Метрики качества модели

Qwen2.5-Max
суб, 15-фев-2025

1. Матрица ошибок

Метрики классификации, такие как **точность (Accuracy)**, **полнота (Recall)**, **точность (Precision)** и **F1-мера**, используются для оценки качества модели машинного обучения в задачах классификации. Каждая из этих метрик имеет свое значение и интерпретацию, зависящую от контекста задачи.
Матрица ошибок — это таблица, которая показывает, как модель классифицировала объекты по сравнению с их истинными метками. Для бинарной классификации матрица выглядит так:

	Предсказано положительно (Positive)	Предсказано отрицательно (Negative)
Истинно положительно (True Positive, TP)	TP	FN
Истинно отрицательно (True Negative, TN)	FP	TN

где:
- **TP (True Positives)**: Количество объектов, которые действительно принадлежат к положительному классу и были правильно классифицированы.
- **TN (True Negatives)**: Количество объектов, которые действительно принадлежат к отрицательному классу и были правильно классифицированы.
- **FP (False Positives)**: Количество объектов, которые не принадлежат к положительному классу, но были ошибочно классифицированы как положительные.
- **FN (False Negatives)**: Количество объектов, которые принадлежат к положительному классу, но были ошибочно классифицированы как отрицательные.

2. Основные метрики качества

1. Точность (Accuracy)

Точность — это доля правильных предсказаний среди всех сделанных предсказаний.

Формула:

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

Интерпретация:

- Точность показывает общее качество модели по всем классам.
- Хорошо работает, когда данные сбалансированы (т.е., примеры разных классов распределены равномерно).
- Если данные несбалансированы (например, один класс значительно преобладает), точность может быть обманчивой. Например, если модель всегда предсказывает

преобладающий класс, она может показать высокую точность, но при этом плохо работать на меньшем классе.

2. Точность (Precision)

Точность — это доля верно предсказанных положительных примеров среди всех примеров, которые модель пометила как положительные.

Формула:

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

Интерпретация:

- Показывает, насколько надёжны предсказания положительного класса.
- Высокая точность означает, что большинство объектов, отнесённых моделью к положительному классу, действительно принадлежат этому классу.
- Важна в ситуациях, где важно минимизировать ложноположительные ошибки (FP). Например, в диагностике болезней или фильтрации спама.

3. Полнота (Recall)

Полнота — это доля верно предсказанных положительных примеров среди всех реальных положительных примеров.

Формула:

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

Интерпретация:

- Показывает, насколько хорошо модель находит все положительные примеры.
- Высокая полнота означает, что модель редко пропускает положительные примеры (FN).
- Важна в ситуациях, где важно минимизировать ложноотрицательные ошибки (FN). Например, в диагностике серьезных заболеваний, где важно найти все больных пациентов.

4. F1-мера

F1-мера — это гармоническое среднее между точностью и полнотой. Она учитывает оба показателя и помогает сбалансировать их.

Формула:

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

Интерпретация:

- F1-мера особенно полезна, когда важны как точность, так и полнота.
- Хороша для несбалансированных данных, где одна из метрик (точность или полнота) может быть более значимой.
- Чем ближе F1-мера к 1, тем лучше модель.

5. Specificity (Специфичность)

Specificity измеряет долю верно предсказанных отрицательных примеров среди всех реальных отрицательных примеров.

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

- **Интерпретация**: Насколько хорошо модель распознает отрицательный класс.
- **Применение**: Важна в задачах, где важно минимизировать ложноположительные ошибки.

6. ROC-AUC (Receiver Operating Characteristic- Area Under Curve)

ROC-AUC — это площадь под кривой ROC, которая строится на основе двух параметров:

- **True Positive Rate (TPR)**: То же, что и Recall ($\text{TPR} = \frac{\text{TP}}{\text{TP+FN}}$).
- **False Positive Rate (FPR)**: Доля ложноположительных предсказаний среди всех реальных отрицательных примеров ($\text{FPR} = \frac{\text{FP}}{\text{FP+TN}}$).
- **Интерпретация**: Чем ближе AUC к 1, тем лучше модель различает классы.
- **Применение**: Используется для оценки качества работы модели на разных порогах классификации.

7. Log Loss (Логарифмическая потеря)

Log Loss измеряет качество предсказаний вероятностей. Она штрафует модель за уверенность в неправильных предсказаниях.

$$\text{Log Loss} = -\frac{1}{N}\sum_{i=1}^{N}[y_i \cdot \log(p_i) + (1 - y_i) \cdot \log(1 - p_i)]$$

где:

- N — количество объектов,
- y_i — истинная метка (0 или 1),
- p_i — предсказанная вероятность принадлежности к положительному классу.

Интерпретация: Чем меньше Log Loss, тем лучше модель.
Применение: Полезна для моделей, которые предсказывают вероятности (например, логистическая регрессия).

Сравнение метрик:

Метрика	Что показывает?	Когда важна?
Accuracy	Общее качество модели по всем классам	При сбалансированных данных
Precision	Насколько надёжны предсказания положительного класса	Когда важно минимизировать ложноположительные ошибки (FP)
Recall	Насколько хорошо модель находит все положительные примеры	Когда важно минимизировать ложноотрицательные ошибки (FN)
F1-мера	Сбалансированное сочетание точности и полноты	При необходимости сбалансировать точность и полноту (особенно при несбалансированных данных
Specificity	Насколько надёжны предсказания отрицательного класса	Когда важно минимизировать ложноположительные ошибки (FP)
ROC-AUC	Для оценки общей способности модели различать классы	
Log Loss	Для оценки качества предсказаний вероятностей	

Пример

Рассмотрим задачу диагностики болезни:
- **True Positives (TP):** 90 человек правильно диагностированы как больные.
- **False Positives (FP):** 10 здоровых людей ошибочно диагностированы как больные.
- **False Negatives (FN):** 5 больных людей ошибочно диагностированы как здоровые.
- **True Negatives (TN):** 900 здоровых людей правильно диагностированы как здоровые.

Вычисления:

1. **Accuracy:**

$$\text{Accuracy} = \frac{90 + 900}{90 + 10 + 900 + 5} = \frac{990}{1005} \approx 0.985$$

2. **Precision:**

$$\text{Precision} = \frac{90}{90 + 10} = \frac{90}{100} = 0.9$$

3. **Recall:**

$$\text{Recall} = \frac{90}{90 + 5} = \frac{90}{95} \approx 0.947$$

4. **F1-мера:**

$$F1 = 2 \cdot \frac{0.9 \cdot 0.947}{0.9 + 0.947} \approx 2 \cdot \frac{0.852}{1.847} \approx 0.916$$

5. **Specificity:**

$$\text{Specificity} = \frac{900}{900 + 10} = \frac{900}{910} \approx 0.989$$

Интерпретация:

- Точность (Accuracy) высока (98.5%), но это может быть обманчиво, если мы хотим убедиться, что все больные люди найдены.
- Точность (Precision) показывает, что модель редко ошибается, предсказывая болезнь.
- Полнота (Recall) показывает, что модель находит почти всех больных людей (94.7%).
- F1-мера (91.6%) дает сбалансированную оценку, учитывая оба аспекта.
- Специфичность (Specificity) показывает, что модель редко ошибается, предсказывая, что человек здоров (98.9%).

Заключение

Каждая из метрик имеет свою область применения, и выбор конкретной метрики зависит от целей задачи. Для задач с несбалансированными данными рекомендуется использовать **Precision**, **Recall** и **F1-меру**, чтобы получить более полную картину работы модели. Например:
- Если важны оба класса, используйте **F1-меру**.
- Если важен только один класс, используйте **Precision** или **Recall**.
- Если нужно оценить общее качество, используйте **Accuracy** или **ROC-AUC**.

Accuracy, Precision, Recall и F1-мера дополняют друг друга и позволяют оценить разные аспекты качества модели.