Лекція 6

Метрики якості. Бібліотека scikit-learn

§31 Об'єднання точності й повноти

У деяких задачах є обмеження на одну із цих метрик, тоді як за другою метрикою буде проводитись оптимізація.

Але в деяких випадках потрібно максимізувати й точність, і повноту одночасно. Встає питання про об'єднання цих двох метрик.

Нагадую, що **precision (точність спацьовування класифікатора 1**)

$$precision(a, X) = \frac{TP}{TP + FP}.$$

recall (повнота визначення об’єктів класу 1)

$$recall(a, X) = \frac{TP}{TP + FN}.$$

	$y = 1$	$y = -1$
$a(x) = 1$	True Positive (TP)	False Positive (FP)
$a(x) = -1$	False Negative (FN)	True Negative (TN)

Арифметичне середнє

Єдина метрика може бути отримана як арифметичне середнє точності й повноти:

$$A = \frac{1}{2}(\mathbf{precision} + \mathbf{recall})$$

Нехай є алгоритм, точність якого дорівнює 10%, а повнота — 100%:

precision = 0.1   recall = 1.

Це може бути **випадок, коли у вибірці всього 10% об'єктів класу +1**, а алгоритм є константним і завжди повертає +1.

Очевидно, що цей **алгоритм некорисний**, але уведена вище метрика для нього дорівнює **A = 0.55**.

У свою чергу **інший**, набагато більше кращий алгоритм, з precision = 0.55 і recall = 0.55 також характеризується **A = 0.55**.

Ситуація, коли константний і корисний алгоритми можуть характеризуються однаково, є не припустимим, тому варто шукати інший спосіб побудови єдиної метрики.

Мінімум

Щоб константний і розумний алгоритми не характеризувались однаковим значенням, можна розглядати:

$$M = \min(precision, recall).$$

Даний підхід вирішує вищезгадану проблему, наприклад:

$$precision = 0.5, \quad recall = 1 \quad ==> \quad M = 0.5$$

Але є інший нюанс: два алгоритми, для яких точності однакові, але відрізняються значенням повноти, будуть мати однакое значення M :

$$precision = 0.4, \quad recall = 0.5 \quad \Longrightarrow \quad M = 0.4,$$
$$precision = 0.4, \quad recall = 0.9 \quad \Longrightarrow \quad M = 0.4.$$

Таке теж неприпустимо, тому що другий алгоритм істотно краще першого.

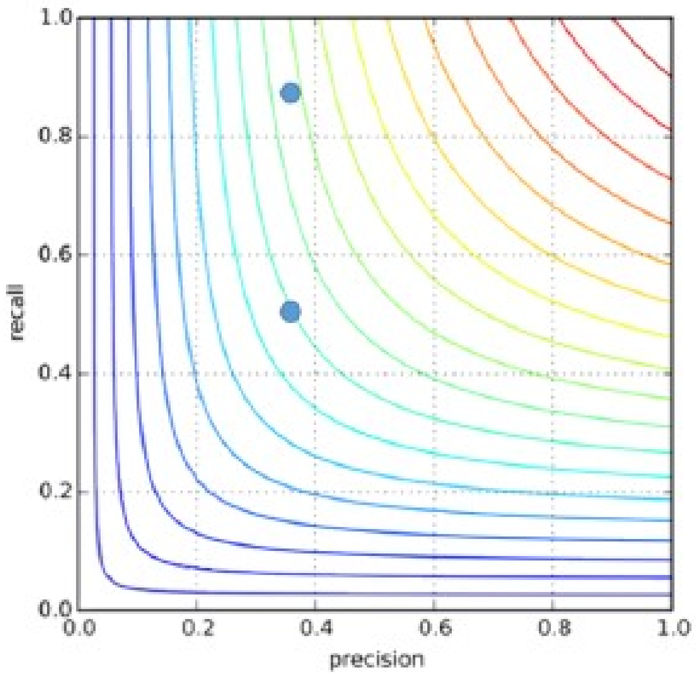
F-міра

Гармонійне середнє, або F-міри:

$$F = \frac{2 \cdot \text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

- $\text{precision} = 0.4$
- $\text{recall} = 0.5$
- $F = 0.44$

- $\text{precision} = 0.4$
- $\text{recall} = 0.9$
- $F = 0.55$



Якщо необхідно віддати перевагу точності або повноті, варто використовувати розширену F-міру, у якій є параметр β:

$$F = (1 + \beta^2)\frac{precision \cdot recall}{\beta^2 \cdot precision + recall}.$$

Наприклад, при β = 0.5 важливіше виявляється повнота, а у випадку β = 2, навпаки, важливіше виявляється точність.

§32 Якість оцінок належності до класу

Оцінка належності

Багато алгоритмів бінарної класифікації улаштовані в такий спосіб: спочатку обчислюється деяке дійсне число b(x), що порівнюється з порогом *t*.

$$a(x) = [b(x) > t],$$

де b(x) — оцінка належності класу +1. (оцінка впевненості, що x належить класу +1).

У випадку лінійного класифікатора $a(x) = [\langle w, x \rangle > t]$ оцінка належності класу +1 має вигляд $b(x) = \langle w, x \rangle$.

Часто буває необхідно **оцінити якість саме оцінки належності**, а поріг вибирається пізніше з міркувань на точність або повноту.

Оцінка належності в задачі кредитного скорінгу

Нехай розглядається задача кредитного скорінгу й була побудована деяка функція b(x), що оцінює ймовірність повернення кредиту клієнтом x. Далі класифікатор будується в такий спосіб:

$$a(x) = [b(x) > 0.5]$$

При цьому отримуємо, що точність (precision) дорівнює 10%, а повнота (recall) - 70%. Це дуже поганий алгоритм, тому що 90% клієнтів, яким буде виданий кредит, не повернуть його.

При цьому не зрозуміло, у чому справа: був погано обраний поріг або алгоритм не підходить для рішення цієї задачі.

Саме для цього необхідно вимірювати якість самих оцінок b(x).

PR-крива

Перший спосіб оцінки належності класу заснований на використанні кривої точності-повноти. Кожній точці на цій кривій буде відповідати класифікатор з деяким значенням порога.

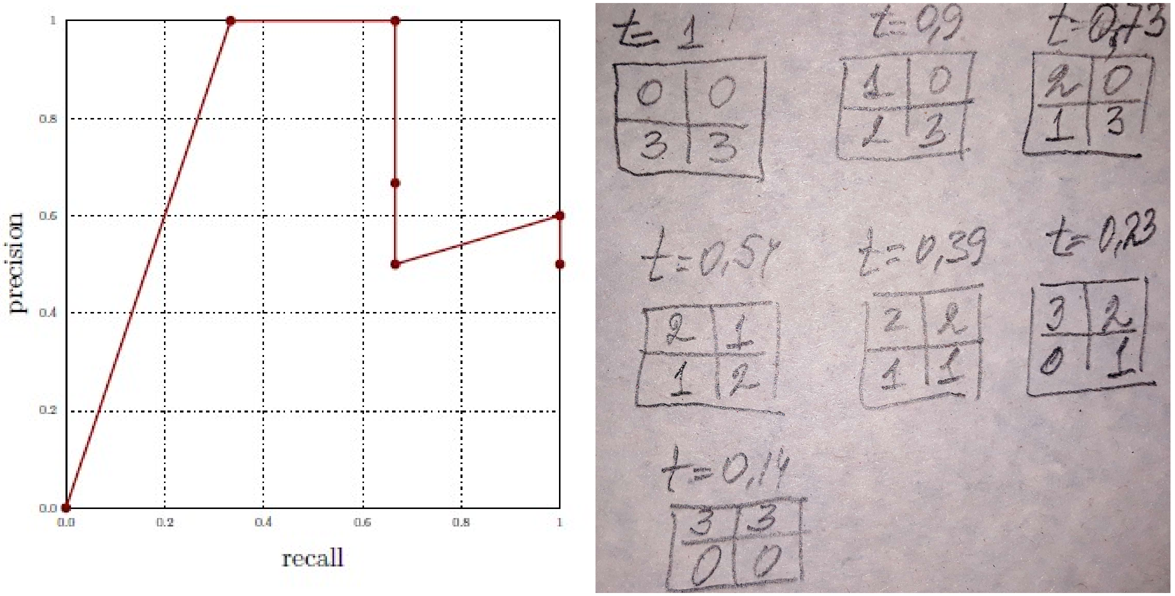


Рис. 4.5: Крива повноти-точності

Для прикладу наведена побудова PR-кривої для вибірки з 6 об'єктів, три з яких відносяться до класу y=1 і три — до класу y=0.

b(x)	0.14	0.23	0.39	0.54	0.73	0.90
y	0	1	0	0	1	1

- При досить великому порозі жоден об'єкт не буде віднесений до класу 1. У цьому випадку й точність і повнота рівні 0.
- При такому порозі, що рівно один об'єкт віднесений до класу 1, точність буде 100% (оскільки цей об'єкт дійсно з 1 класу), а повнота — 1/3 (оскільки всього 3 об'єкти 1 класу).
- При подальшому зменшенні порога вже два об'єкти віднесені до класу 1, точність також залишається 100%, а повнота стає 2/3.
- При такому порозі, що вже три об'єкти будуть віднесені до класу 1, точність стає рівної 2/3, а повнота залишається такий же.
- При такому порозі, що чотири об'єкти віднесені до класу 1, точність зменшиться до 0.5, а повнота знову не зміниться.
- При подальшому зменшенні порога вже 5 об'єктів будуть віднесені до 1 класу, повнота стане рівної 100%, а точність — 3/5.

У реальних задачах із числом об'єктів порядку декількох тисяч або десятків тисяч, крива точності- повноти виглядає приблизно так.

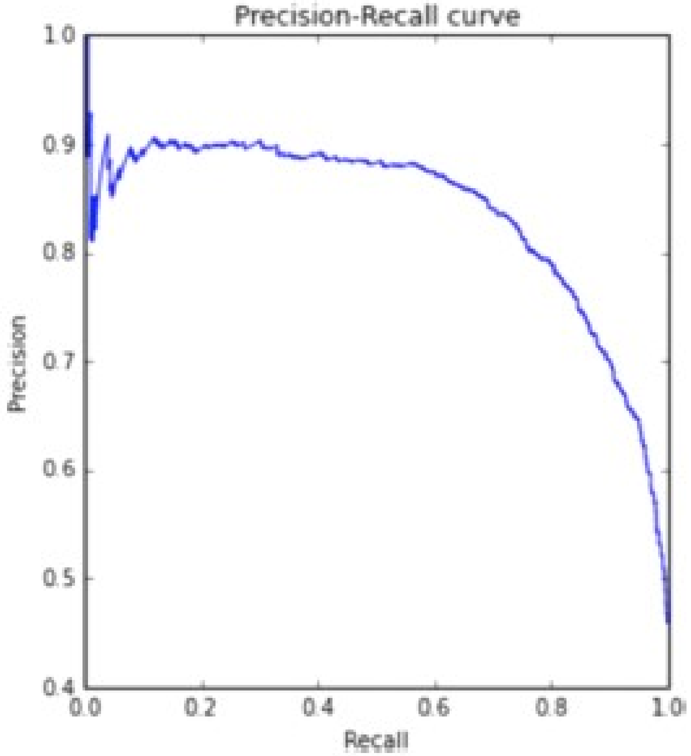


Рис. 4.6: Крива точності-повноти в реальних задачах з десятками тисяч об'єктів

Слід зазначити, що починається PR-крива завжди із точки (0, 0), а закінчується точкою (1, r), де r — частка об'єктів класу 1.

У випадку ідеального класифікатора, тобто якщо існує такий поріг, що й точність, і повнота дорівнюють 100%, крива буде проходити через точку (1, 1). Таким чином, чим ближче крива пройде до цієї точки, тим краще оцінки.

Площа під цією кривою може бути гарною мірою якості оцінок належності до класу 1. Така метрика **називається AUC-PRC**, або площа під PR-кривою.

ROC-крива

Інший спосіб виміряти якість оцінок належності до класу 1 — ROC-крива, що будується в осях False Positive Rate (X) і True Positive Rate (Y):

$$FPR = \frac{FP}{FP + TN}, \qquad TPR = \frac{TP}{TP + FN}.$$

ROC-крива будується аналогічно PR-кривій: поступово розглядаються випадки різних значень порогів і відзначаються точки на графіку. Для згаданої вище вибірки ROC-крива має такий вигляд:

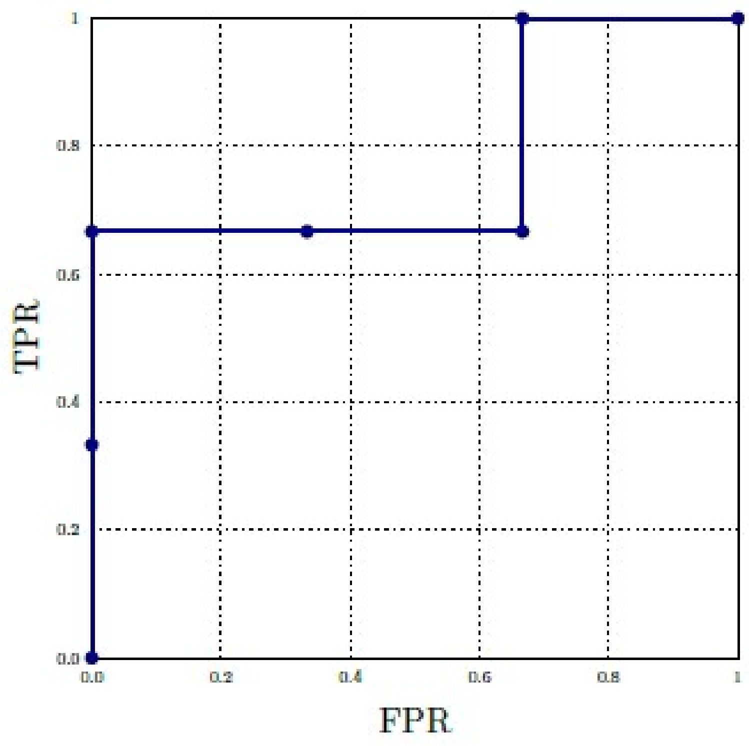


Рис. 4.7: ROC-крива

	$y = 1$	$y = -1$
$a(x) = 1$	True Positive (TP)	False Positive (FP)
$a(x) = -1$	False Negative (FN)	True Negative (TN)

t = 1

0	0
3	3

t = 0,9

1	0
2	3

t = 0,73

2	0
1	3

t = 0,57

2	1
1	2

t = 0,39

2	2
1	1

t = 0,23

3	2
0	1

t = 0,14

3	3
0	0

У випадку з великою вибіркою ROC-крива виглядає в так:

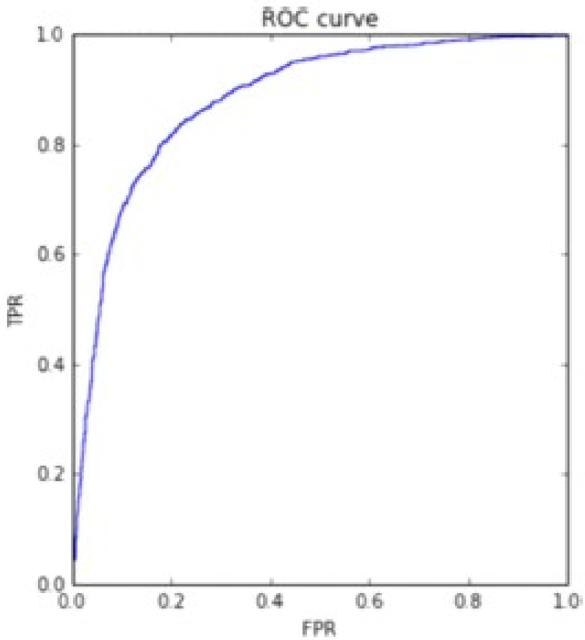


Рис. 4.8: Крива ROC у реальних задачах з десятками тисяч об'єктів

Крива стартує із точки (0, 0) і приходить у точку (1, 1). При цьому, якщо існує ідеальний класифікатор, крива повинна пройти через точку (0, 1). Чим ближче крива до цієї точки, тим краще будуть оцінки, а площа під кривою буде характеризувати якість оцінок належності до першого класу. Така метрика називається AUC-ROC, або площа під ROC-кривою.

Особливості AUC-ROC

Як було вказано вище, ROC-крива будується в осях FPR і TPR, які нормуються на розміри класів:

$$FPR = \frac{FP}{FP + TN}, \qquad TPR = \frac{TP}{TP + FN}.$$

Отже, при **зміні балансу класів** величина AUC-ROC і **незмінних властивостях об'єктів** вибірки **площа** під ROC-кривою **не зміниться**.

У випадку ідеального алгоритму AUC ROC = 1, а у випадку найгіршого AUC ROC = 1/2 .

Значення AUC ROC має **зміст імовірності** того, що якщо були обрані випадковий позитивний і випадковий негативний об'єкти вибірки, **позитивний об'єкт одержить оцінку належності до класу 1 з імовірністю AUC_ROC**

AUC=0,7 можна інтерпретувати як те, що моделі вдасться вірно розпізнати клас об’єкту імовірністю 0,7.

Особливості AUC-PRC

PR-Крива будується в осях precision і recall:

$$precision = \frac{TP}{TP+FP}, \qquad recall = \frac{TP}{TP+FN},$$

а отже змінюється при зміні балансу класів.

§33 Вбудовані датасети. Sklearn.datasets

Див. JNotebook «2_1Вбудовані_датасети_Sklearn_datasets.ipynb»