ИНТЕЛЛЕКТУАЛЬНОЕ МОДЕЛИРОВАНИЕ

Образовательная программа «6B06103 – Компьютерная инженерия»

Лекция 11

Метрики качества моделей регрессии и классификации

Неделя 11 · Tan, Steinbach, Kumar: 4.2, 5.7

Цель и план лекции

Цель: выбирать метрики, адекватные задаче, вычислять и интерпретировать их

1. Метрики регрессии: MAE, MSE, RMSE, MAPE, R^2
2. Матрица ошибок и accuracy
3. Precision, recall, specificity, F1
4. Порог, ROC-кривая и AUC
5. Многоклассовые задачи, базовая модель, scikit-learn

Зачем нужны метрики

- Оценить, годится ли модель для практики
- Сравнить модели и выбрать лучшую
- Подобрать гиперпараметры (неделя 12)
- Объяснить заказчику, чего ожидать

Оцениваем **ошибку обобщения** — на тестовой выборке или кросс-валидацией, а не на обучающих данных

Метрики регрессии

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

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

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n} \left|\frac{y_i - \hat{y}_i}{y_i}\right|$$

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

- MSE сильнее штрафует большие ошибки; $\mathrm{RMSE} \geq \mathrm{MAE}$
- MAPE неприменима при нулевых y
- $R^2 < 0$ на тесте — модель хуже среднего; для разного числа признаков — скорректированный R^2

Пример расчёта

y_i	$\hat{y}_i$	$\lvert e_i \rvert$	e_i^2	$\frac{\lvert e_i \rvert}{y_i}$
3	2,5	0,5	0,25	0,167
5	5,0	0,0	0,00	0,000
2	3,0	1,0	1,00	0,500
7	6,0	1,0	1,00	0,143
4	4,5	0,5	0,25	0,125

$\mathrm{MAE} = \mathbf{0{,}6}$

$\mathrm{MSE} = \mathbf{0{,}5}\ \cdot$
$\mathrm{RMSE} \approx \mathbf{0{,}707}$

$\mathrm{MAPE} \approx \mathbf{18{,}7\ \%}$

$R^2 = 1 - \frac{2{,}5}{14{,}8} \approx \mathbf{0{,}831}$

Какую метрику регрессии выбрать

Метрика	Единицы	Выбросы	Когда использовать
MAE	Как y y	Слабо	Понятная «средняя ошибка»
MSE	Квадрат y	Сильно	Функция потерь при обучении
RMSE	Как y y	Сильно	Акцент на больших промахах
MAPE	%	Средне	Разный масштаб величин, нет нулей
R^2	Безразм.	Сильно	Доля объяснённой изменчивости

Матрица ошибок

	Предсказан +	**Предсказан –**
Фактически +	TP — верно найден	FN — пропуск
Фактически –	FP — ложная тревога	TN — верно отвергнут

FN дороже

Медицина: пропустить болезнь опаснее

FP дороже

Спам-фильтр: важное письмо в спаме

Точность (accuracy) и её ловушка

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

$$\text{Error rate} = 1 - \text{Accuracy}$$

Мошеннических операций 1 %. Модель «всегда не мошенничество»:

99 %

accuracy

0

найденных мошенничеств

Precision, recall, specificity, F1

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

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

$$\text{Specificity} = \text{TNR} = \frac{\text{TN}}{\text{TN} + \text{FP}}, \quad \text{FPR} = 1 - \text{TNR}$$

$$F_1 = \frac{2 \cdot P \cdot R}{P + R} = \frac{2\text{TP}}{2\text{TP} + \text{FP} + \text{FN}}$$

- Precision и recall конфликтуют; F1 — их среднее гармоническое
- Для 1 и 5: арифм. 3, геом. 2,236, гарм. 1,667 — ближе к меньшему
- Fβ: β > 1 — важнее recall, β < 1 — precision

Пример: скрининг 1000 пациентов, 50 больны

	«болен»	«здоров»
Болен	$\mathrm{TP} = 40$	$\mathrm{FN} = 10$
Здоров	$\mathrm{FP} = 20$	$\mathrm{TN} = 930$

- $\mathrm{Accuracy} = \frac{970}{1000} = 0{,}970$
- $\mathrm{Recall} = \frac{40}{50} = 0{,}800$
- $\mathrm{Precision} = \frac{40}{60} \approx 0{,}667$
- $\mathrm{Specificity} = \frac{930}{950} \approx 0{,}979$
- $F_1 = \frac{80}{110} \approx 0{,}727$

97 % впечатляет, но модель «всегда здоров» дала бы 95 %

Какую метрику классификации выбрать

Метрика	Вопрос	Когда важна
Accuracy	Доля верных прогнозов?	Классы сбалансированы
Recall	Какую долю «+» нашли?	Пропуск дорог
Precision	Можно ли верить «+»?	Ложная тревога дорога
Specificity	Доля верно распознанных «−»?	Не беспокоить «здоровых»
F1	Баланс P и R	Несбалансированные классы
ROC AUC	Качество ранжирования?	Сравнение без порога

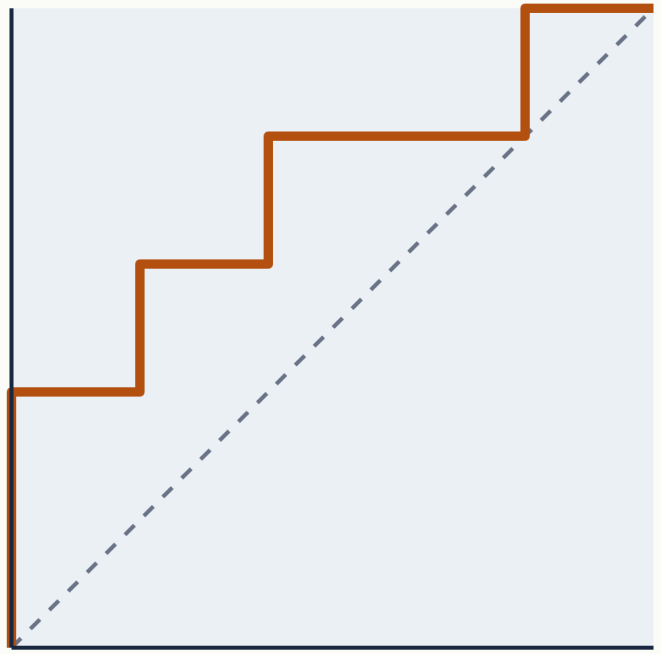
Порог классификации и ROC-кривая

- Модель выдаёт оценку; метка = оценка ≥ порог (по умолчанию 0,5)
- Ниже порог → выше recall, но больше ложных тревог
- **ROC-кривая:** TPR (ось y) против FPR (ось x) при всех порогах
- (0, 1) — идеал; диагональ — случайное угадывание

AUC — площадь под кривой: 1 у идеальной модели, 0,5 у случайной. Вероятность, что случайный «+» получит оценку выше случайного «–»

Построение ROC-кривой по шагам

```
Оценка: 0,95 0,85 0,78 0,66 0,60
Класс:    +    +    -    +    -
Оценка: 0,55 0,53 0,52 0,43 0,25
Класс:    +    -    -    +    -
```



- Опускаем порог по оценкам: «+» → вверх на 0,2, «–» → вправо на 0,2
- **AUC = 0,72**
- Порог 0,5: precision 0,5, recall 0,8
- Порог 0,7: precision ≈ 0,667, recall 0,4

Многоклассовые задачи и базовая модель

Матрица k × k; метрики считаются для каждого класса и усредняются:

macro

простое среднее — классы равноценны

weighted

взвешено по числу объектов

micro

по суммарным TP, FP, FN

Базовая модель: среднее (DummyRegressor) или частый класс (DummyClassifier). Не лучше базовой — бесполезна

Метрики в scikit-learn

```
mean_absolute_error(y_true, y_pred)     # 0.6
r2_score(y_true, y_pred)                # 0.831

pred = (scores ⩾ 0.5).astype(int)
confusion_matrix(y, pred)       # [[TN FP] [FN TP]]
precision_score(y, pred)        # 0.5
recall_score(y, pred)           # 0.8
f1_score(y, pred)               # 0.615
roc_auc_score(y, scores)        # 0.72
```

В scikit-learn первая строка матрицы — класс 0: [[TN, FP], [FN, TP]]

Несбалансированные данные (5 % класса 1)

0,948

accuracy базовой модели

0,947

accuracy логистической регрессии

0,873

ROC AUC — ранжирует хорошо

- Порог 0,5: recall по редкому классу всего **0,10**
- Порог 0,1: recall **0,61**, precision падает до **0,23**

Проблема не в модели, а в пороге; порог выбирают по цене ошибок

Выводы

- Качество оценивают на данных, не участвовавших в обучении
- Регрессия: MAE, MSE/RMSE, MAPE, R^2
- Классификация: матрица ошибок, precision, recall, F1
- ROC AUC оценивает ранжирование; порог — по цене ошибок
- Всегда сравнивайте с базовой моделью

Далее: кросс-валидация, регуляризация, подбор гиперпараметров

Контрольные вопросы

1. Формулы MAE, MSE, RMSE, R^2; чем они различаются?
2. Почему R^2 может быть отрицательным? Зачем скорректированный R^2 ?
3. Элементы матрицы ошибок; когда FN дороже FP и наоборот?
4. Почему accuracy обманчива при дисбалансе? Определите precision, recall, specificity, F1.
5. $\mathrm{TP} = 30$, $\mathrm{FN} = 20$, $\mathrm{FP} = 10$, $\mathrm{TN} = 940$: вычислите все метрики.
6. Постройте ROC-кривую; выведите матрицу и F1 при порогах 0,3, 0,5, 0,7 и ROC AUC.

Литература

1. Tan P.-N., Steinbach M., Kumar V. Introduction to Data Mining. — Pearson, 2014. — 4.2; 5.7.1–5.7.2; D.2.3.
2. Мюллер А., Гвидо С. Введение в машинное обучение с помощью Python. — 2017. — Глава 5.
3. Грас Дж. Data Science. Наука о данных с нуля. — 2020. — Глава 11.
4. Fawcett T. An introduction to ROC analysis // Pattern Recognition Letters. — 2006. — Vol. 27, № 8.
5. Документация scikit-learn: Metrics and scoring.