Шпаргалка по метрикам и проблемам переобучения / Концепции Cheatsheet (XeLaTeX)

Краткий справочник

Содержание

1 Метрики Оценки Задач Регрессии

Зачем нужны метрики регрессии?

Метрики — это численные показатели для **объективной оценки качества** модели регрессии (предсказание непрерывного значения). Они показывают, насколько хорошо предсказания модели ($\hat{y}_i$) соответствуют истинным значениям (y_i).

MAE (Mean Absolute Error) / Средняя Абсолютная Ошибка

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

Описание: Среднее абсолютное отклонение предсказаний от факта. *Интерпретация:* Показывает среднюю ошибку в единицах целевой переменной. Например, MAE = 10 означает, что модель в среднем ошибается на 10 единиц (рублей, градусов и т.д.). *Свойства:* Менее чувствительна к выбросам, чем MSE/RMSE. *Предпочтительна:* Когда важна прямая интерпретация ошибки и устойчивость к выбросам.

MSE (Mean Squared Error) / Среднеквадратичная Ошибка

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

Описание: Среднее квадратов отклонений предсказаний от факта. *Интерпретация:* Единицы измерения - квадрат исходных единиц (сложно интерпретировать напрямую). *Свойства:* Сильно штрафует за большие ошибки из-за возведения в квадрат. Дифференцируема, часто используется как функция потерь при обучении. *Предпочтительна:* Когда большие ошибки крайне нежелательны; удобна для оптимизации.

RMSE (Root Mean Squared Error) / Корень из Среднеквадратичной Ошибки

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

Описание: Корень квадратный из MSE. *Интерпретация:* Возвращает метрику к исходным единицам измерения (как MAE), что упрощает интерпретацию. RMSE всегда больше или равен MAE. *Свойства:* Чувствительна к выбросам (но меньше, чем MSE). Штрафует большие ошибки сильнее, чем MAE. *Предпочтительна:* Когда нужна интерпретация в исходных единицах, но с большим штрафом за большие ошибки, чем у MAE.

R^2 (Коэффициент Детерминации)

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

Описание: Доля дисперсии зависимой переменной (y), объясненная моделью. *Интерпретация:* Варьируется от $(-\infty)$ до 1. Ближе к 1 — лучше. $R^2 = 0.7$ означает, что модель объясняет 70*Свойства:* Увеличивается при добавлении любых признаков (даже бесполезных). Используйте **Adjusted R^2** для учета количества признаков при сравнении моделей. *Предпочтительна:* Для оценки объяснительной силы модели.

2 Метрики Оценки Задач Классификации

2.1. Матрица ошибок (Confusion Matrix)

Основа для метрик бинарной классификации

Матрица ошибок показывает распределение предсказаний модели по сравнению с истинными метками классов. Обычно класс "1" считается положительным (positive, e.g., "болезнь", "спам", "мошенничество"), а класс "0" - отрицательным (negative).

- **TP (True Positive)**: Истинно положительные. Объект класса 1, предсказан как 1. (Верно найден больной).
- **TN (True Negative)**: Истинно отрицательные. Объект класса 0, предсказан как 0. (Верно найден здоровый).
- **FP (False Positive)**: Ложно положительные. **Ошибка I рода**. Объект класса 0, предсказан как 1. (Здоровый ошибочно признан больным).
- **FN (False Negative)**: Ложно отрицательные. **Ошибка II рода**. Объект класса 1, предсказан как 0. (Больной ошибочно признан здоровым).

Структура Матрицы Ошибок:

		Предсказанный класс		**Всего (**
		1 (Positive)	**0 (Negative)**	
Истинный класс	**1 (Positive)**	TP	FN	P =
	0 (Negative)	FP	TN	N =
	Всего (Предсказание)	P TP + FP	N FN + TN	**Tota**

2.2. Основные Метрики Классификации

Accuracy (Доля правильных ответов)

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

Описание: Общая доля верных предсказаний модели. *Проблема:* **Неинформативна при сильном дисбалансе классов!** Модель, всегда предсказывающая мажоритарный класс, будет иметь высокую Accuracy, но будет бесполезна.

Precision (Точность)

$$\text{Precision} = \frac{TP}{TP + FP} = \frac{\text{Верно найденные позитивные}}{\text{Все объекты, названные позитивными}}$$

Вопрос: "Из тех, кого мы назвали классом 1, какая доля действительно принадлежит классу 1?" *Важность:* Высока, когда цена **FP** (Ошибка I рода) велика (e.g., неверный диагноз здоровому, блокировка честного клиента).

Recall (Полнота, Sensitivity, True Positive Rate - TPR)

$$\text{Recall} = \frac{TP}{TP + FN} = \frac{\text{Верно найденные позитивные}}{\text{Все реальные позитивные объекты (P)}}$$

Вопрос: "Какую долю реальных объектов класса 1 мы смогли обнаружить?" *Важность:* Высока, когда цена **FN** (Ошибка II рода) велика (e.g., пропуск больного, пропуск мошенника).

F1-мера (F1-Score)

$$F_1 = 2 \cdot \frac{\text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}} = \frac{2TP}{2TP + FP + FN}$$

Описание: Гармоническое среднее Precision и Recall. Полезна, когда важен баланс между ними. *Свойство:* Близка к нулю, если хотя бы одна из компонент (Precision или Recall) близка к нулю. *Обобщение:* F_β**-мера** позволяет придать больший вес Recall ($\beta > 1$) или Precision ($\beta < 1$).

$$F_\beta = (1 + \beta^2) \cdot \frac{\text{Precision} \cdot \text{Recall}}{(\beta^2 \cdot \text{Precision}) + \text{Recall}}$$

Specificity (Специфичность, True Negative Rate - TNR)

$$\text{Specificity} = \frac{TN}{TN + FP} = \frac{\text{Верно найденные негативные}}{\text{Все реальные негативные объекты (N)}}$$

Вопрос: "Какую долю реальных объектов класса 0 мы правильно определили как класс 0?" *Связь:* Часто используется в паре с Recall (Sensitivity) в медицине.

False Positive Rate (FPR)

$$\text{FPR} = \frac{FP}{TN + FP} = \frac{\text{Ложно названные позитивными}}{\text{Все реальные негативные объекты (N)}} = 1 - \text{Specificity}$$

Вопрос: "Какую долю объектов класса 0 модель ошибочно назвала классом 1?" *Использование:* Используется как ось X в ROC-кривой.

Компромисс Precision-Recall

Часто существует обратная зависимость между Precision и Recall при изменении порога классификации модели.

- Увеличение порога $\Longrightarrow$ меньше объектов объявляются классом 1 $\Longrightarrow$ FP уменьшается (растет Precision), но TP тоже может уменьшиться (падает Recall).
- Уменьшение порога $\Longrightarrow$ больше объектов объявляются классом 1 $\Longrightarrow$ TP растет (растет Recall), но FP тоже может вырасти (падает Precision).

Выбор оптимального порога зависит от задачи и баланса между Precision и Recall (часто максимизируют F1 или выбирают порог на PR-кривой).

2.3. Метрики для Оценки Ранжирования

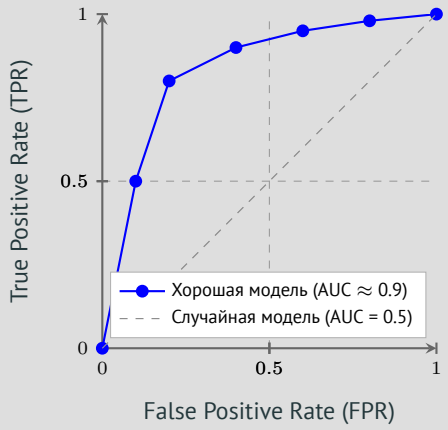
ROC AUC (Receiver Operating Characteristic Area Under Curve)

Метрика, оценивающая качество модели как бинарного классификатора **независимо от порога классификации**. Показывает, насколько хорошо модель способна **ранжировать** объекты.

- **ROC-кривая**: График **TPR (Recall)** vs **FPR** при изменении порога от 1 до 0.
- **AUC**: Площадь под ROC-кривой (от 0 до 1).
  - AUC = 1: Идеал.
  - AUC = 0.5: Случайность.
  - AUC < 0.5: Хуже случайности.
- **Интерпретация AUC**: Вероятность, что случайный объект класса 1 получит скор выше, чем случайный объект класса 0.
- **Свойства**: Относительно устойчив к дисбалансу классов. Сравнивает модели по общей ранжирующей способности.

Аналогия: Насколько хорошо модель отделяет "больных" от "здоровых" при сортировке.

Пример ROC-кривой



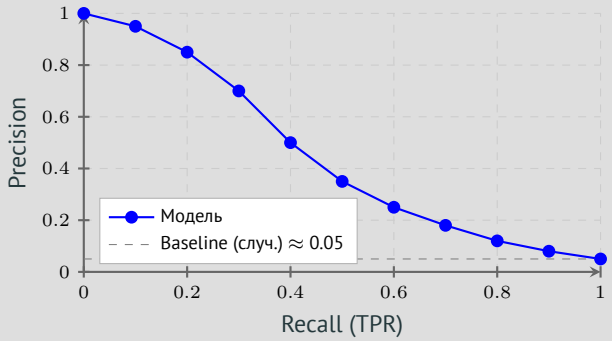
Precision-Recall AUC (PR AUC)

Альтернатива ROC AUC, особенно полезная при **сильном дисбалансе классов** и фокусе на **положительном (миноритарном) классе**.

- **PR-кривая**: График **Precision** vs **Recall (TPR)** при изменении порога.
- **PR AUC**: Площадь под PR-кривой (от 0 до 1). Выше - лучше.
- **Приоритет при дисбалансе**: Более чувствительна к FP, чем ROC AUC. Сравнивает модели по способности находить позитивные примеры с высокой точностью.
- **Baseline**: Равен доле позитивного класса $P/(P+N)$. Низкий baseline при дисбалансе делает PR AUC более показательным для оценки улучшения.

Аналогия: Насколько точно мы находим "иголки" (класс 1), когда пытаемся найти разную их долю (Recall) в "стоге сена" (все данные)?

Пример PR-кривой (Сильный дисбаланс)



2.4. Выбор Метрики Классификации

Ключевой Аспект: Бизнес-Задача

Выбор основной метрики **критически зависит от бизнес-задачи** и стоимости ошибок FP и FN.

- **Задача: Диагностика опасной болезни** *Цель:* Не пропустить больных (максимизировать **Recall**). *Метрики:* Recall, F-мера ($\beta > 1$), PR AUC.
- **Задача: Спам-фильтр** *Цель:* Не терять важные письма (максимизировать **Precision**). *Метрики:* Precision, F-мера ($\beta < 1$).
- **Задача: CTR prediction** *Цель:* Точное ранжирование и калибровка вероятностей. *Метрики:* **ROC AUC**, **LogLoss**.
- **Задача: Поиск мошенников (дисбаланс)** *Цель:* Баланс между поиском мошенников (Recall) и точностью (Precision). Accuracy бесполезна. *Метрики:* **F1-Score**, **PR AUC**, анализ матрицы ошибок.

Вывод: Всегда обсуждайте с заказчиком последствия ошибок FP и FN для выбора адекватной метрики. Часто нужно отслеживать несколько метрик.

2.5. Offline vs Online Метрики

Где и как измеряем качество

- **Offline метрики**: Расчет на отложенной выборке (Test set, CV folds). Примеры: Accuracy, F1, AUC, MSE. Назначение: разработка, валидация, выбор модели.
- **Online метрики**: Расчет на реальных данных в production. Примеры: CTR, конверсия, доход. Назначение: оценка реального бизнес-эффекта, A/B тестирование.

Важно: Улучшение offline-метрик не всегда гарантирует улучшение online-метрик.

3 Статистические Основы Оценки

Зачем нужна статистика в ML?

Работа с **ограниченными выборками** данных означает, что любая посчитанная метрика — это лишь **оценка** истинного значения. Статистика помогает:

- Оценить **неопределенность** измерений (точность оценки).
- Проверить **надежность** выводов (значимость различий).
- Принимать обоснованные решения.

3.1. Дисперсия и Стандартное Отклонение Данных

Измерение Разброса Данных

Дисперсия (Variance, s^2): Мера разброса значений относительно среднего ($\bar{x}$).

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2$$

Стандартное отклонение (Standard Deviation, SD, s): Корень из дисперсии ($s = \sqrt{s^2}$). Измеряется в исходных единицах.
Интуиция: Низкая дисперсия/SD $\implies$ данные сгруппированы. Высокая $\implies$ данные разбросаны.
Применение в ML: Анализ изменчивости данных, оценка стабильности метрик на CV (SD по фолдам), компонент R^2, стандартной ошибки (SE).

3.2. Доверительные Интервалы (Confidence Intervals, CI)

Что такое Доверительный Интервал?

CI — это диапазон значений, который с заданной **уверенностью** (e.g., 95%) содержит **истинное значение** параметра (e.g., истинный средний AUC), оцененного по выборке.
Построение (схематично):

$$CI = [\text{Оценка} - \text{Погрешность}, \quad \text{Оценка} + \text{Погрешность}]$$

Погрешность (Margin of Error) = Крит. Значение$\times$Стандартная Ошиб

- **Оценка**: Значение метрики по выборке (e.g., среднее по фолдам CV).
- **Крит. Значение**: Зависит от уровня доверия и распределения (e.g., ≈ 1.96 для 95% CI, нормальное распределение).
- **SE (Standard Error)**: Точность оценки (e.g., $SE = SD/\sqrt{n}$ для среднего).

Интерпретация и Применение CI в ML

Интерпретация 95% CI [A, B]: "Мы на 95% уверены, что истинное значение параметра находится между A и B."
Применение:

- **Оценка надежности:** Широкий CI $\implies$ большая неопределенность.
- **Сравнение моделей/групп:** Строим CI для **разницы** метрик

(metric$_B$ − metric$_A$).

- Если CI **не включает 0** (e.g., [0.01, 0.05]): разница статистически значима на выбранном уровне.
- Если CI **включает 0** (e.g., [-0.02, 0.03]): нет оснований утверждать о значимой разнице.

3.3. Проверка Статистических Гипотез и p-value

Процесс Проверки Статистических Гипотез

Цель: определить, является ли наблюдаемый эффект реальным или случайным. **Шаги:**

1. **Формулировка Гипотез:**
   - H_0 (Нулевая): Гипотеза об отсутствии эффекта (e.g., $\mu_A = \mu_B$).
   - H_1 (Альтернативная): Гипотеза о наличии эффекта (e.g., $\mu_A \neq \mu_B$ или $\mu_B > \mu_A$).
2. **Выбор Уровня Значимости (α):** Порог ошибки I рода (обычно 0.05).
3. **Расчет Статистики Теста:** Мера отклонения данных от H_0.
4. **Расчет p-value.**

P-value (P-значение)

Определение: Вероятность получить наблюдаемые данные (или еще более экстремальные), **если предположить, что H_0 верна**.

Правило принятия решения:

- **p-value < α**: Отвергаем H_0. Результат статистически значим. (Данные маловероятны при H_0).
- **p-value $\geq \alpha$**: Не отвергаем H_0. Результат не является статистически значимым. (Данные совместимы с H_0).

Важно: Не отвергнуть H_0 не значит доказать её истинность!

Типы Ошибок при Проверке Гипотез

- **Ошибка I рода (α)**: Отвергнуть H_0, когда она верна (False Positive). Вероятность $\leq \alpha$.
- **Ошибка II рода (β)**: Не отвергнуть H_0, когда она ложна (False Negative).
- **Мощность теста (Power = $1 - \beta$)**: Вероятность правильно отвергнуть ложную H_0 (обнаружить реальный эффект).

3.4. Основные Статистические Тесты для Сравнения

Выбор подходящего теста

Зависит от: цели, числа групп, зависимости выборок, типа и распределения данных, размера выборки.

Z-тест (для средних или долей)

- **Применение**: Сравнение долей (конверсий, CTR) в A/B тестах.
- **Предположения**: Очень большие выборки ($n > 30..50$, часто тысячи).

T-тест (для средних)

- **Применение**: Сравнение средних (метрики, бизнес-показатели).
- **Предположения**: Данные примерно нормальны (или $n \geq 15..30$); дисперсии неизвестны. Используется Welch's t-test по умолчанию для независимых выборок.
- **Виды:**
  - *Независимый*: Сравнение средних 2х независимых групп (A/B тест).
  - *Парный*: **Критичен для ML!** Сравнение метрик 2х моделей на **одних и тех же фолдах CV**. Тестирует H_0 : $\mu_{\text{diff}} = 0$.

Тест Манна-Уитни (Mann-Whitney U / Wilcoxon Rank-Sum)

- **Применение**: Сравнение распределений/медиан 2х **независимых** выборок. Непараметрический аналог независимого t-теста.
- **Предположения**: Независимые выборки, данные мин. порядковые. **Нормальность не требуется!**
- **Использование в ML**: A/B тесты с ненормальными данными (время на сайте, доход).
- **Парный аналог (для ненормальных парных данных)**: **Тест Уилкоксона для связанных выборок (Wilcoxon signed-rank test)**. Можно использовать для сравнения моделей на фолдах CV, если разности метрик не нормальны.

Резюме по Статистике

Стат. инструменты (CI, тесты) нужны для оценки надежности метрик и значимости различий между моделями/группами.

4 Практические Аспекты Валидации и Оценки

4.1. Кросс-валидация (Cross-Validation, CV)

Общая Идея Кросс-валидации

Метод оценки обобщающей способности модели на независимых данных и получения более надежной оценки метрики, чем на единственном test-сплите. Используется для оценки модели и настройки гиперпараметров.

- **Принцип:** Обучающая выборка многократно делится на train fold и validation fold. Модель обучается на train fold, оценивается на validation fold. Процесс повторяется.
- **Результат:** Набор оценок метрики (по одной на фолд). Итоговая оценка - среднее по фолдам. Стандартное отклонение по фолдам показывает стабильность.

K-Fold CV

- **Метод:** Данные делятся на K фолдов. На итерации k, модель обучается на $K-1$ фолдах, валидируется на k-ом фолде. Повторяется K раз.
- **Параметры:** Обычно $K = 5$ или $K = 10$.
- **Применение:** Стандартный метод для регрессии и классификации (при балансе классов).

Stratified K-Fold CV

- **Метод:** Модификация K-Fold для **классификации**. Гарантирует сохранение **соотношения классов** в каждом фолде.
- **Применение: Обязательно использовать при дисбалансе классов**. Рекомендуется по умолчанию для задач классификации.

Leave-One-Out CV (LOOCV)

- **Метод:** K-Fold с $K = n$ (число объектов). Обучение на $n-1$, валидация на 1.
- **Плюсы:** Почти несмещенная оценка ошибки.
- **Минусы:** Вычислительно дорого, высокая дисперсия оценки.
- **Применение:** Очень маленькие датасеты.

Time Series CV (Кросс-валидация для Временных Рядов)

- **Проблема:** Стандартный K-Fold нарушает временную структуру ("заглядывание в будущее").
- **Методы:** Используются схемы, сохраняющие порядок данных:
  - **Rolling / Sliding Window:** Обучение на окне данных, валидация на следующем блоке. Окно сдвигается вперед.
  - **Expanding Window:** Обучение на всех данных до точки t, валидация на следующем блоке. Обучающая выборка растет.
- **Применение:** Задачи прогнозирования временных рядов.

4.2. Работа с Несбалансированными Данными

Проблема Дисбаланса Классов

Ситуация значительного преобладания одного класса над другим(и). **Основные Проблемы:** Неинформативность Accuracy, игнорирование миноритарного класса моделью, проблемы с валидацией (K-Fold).

Подход 1: Использовать Адекватные Метрики

Фокусироваться на метриках, чувствительных к ошибкам на миноритарном классе:

- **Precision, Recall, F1-Score**
- **ROC AUC** (но может быть обманчив при сильном дисбалансе)
- **PR AUC** (часто более предпочтителен при сильном дисбалансе)
- Анализ **Матрицы ошибок**.

Подход 2: Изменение Выборки (Resampling)

Применяется **только к обучающим данным внутри фолдов CV!**

- **Undersampling**: Удаление части мажоритарного класса (Random, NearMiss). *Риск:* Потеря информации.
- **Oversampling**: Увеличение миноритарного класса.
  - *Random Oversampling:* Дублирование. *Риск:* Переобучение.
  - *SMOTE (и варианты)*: Генерация синтетических миноритарных объектов. Часто предпочтительнее Random Oversampling.

Важнейшее правило ресемплинга

Методы ресемплинга (Under/Over-sampling, SMOTE) должны применяться **только к обучающей части данных внутри каждого фолда CV**. **Никогда** не применяйте их ко всей выборке до CV или к validation/test set! Это ведет к **утечке данных** и завышенным оценкам.

Подход 3: Взвешивание Классов/Примеров (Weighting)

Назначение большего веса объектам миноритарного класса в функции потерь модели.

- Заставляет модель уделять больше внимания ошибкам на миноритарном классе.
- Поддерживается многими алгоритмами (LogReg, SVM, Trees, Boostings).
- Параметры: `class_weight='balanced'` (авто-подбор), `scale_pos_weight` (ручной множитель для позитивного класса в бустингах).
- Часто проще и эффективнее ресемплинга.

Подход 4: Использование Алгоритмов, Устойчивых к Дисбалансу

- Некоторые модели лучше справляются по умолчанию (e.g., ансамбли деревьев).
- Существуют специализированные ансамблевые методы (EasyEnsemble, BalanceCascade).

Подход 5: Изменение Порога Классификации

- Если модель выдает вероятности, можно подобрать оптимальный порог (не обязательно 0.5) для бинаризации.
- Порог выбирается на валидационной выборке для достижения нужного баланса Precision/Recall (e.g., максимизация F1).

Подход 6: Использовать Stratified K-Fold

Как уже упоминалось, Stratified K-Fold гарантирует репрезентативность классов в фолдах CV, что критически важно при дисбалансе.