболее частое значение).
- Для временных рядов можно брать соседнее значение соседей.

Работа с пропущенными данными



Более сложные решения:

- Можно использовать модель для заполнения пропущенных данных (например алгоритм k ближайших соседей).



Отбор признаков

Все ли признаки полезны?

Отбор признаков, зачем?



- Меньше признаков - выше производительность.
- Меньше признаков - проще их сбор и преобразование.
- Снижение количества признаков может приводить как к снижению так и к росту точности модели.

Каждый признак - это сигнал + шум



- Выбрасываем признаки, значение которых константно на тренировочном наборе данных (всем или большей части)
- Выбрасываем признаки, которые слабо статистически связаны с целевой переменной (например: корреляция)

Проблема статистической связи



Признак 1	Признак 2	Целевая переменная
1	0	1
1	0	1
1	0	1
0	1	1
1	1	0
0	0	0
0	0	0
0	0	0

Отбор признаков по важности для модели



Смотрим на какие признаки модель опирается при принятии решений.

Пример: веса при признаках для линейной модели

Последовательный набор признаков



Пусть дано M основных и N дополнительных признаков.

1. Из N дополнительных признаков по очереди добавляем каждый к основному набору из M признаков и таким образом получаем N новых наборов признаков.
2. Оцениваем качество модели на каждом из N наборов признаков.
3. Выбираем набор с наилучшей точностью.
4. Переходим на шаг 1 (опционально если точность улучшилась).

Последовательный отброс признаков



Пусть N - количество признаков

1. Из полного набора признаков по очереди выбрасываем каждый и таким образом получаем N новых наборов признаков.
2. Оцениваем качество модели на каждом из N наборов признаков.
3. Выбираем набор с наилучшей точностью.
4. Переходим на шаг 1 (опционально если точность улучшилась).

Метод "Permutation Importance"



1. Делим выборку на train и test
2. Обучаем модель на train
3. Выполняем предсказание на test, измеряем качество и записываем его как базовый скор
4. Для каждого признака в test выполняем перемешивание значений в колонке и выполняем предсказание, важность признака оцениваем как полученный скор минус базовый скор



**Спасибо за
внимание!**

Евгений Некрасов

e.nekrasov@corp.mail.ru