

Лекция 5. Оценка качества моделей

Основы интеллектуального анализа данных

Полузёров Т. Д.

БГУ ФПМИ

- 1 Обобщающая способность
 - Функция потерь
 - Понятие метрики
- 2 Метрики классификации
 - Метки классов
 - Порядок объектов
- 3 Метрики регрессии
 - Абсолютные
 - Относительные
- 4 Подбор гиперпараметров
 - Подбор гиперпараметров

Дано

Имеется выборка объектов $X = (x_1, \dots, x_\ell)$, $X \in \mathbb{X}$ и соответствующие им метки $Y = (y_1, \dots, y_\ell)$, $Y \in \mathbb{Y}$.

Построена модель $a : \mathbb{X} \rightarrow \mathbb{Y}$ по данным $D^\ell = (X, Y)$ с помощью метода обучения $\mu : \mathbb{X} \times \mathbb{Y} \rightarrow \mathbb{A}$ – минимизации эмпирического риска

$$Q(a, D^\ell) = \frac{1}{\ell} \sum_{i=1}^{\ell} \mathcal{L}(y_i, a(x_i, \theta)) \rightarrow \min_{\theta}$$

где $\mathcal{L} : \mathbb{Y} \times \mathbb{Y} \rightarrow \mathbb{R}$ - функция ошибки модели на объекте

Насколько адекватно работает модель на других объектах из \mathbb{X} ?

Обобщающая способность

Обобщающая способность алгоритма a

$$G(a) = E_{\mathbb{X}}\{\mathcal{L}(y, a(x))\}$$

На практике используется эмпирическая оценка

$$\hat{G}(a) = Q(a, \hat{D}) = \frac{1}{|\hat{D}|} \sum_{i=1}^{|\hat{D}|} \mathcal{L}(y_i, a(x_i))$$

показывает среднюю ошибку модели a , обученной по выборке D , на других данных $\hat{D} \in (\mathbb{X} \times \mathbb{Y})$.

При $|\hat{D}| \rightarrow \infty$, согласно закону больших чисел, $\hat{G} \rightarrow G$.

Метод отложенной выборки

Случайное разбиение выборки $D = D_{train} \sqcup D_{test}$.

Обучение на D_{train} , а оценка на D_{test} .

$$\hat{G}(a) = Q(\mu(D_{train}), D_{test})$$

- Обычно пропорции train - 80%, test- 20%
- Для задачи классификации случайное разбиение сохраняющее пропорции классов – **стратифицированное**
- Вычислительно эффективно – всего 1 обучение
- Может быть нестабильной оценкой

Функция потерь \neq метрика качества

Имея алгоритм a хочется оценить его реальный бизнес эффект. Метрика M отражает целевой эффект от модели a . Может порождаться иерархия метрик.

Пример иерархии метрик для задачи рекомендаций фильмов:

- 1 Доход сервиса – глобальная цель
- 2 Среднее число просматриваемых фильмов в неделю – косвенный показатель релевантности рекомендаций
- 3 Точность классификации <кликнет - не кликнет> – можно рассчитать по имеющимся данным
- 4 Функция потерь (\mathcal{L}) – по ней и строим модель a

Функция потерь как аппроксимация метрики

Метрика M - внешний, объективный показатель качества алгоритма.

Функция потерь \mathcal{L} - математически удобная функция, которая служит лишь для построения алгоритма.

- Глобальная задача стоит в оптимизации метрики, а функция потерь выступает как прокси.
- В некоторых задачах они могут совпадать - тогда метрика оптимизируется напрямую.
- Прямые верхнеуровневые метрики бывает невозможно измерить в моменте, поэтому используются косвенные

Online и Offline метрики

Online - метрики, которые рассчитываются по уже работающей системе.

Offline - рассчитываются до введения в эксплуатацию, могут быть рассчитаны по имеющимся данным.

На этапе построения моделей и их сравнения доступны только оффлайн метрики. $(X, Y) \in (\mathbb{X}, \mathbb{Y})$, $\mathcal{M} : \mathbb{Y} \times \mathbb{Y} \rightarrow \mathbb{R}$

Доля ошибочных классификаций

Доля ошибочных классификаций (Accuracy):

$$Accuracy(y, \hat{y}) = \frac{1}{\ell} \sum_i^N [y_i = \hat{y}_i]$$

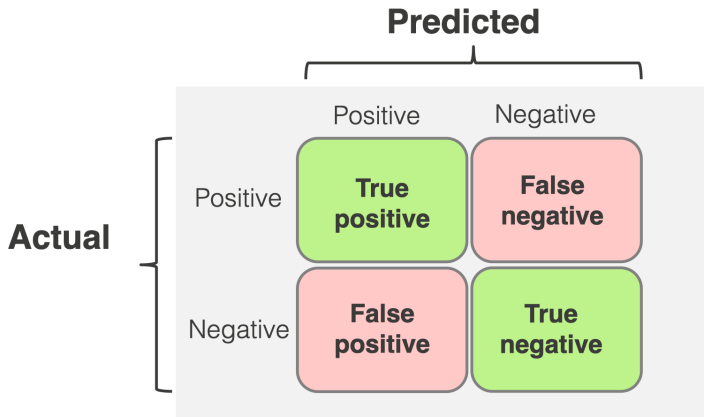
Или сопряженная с ней **доля ошибочных классификаций** (Error Rate):

$$ErrorRate = 1 - Accuracy$$

- неприменима при сильном дисбалансе классов
- не учитывает величину ошибки на объектах разных классов

Матрица ошибок

Матрица ошибок (Confusion Matrix) – кросс-таблица из истинных меток и прогнозов.



Элементы матрицы ошибок

True/False - допустил ли классификатор ошибку

Positive/Negative - ответ классификатора

- TP - верно определенный положительный класс
- TN - верный негативный класс
- FP - ложное срабатывание (ошибочная положительная классификация)
- FN - пропуск объекта(ошибочное не срабатывание)

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

Метрики по матрице ошибок

Ошибка 1-го рода(False Positive Rate, FPR).

Вероятность ложного срабатывания:

$$P(\hat{y} = +1|y = -1) = \frac{FP}{FP + TN}$$

Ошибка 2-го рода(False Negative Rate, FNR).

Вероятность пропуска:

$$P(\hat{y} = -1|y = +1) = \frac{FN}{FN + TP}$$

Точность и полнота

Точность – доля верных классификаций при срабатывании

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

Полнота – доля всех положительных объектов на которых сработал классификатор

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

Усреднение точности и полноты

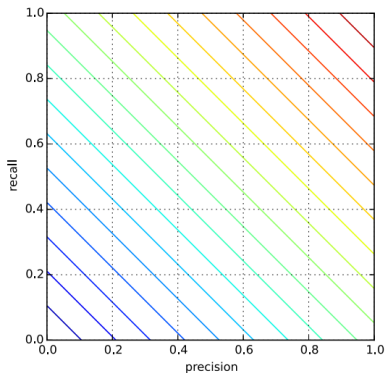
Одновременно учитывать Precision и Recall можно с помощью среднего гармонического:

$$F_1 = \frac{2}{\frac{1}{Precision} + \frac{1}{Recall}} = 2 \frac{Precision \cdot Recall}{Precision + Recall}$$

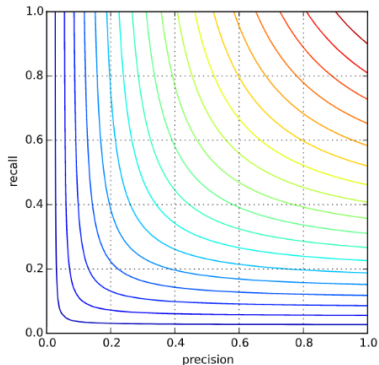
Важность одной из метрик можно учесть с помощью коэффициента β

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

Линии уровня



Среднее арифметическое



Среднее гармоническое

Порог классификации

Классификаторы вида $a(x, t) = \text{sign}(f(x) - t)$ допускают настройку порога отсечения t .

При $t \rightarrow -\infty$:

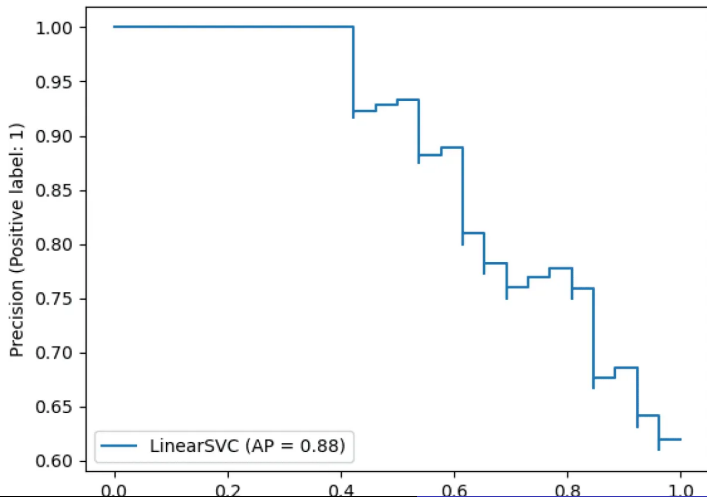
- 1 все объекты классифицируются в **положительный** класс
- 2 $Recall = 1$
- 3 $Precision$ = доле положительных объектов в выборке

При $t \rightarrow +\infty$:

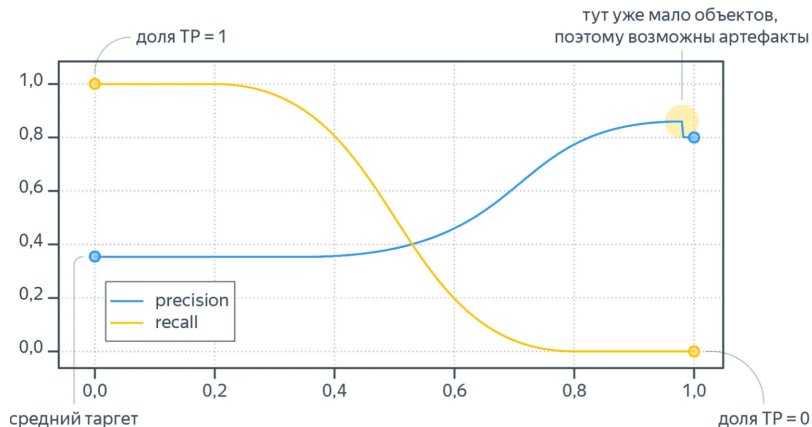
- все объекты классифицируются в **отрицательный** класс
- $Recall = 0$
- $Precision$ - не определен

Precision-Recall кривая

2-class Precision-Recall curve: AP=0.88



Зависимость от порога



Оптимизация порога

Перебирая порог t можно добиться **любого значения** $Recall \in [0, 1]$. При этом будет достигаться некоторый $Precision$.

Оптимизация $Precision - Recall$ соотношения происходит так:

- 1 Строится алгоритм базовый алгоритм $b : \mathbb{X} \rightarrow \mathbb{R}$
- 2 по отдельной валидационной выборке строятся кривые порога t
- 3 определяется оптимальное соотношение $Precision$ и $Recall$ при t^*
- 4 фиксируется итоговый алгоритм $a(x) = \text{sign}(b(x) - t^*)$

Средняя точность

Существует ряд метрик отражающих общее качество модели, инвариантно относительно порога t .

$$AveragePrecision = \int_0^1 p(r)dr$$

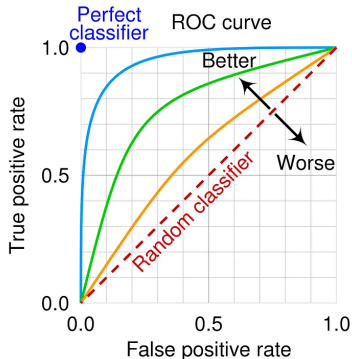
где $p(r)$ - значение *Precision* при $Recall = r$.

Эта метрика соответствует площади под P-R кривой.

ROC-кривая

TPR (True Positive Rate) - полнота (Recall)

FPR (False Positive Rate) - доля отрицательных объектов на которых ложно сработал классификатор



$$TPR = \frac{TP}{TP + NP}$$

$$FPR = \frac{FP}{FP + TN}$$

Алгоритм построения ROC

- ❶ сортируем объекты по уверенности классификатора
- ❷ стартуем из точки $(0, 0)$ и перебираем объекты выборки
- ❸ делаем шаг в
 - вверх, если объект правильно классифицирован
 - право, если допущена ошибка

Классика

Стандартный набор метрик в задачах регрессии

- 1 Средний квадрат ошибки

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

- 2 Иногда MSE приводят к той же размерности что и ответы

$$RMSE = \sqrt{MSE}$$

- 3 Средняя абсолютная ошибка

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

Относительные ошибки

Mean Absolute Percentage Error

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

Symmetric MAPE

$$SMAPE = \frac{1}{\ell} \sum_{i=1}^{\ell} \frac{2|y_i - \hat{y}_i|}{y_i + \hat{y}_i}$$

Weighted Average PE

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

Подбор гиперпараметров

Параметры в модели бывают двух типов:

- 1 Параметры - те, что настраиваются в ходе решения
 $Q \rightarrow \min$
- 2 Гиперпараметры - фиксируются до обучения

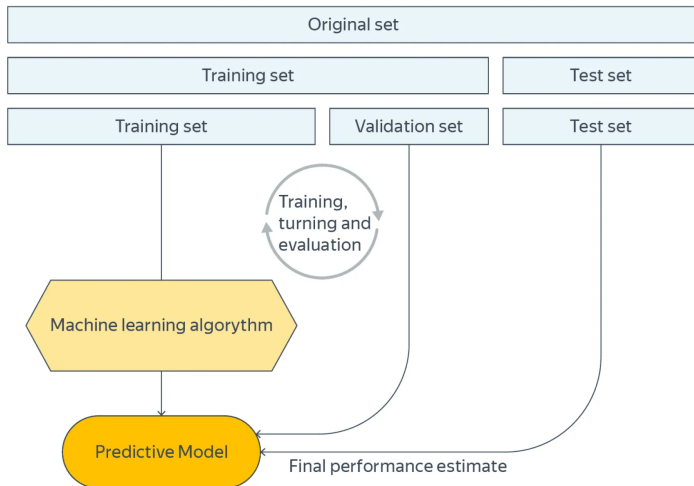
Пример:

$$a_d(x, \omega) = \omega_0 + \omega_1 x + \dots + \omega_d x^d$$

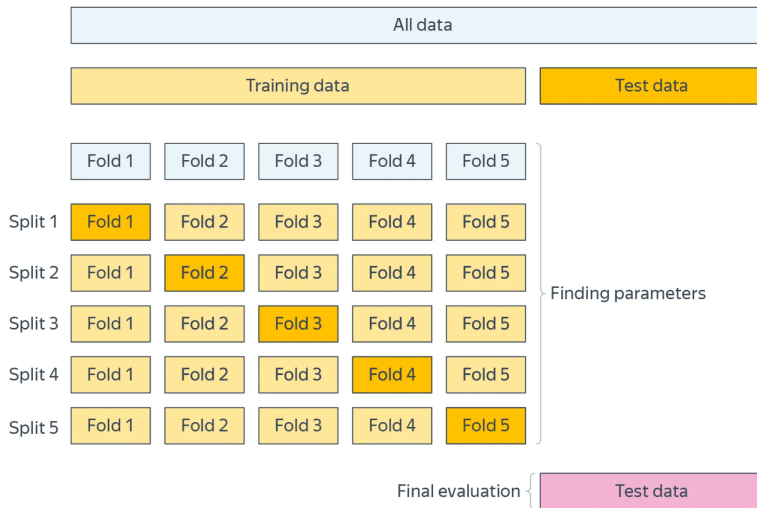
ω - параметры, а d - гиперпараметр

Как определить оптимальные значения гиперпараметров?

Отложенные выборки



Кросс-валидация



Заключение

- Метрики служат для сравнения обученных моделей
- Метрики должны отражать бизнес требования
- Функция потерь – математическое приближение бизнес метрики
- Обучение модели, подбор гиперпараметров и оценка качества - на разных наборах данных