Семинар 2

ИИ и БА, УБ 3 курс

Содержание

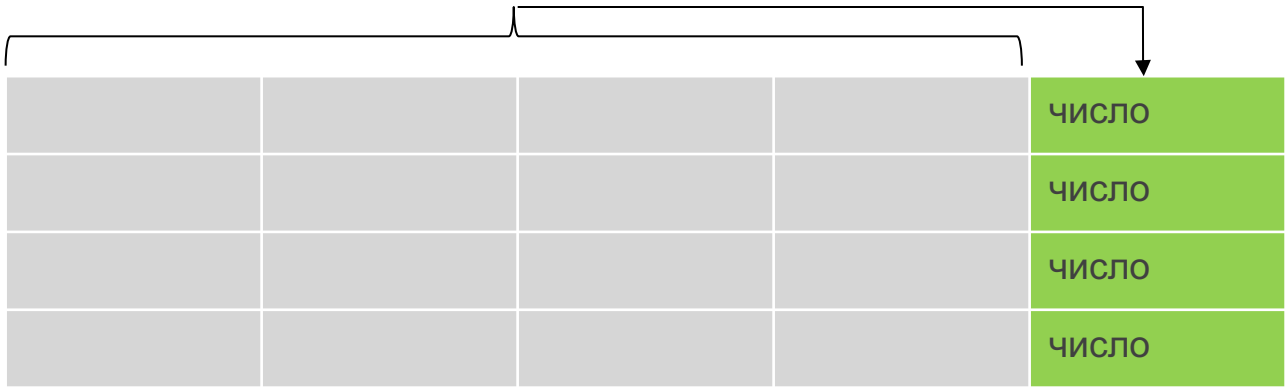
Основные типы задач машинного обучения и по каждому:

- Примеры практического применения
- Основные метрики эффективности
- Популярные алгоритмы
- Типичные подводные камни и рекомендации как на них не напороться

Регрессия

Definition

Регрессия — это задача предсказания непрерывного числового значения на основе имеющегося набора признаков (переменных)



Примеры бизнес-кейсов

	ЦИАН	Пятёрочка	Альфа Банк
Тех задача	Прогнозирование цены продажи квартиры	Прогнозирование спроса на товары (сколько купят)	Прогноз доходов клиентов в банковской сфере
Бизнес задача	Поставить в объявлении цену, за которую купят и захотят продать	Привезти столько товара, сколько нужно	Спрориоритизировать обработку самых платежеспособных

Как оценивать модели? (математика)

MAE
(Mean Absolute Error)

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i|$$

MSE
(Mean Squared Error)

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2$$

+2

RMSE
(Root Mean Squared Error)

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$$

+√

R^2
(коэффициент детерминации)

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$$

- n — количество наблюдений,
- $\hat{y}_i$ — предсказанное значение,
- y_i — истинное значение,
- $\bar{y}$ — среднее значение целевой переменной,
- $|y_i - \hat{y}_i|$ — абсолютная ошибка для каждого наблюдения.

Как оценивать модели? (смысл)

MAE
(Mean Absolute Error)

насколько в среднем предсказания отличаются от реальных значений

MSE
(Mean Squared Error)

возводит отклонения в квадрат, чувствительнее к выбросам

RMSE
(Root Mean Squared Error)

имеет ту же размерность, что и целевая переменная (единицы измерения), так как обратно извлекает корень

R^2
(коэффициент детерминации)

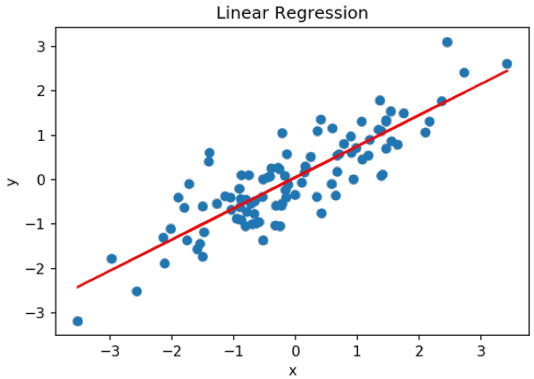
какую долю дисперсии целевой переменной объясняет модель

Меньше - лучше

Ближе к 1 - лучше

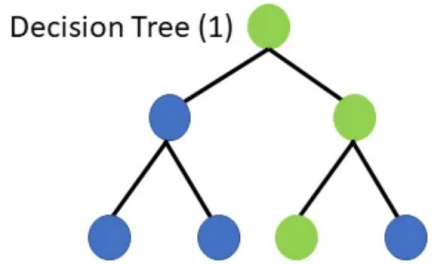
Основные алгоритмы

1.
Линейная регрессия



Эволюция:
Регуляризованные решения для функции потерь
(Ridge, Lasso)

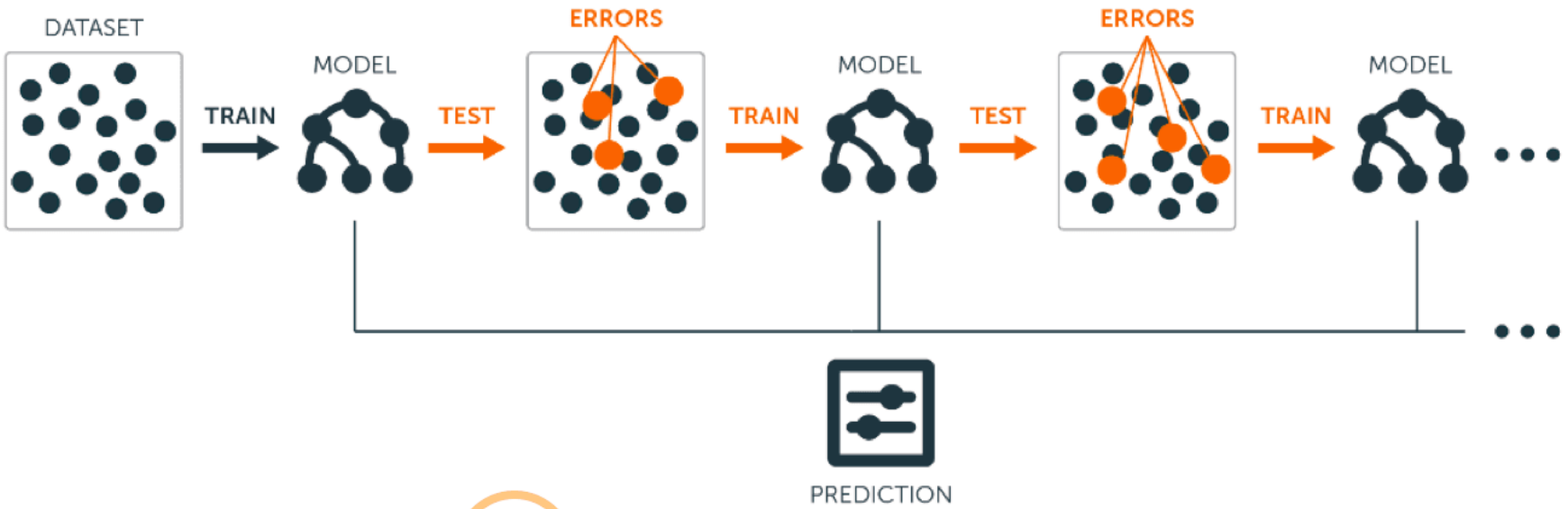
2.
Деревья решений



Эволюция:
Random forest – ансамбль деревьев

3.
Эволюция ансамблей:
Градиентный бустинг (XGBoost, LightGBM, CatBoost)

Детализация по градиентному бустингу



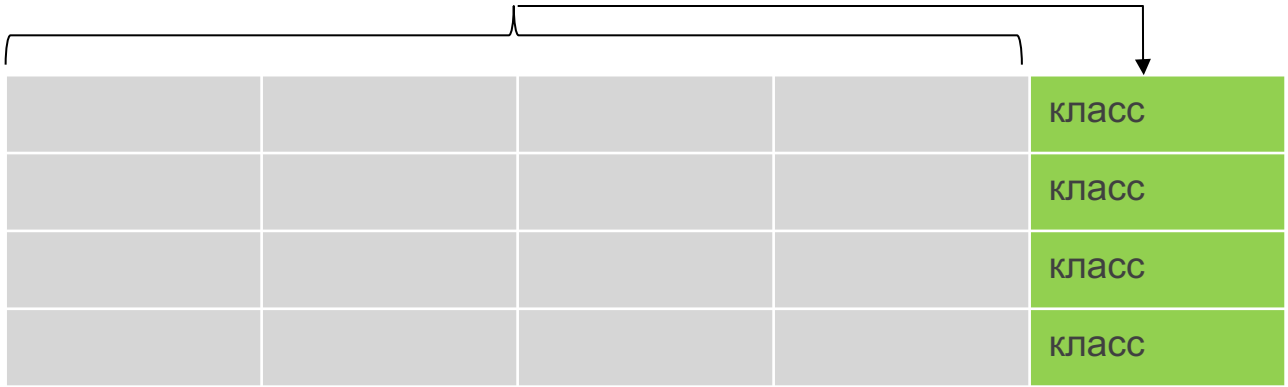
Нюансы

- **Выбросы (outliers)**
  Могут сильно влиять на параметры моделей, особенно линейных
- **Мультиколлинеарность признаков (созависимость)**
  Влияет на интерпретацию регрессионных коэффициентов.
- **Переобучение**
  Сложные модели (бустинг, нейронные сети) могут запоминать обучающую выборку и потом хуже перфомить на тесте
- **Недообучение**
  Слишком простые модели (простая линейная регрессия без учёта важных факторов) могут давать большие ошибки.
- **Правильная интерпретация**
  Даже высокий R2 не гарантирует причинно-следственную связь, а лишь корреляцию

Классификация

Definition

Классификация — это задача предсказания категориальной (качественной) переменной, то есть отнесение объекта к одному из нескольких классов



Примеры бизнес-кейсов

	Альфа Банк	Beeline	OZON
Тех задача	Кредитный скоринг	Определение фродовых звонков	Определение клиентов склонных к оттоку
Бизнес задача	Не дать кредит тому, кто его не вернет или поставить % окупающий риск	Не пропустить спам / мошенников до абонентов	Удержать клиентов (например дать скидку)

Как оценивать модели? (математика)

Accuracy
(точность классификации)

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

Precision, Recall, F1-score

$$\text{Precision} = \frac{TP}{TP + FP}$$

$$\text{Recall} = \frac{TP}{TP + FN}$$

$$\text{F1-score} = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}$$

ROC AUC

$$AUC = \int TPR \, d(FPR)$$

TP (True Positive)
TN (True Negative)
FP (False Positive)
FN (False Negative)

Как рисовать roc кривую кривую

id	оценка	класс
1	0.5	0
2	0.1	0
3	0.2	0
4	0.6	1
5	0.2	1
6	0.3	1
7	0.0	0

Табл. 1

id	оценка	класс
4	0.6	1
1	0.5	0
6	0.3	1
3	0.2	0
5	0.2	1
2	0.1	0
7	0.0	0

Табл. 2

id	> 0.25	класс
4	1	1
1	1	0
6	1	1
3	0	0
5	0	1
2	0	0
7	0	0

Табл. 3

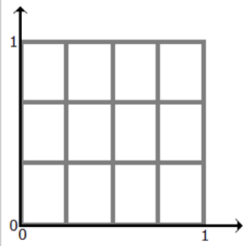
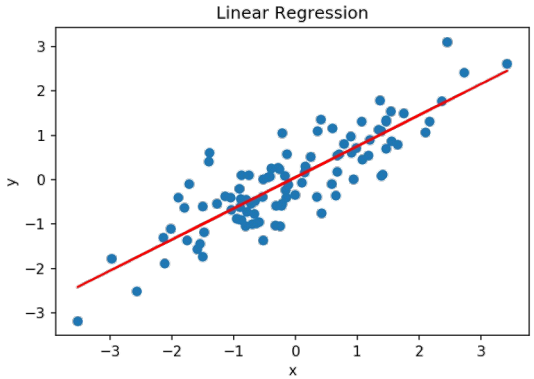


Рис.1. Построение ROC-кривой.

https://alexanderdyakonov.wordpress.com/2017/07/28/auc-roc-%D0%BF%D0%BB%D0%BE%D1%89%D0%B0%D0%B4%D1%8C-%D0%BF%D0%BE%D0%B4-%D0%BA%D1%80%D0%B8%D0%B2%D0%BE%D0%B9-%D0%BE%D1%88%D0%B8%D0%B1%D0%BE%D0%BA/

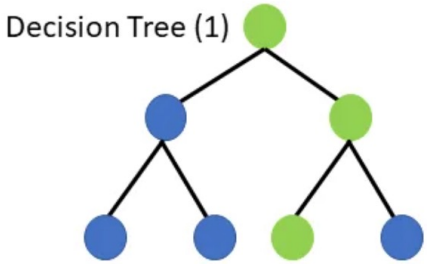
Основные алгоритмы

1.
Логистическая регрессия



Эволюция:
Регуляризованные решения для функции потерь
(Ridge, Lasso)

2.
Деревья решений



Эволюция:
Random forest – ансамбль деревьев

3.
Эволюция ансамблей: Градиентный бустинг (XGBoost, LightGBM, CatBoost)

4.
Метод опорных векторов (SVM)

Нюансы

- **Дисбаланс классов**
  Нужно рассматривать дополнительные метрики (Precision, Recall, AUC).
- **Выбор порога классификации**
  Меняя порог, можно балансировать между точностью и полнотой
- **Переобучение на некачественных данных.**
  Требуются методы регуляризации и кросс-валидация
- **Интерпретируемость моделей**
  Сложные модели (бустинг, нейронные сети) труднее объяснять