ГУАП

Основы машинного обучения

Поляк Марк Дмитриевич

2025

Оценивание качества моделей

Лекция 9

Оценки качества классификации

Чувствительность, специфичность, ROC, AUC

Оценка результата внедрения модели машинного обучения

Иерархия метрик

1. Верхний уровень: повышение эффективности бизнеса. Например, рост дохода/прибыли. Невозможно измерить в моменте
2. Количественные показатели «удовлетворенности» пользователя. Например, длительность сессии, средний чек и т.п. Косвенно влияют на критерий верхнего уровня
3. Доля удовлетворенных качеством предсказаний модели асессоров, на которых модель протестирована до выставления на суд пользователей
4. Функция потерь, использованная при обучении модели

Online-метрики вычисляются по данным, собираемым с работающей системы.

Offline-метрики могут быть измерены до введения модели в эксплуатацию, например, по историческим данным или с привлечением специальных людей, асессоров.

Оценка качества модели

- Функционал качества: эмпирический риск $Q_{ERM}(\boldsymbol{\theta})$, правдоподобие $Q_{MLE}(\boldsymbol{\theta})$)
- Функция потерь $\mathcal{L}(\boldsymbol{\theta}, x_i)$: связана с решением задачи оптимизации
- Метрика качества: внешний, объективный критерий качества, обычно зависящий не от параметров модели, а только от предсказанных меток.

Функция потерь ≠ метрика качества

Но не всегда. Например, в задаче регрессии MSE может быть как метрикой, так и функцией потерь.

Анализ ошибок классификации

Задача классификации на два класса: $y_i, g(x_i) \in \{-1, +1\}$

	Модель классификации	учитель
TP, True Positive	$g(x_i) = +1$	$y_i = +1$
TN, True Negative	$g(x_i) = -1$	$y_i = -1$
FP, False Positive	$g(x_i) = +1$	$y_i = -1$
FN, False Negative	$g(x_i) = -1$	$y_i = +1$

Матрица ошибок:

n=165	Predicted: NO	Predicted: YES	
Actual: NO	TN = 50	FP = 10	60
Actual: YES	FN = 5	TP = 100	105
	55	110	

FP: ложноположительно, ошибка I рода, «ложная тревога»

FN: ложноотрицательно, ошибка II рода, «пропуск цели»

Правильность классификации (чем больше, тем лучше):

$$Accuracy = \frac{1}{\ell}\sum_{i=1}^{\ell}[g(x_i) = y_i] = \frac{TP + TN}{FP + FN + TP + TN}$$

Недостаток: не учитывается дисбаланс численности классов, различие цены ошибки I и II рода

Функции потерь, зависящие от штрафов за ошибку

Задача классификации на два класса: $y_i \in \{-1, +1\}$

Модель классификации $g(x; \boldsymbol{\theta}, \theta_0) = \text{sign}(h(x, \boldsymbol{\theta}) - \theta_0)$

Чем больше θ_0, тем больше x_i таких, что $g(x_i) = -1$.

Пусть λ_y – штраф за ошибку на объекте класса y.

Функция потерь теперь зависит от штрафов:

$$\mathcal{L}(\boldsymbol{\theta}, x_i) = \lambda_{y_i}[g(x_i; \boldsymbol{\theta}, \theta_0) \neq y_i] = \lambda_{y_i}[(g(x_i, \boldsymbol{\theta}) - \theta_0)y_i < 0]$$

Проблема

На практике штрафы $\{\lambda_y\}$ могут пересматриваться

- Нужен удобный способ выбора θ_0 в зависимости $\{\lambda_y\}$, не требующий построения модели (поиска вектора $\boldsymbol{\theta}$) заново.
- Нужна характеристика качества модели $h(x, \boldsymbol{\theta})$, не зависящая от штрафов $\{\lambda_y\}$ и численности классов.

Определение ROC-кривой

Кривая ошибок ROC (receiver operating characteristic).

Каждая точка кривой соответствует некоторому $g(x; \boldsymbol{\theta}, \theta_0)$

- по оси X: доля ошибочных положительных классификаций (FPR – false positive rate)

$$FPR = \frac{FP}{FP + TN} = \frac{\sum_{i=1}^{\ell}[y_i = -1][g(x; \boldsymbol{\theta}, \theta_0) = +1]}{\sum_{i=1}^{\ell}[y_i = -1]}$$

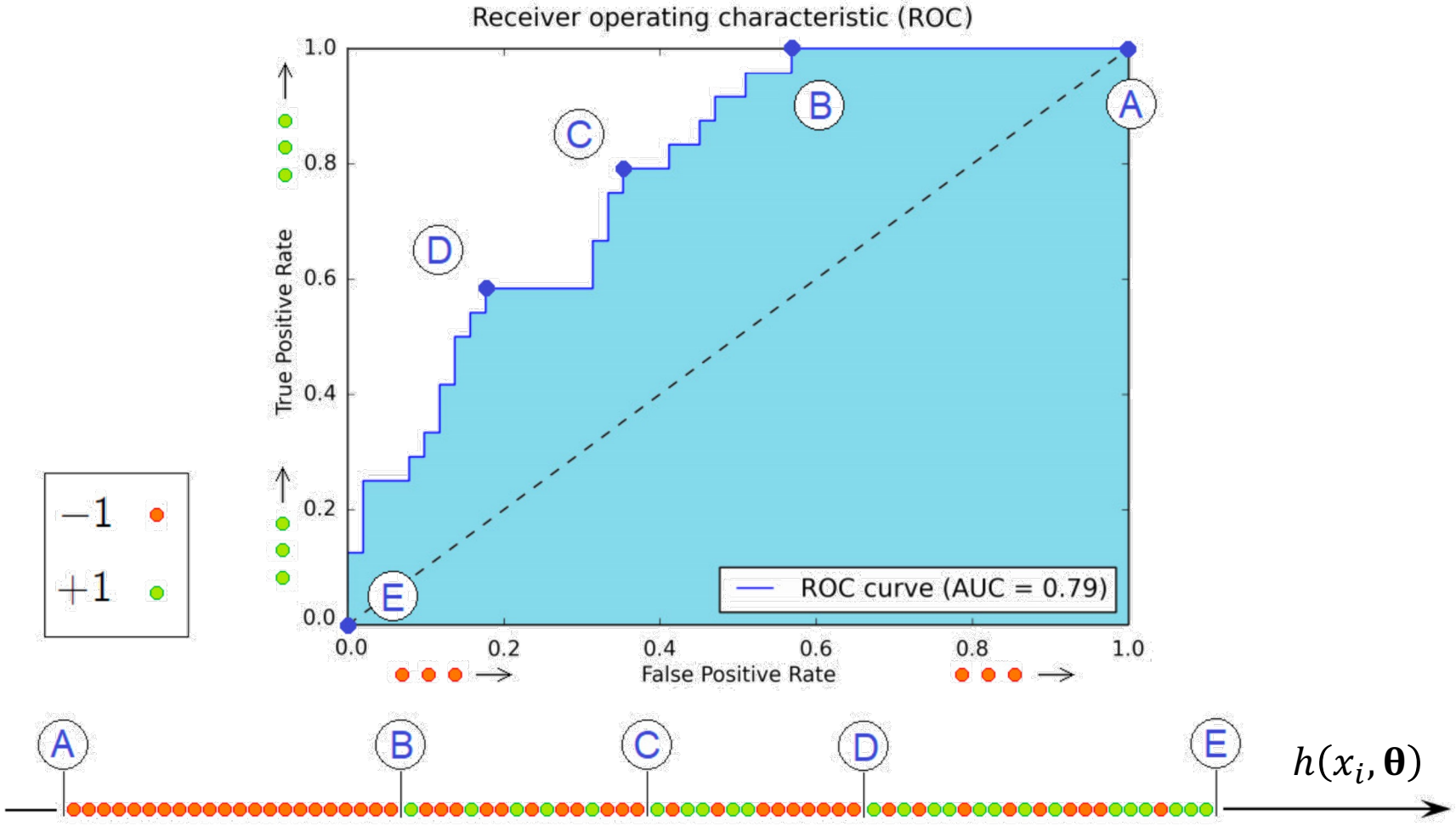
  $1 - FPR$ называется *специфичностью* алгоритма g

- по оси Y: доля правильных положительных классификаций (TPR – true positive rate)

$$TPR = \frac{TP}{TP + FN} = \frac{\sum_{i=1}^{\ell}[y_i = +1][g(x; \boldsymbol{\theta}, \theta_0) = +1]}{\sum_{i=1}^{\ell}[y_i = +1]}$$

TPR называется также *чувствительностью* алгоритма g

ROC-кривая и площадь под ней AUC (Area Under Curve)



ABCDE – положение порога θ_0 на оси значений функции h

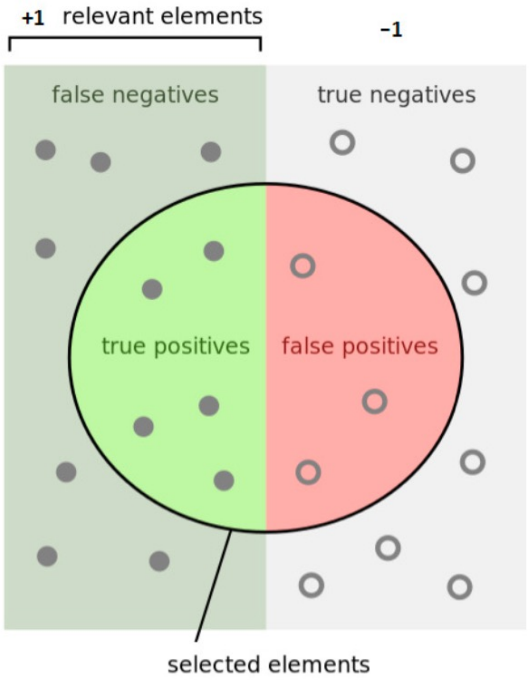
Точность и полнота бинарной классификации

В информационном поиске не важен TN:

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

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

Precision — доля релевантных среди найденных
Recall — доля найденных среди релевантных

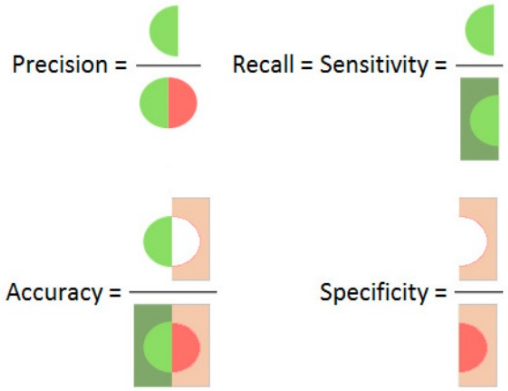


В медицинской диагностике:

$$\text{Чувствительность, Sensitivity} = \frac{TP}{TP+FN}$$

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

Sensitivity — доля верных положительных диагнозов
Specificity — доля верных отрицательных диагнозов



Точность и полнота многоклассовой классификации

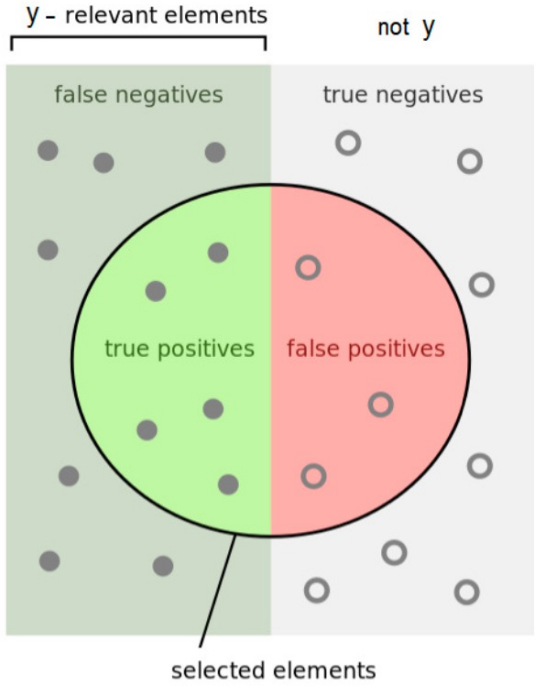
Для каждого класса $y \in Y$:
TP_y — верные положительные
FP_y — ложные положительные
FN_y — ложные отрицательные

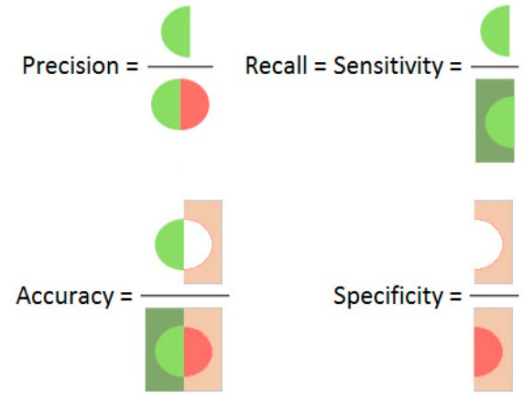
Точность и полнота с микроусреднением:

$$\text{Precision: } P = \frac{\sum_y \mathrm{TP}_y}{\sum_y (\mathrm{TP}_y + \mathrm{FP}_y)};$$

$$\text{Recall: } R = \frac{\sum_y \mathrm{TP}_y}{\sum_y (\mathrm{TP}_y + \mathrm{FN}_y)};$$

Микроусреднение не чувствительно
к ошибкам на малочисленных классах





Точность и полнота многоклассовой классификации

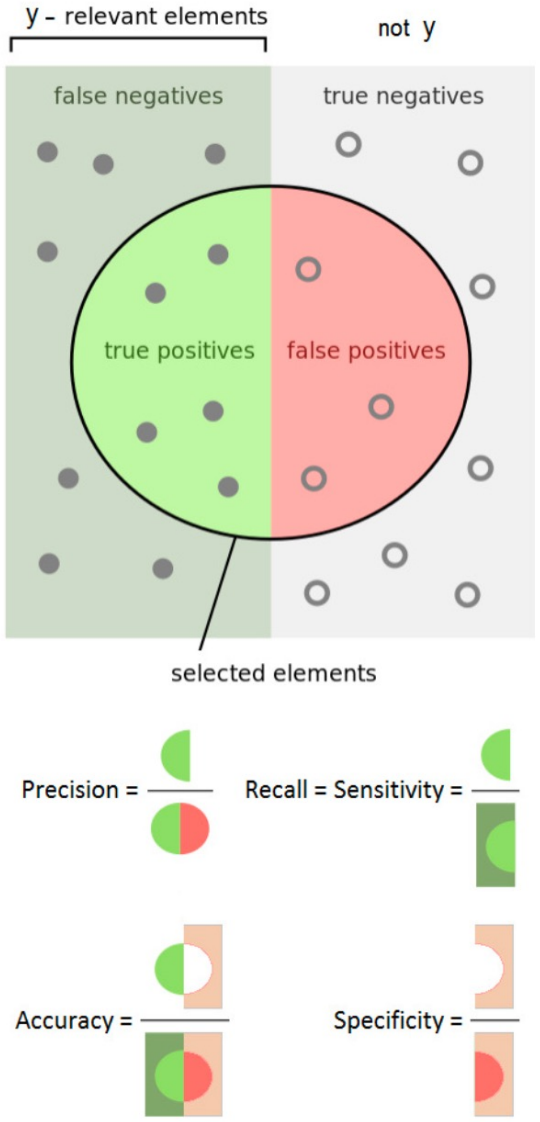
Для каждого класса $y \in Y$:
TP_y — верные положительные
FP_y — ложные положительные
FN_y — ложные отрицательные

Точность и полнота с макроусреднением:

$$\text{Precision: } P = \frac{1}{|Y|} \sum_y \frac{\mathrm{TP}_y}{\mathrm{TP}_y + \mathrm{FP}_y};$$

$$\text{Recall: } R = \frac{1}{|Y|} \sum_y \frac{\mathrm{TP}_y}{\mathrm{TP}_y + \mathrm{FN}_y};$$

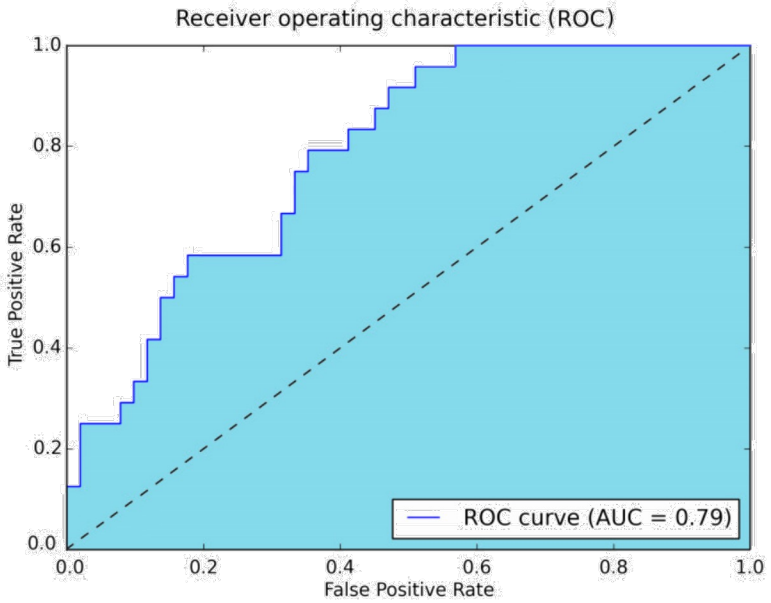
Макроусреднение чувствительно
к ошибкам на малочисленных классах

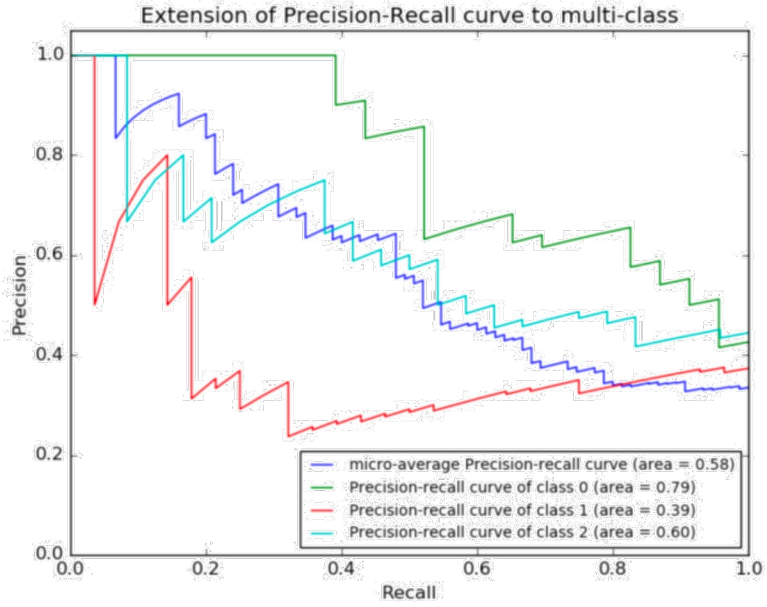


Кривые ROC и Precision-Recall

Модель классификации $g(x) = \text{sign}(\langle x, \boldsymbol{\theta} \rangle - \theta_0)$

Каждая точка кривой соответствует значению порога θ_0





AUROC — площадь под ROC-кривой

AUPRC — площадь под кривой Precision-Recall

Примеры из Python scikit learn: `http://scikit-learn.org/dev`

Резюме. Оценки качества классификации

- Чувствительность и специфичность лучше подходят для задач с несбалансированными классами
- Логарифм правдоподобия (log-loss) лучше подходит для оценки качества вероятностной модели классификации.
- Точность и полнота лучше подходят для задач поиска, когда доля объектов релевантного класса очень мала.

Агрегированные оценки:

- AUC лучше подходит для оценивания качества, когда соотношение цены ошибок не фиксировано.
- AUPRC — площадь под кривой точность–полнота.
- $F_1 = \frac{2PR}{P+R}$ — F-мера, другой способ агрегирования P и R.
- $F_\beta = \frac{(1+\beta^2)PR}{\beta^2 P+R}$ — F_β-мера: чем больше β, тем важнее R.

Оценка качества регрессии

Метрики оценки качества регрессии

$$MSE(y^{true}, y^{pred}) = \frac{1}{N}\sum_{i=1}^{N}(y_i - f(x_i))^2$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(y_i - f(x_i))^2}{\sum_{i=1}^{N}(y_i - \bar{y})^2}$$

$$MAE(y^{true}, y^{pred}) = \frac{1}{N}\sum_{i=1}^{N}|y_i - f(x_i)|$$

$$MAPE(y^{true}, y^{pred}) = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - f(x_i)|}{|y_i|}$$