Методы машинного обучения. Критерии качества и выбор моделей

Воронцов Константин Вячеславович
www.MachineLearning.ru/wiki?title=User:Vokov
вопросы к лектору: vokov@forecsys.ru

материалы курса:
github.com/MSU-ML-COURSE/ML-COURSE-21-22
орг.вопросы по курсу: ml.cmc@mail.ru

ВМК МГУ • 12 октября 2021

Содержание

Анализ ошибок классификации

Задача классификации на два класса, $y_i \in \{-1, +1\}$.
Алгоритм классификации $a(x_i) \in \{-1, +1\}$

	ответ классификатора	правильный ответ
TP, True Positive	$a(x_i) = +1$	$y_i = +1$
TN, True Negative	$a(x_i) = -1$	$y_i = -1$
FP, False Positive	$a(x_i) = +1$	$y_i = -1$
FN, False Negative	$a(x_i) = -1$	$y_i = +1$

Доля правильных классификаций (чем больше, тем лучше):

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

Недостаток: не учитывается ни численность (дисбаланс) классов, ни цена ошибки на объектах разных классов.

Функции потерь, зависящие от штрафов за ошибку

Задача классификации на два класса, $y_i \in \{-1, +1\}$.
Модель классификации: $a(x; w, w_0) = \text{sign}\bigl(g(x, w) - w_0\bigr)$.
Чем больше w_0, тем больше x_i таких, что $a(x_i) = -1$.

Пусть λ_y — штраф за ошибку на объекте класса y.
Функция потерь теперь зависит от штрафов:

$$\mathscr{L}(a, y) = \lambda_{y_i}\bigl[a(x_i; w, w_0) \neq y_i\bigr] = \lambda_{y_i}\bigl[(g(x_i, w) - w_0)y_i < 0\bigr].$$

Проблема

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

- Нужен удобный способ выбора w_0 в зависимости от $\{\lambda_y\}$, не требующий построения w заново.
- Нужна характеристика качества модели $g(x, w)$, не зависящая от штрафов $\{\lambda_y\}$ и численности классов.

Определение ROC-кривой

Кривая ошибок ROC (receiver operating characteristic).
Каждая точка кривой соответствует некоторому $a(x; w, w_0)$.

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

$$\text{FPR} = \frac{\sum_{i=1}^{\ell} [y_i = -1][a(x_i; w, w_0) = +1]}{\sum_{i=1}^{\ell} [y_i = -1]};$$

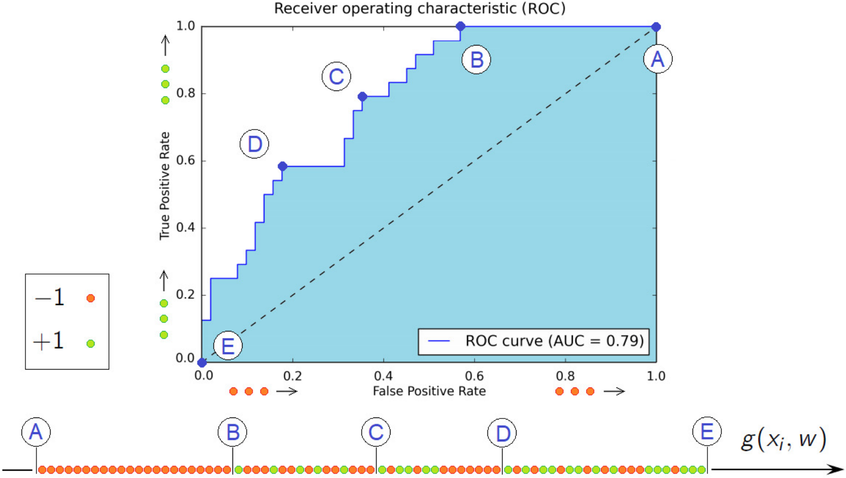
  $1 - \text{FPR}$ называется *специфичностью* алгоритма a.

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

$$\text{TPR} = \frac{\sum_{i=1}^{\ell} [y_i = +1][a(x_i; w, w_0) = +1]}{\sum_{i=1}^{\ell} [y_i = +1]};$$

  TPR называется также *чувствительностью* алгоритма a.

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



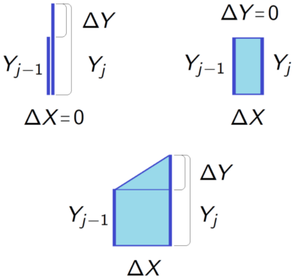
ABCDE — положения порога w_0 на оси значений функции g

Алгоритм эффективного построения ROC-кривой

Вход: выборка $\{x_i\}_{i=1}^{\ell}$; дискриминантная функция $g(x, w)$;
Выход: ROC-кривая $(X_j, Y_j)_{j=0}^{k}$, $k \leqslant \ell$ и площадь AUC

$\ell_y := \sum_{i=1}^{\ell}[y_i = y]$, для всех $y \in Y$;
упорядочить $\{x_i\}$ по убыванию $g_i = g(x_i, w)$: $g_1 \geqslant \ldots \geqslant g_\ell$;
$(X_0, Y_0) := (0, 0)$; $\mathrm{AUC} := 0$; $\Delta X := 0$; $\Delta Y := 0$; $j := 1$;

для $i := 1, \ldots, \ell$
- $\Delta X := \Delta X + \frac{1}{\ell_-}[y_i = -1]$;
- $\Delta Y := \Delta Y + \frac{1}{\ell_+}[y_i = +1]$;
- **если** $(g_i \neq g_{i-1})$ **то**
  - $X_j := X_{j-1} + \Delta X$;
  - $Y_j := Y_{j-1} + \Delta Y$;
  - $\mathrm{AUC} := \mathrm{AUC} + \frac{1}{2}\bigl(Y_{j-1} + Y_j\bigr)\Delta X$;
  - $j := j + 1$; $\Delta X := 0$; $\Delta Y := 0$;



Градиентная максимизация AUC

Модель классификации: $a(x_i, w, w_0) = \operatorname{sign}\bigl(g(x_i, w) - w_0\bigr)$.

AUC — это доля правильно упорядоченных пар (x_i, x_j):

$$\text{AUC}(w) = \frac{1}{\ell_-}\sum_{i=1}^{\ell}\bigl[y_i = -1\bigr]\text{TPR}_i =$$
$$= \frac{1}{\ell_-\ell_+}\sum_{i=1}^{\ell}\sum_{j=1}^{\ell}\bigl[y_i < y_j\bigr]\bigl[g(x_i, w) < g(x_j, w)\bigr] \to \max_w.$$

Явная максимизация аппроксимированного AUC:

$$1 - \text{AUC}(w) \leqslant Q(w) = \sum_{i,j\colon y_i<y_j} \mathscr{L}\bigl(\underbrace{g(x_j, w) - g(x_i, w)}_{M_{ij}(w)}\bigr) \to \min_w,$$

где $\mathscr{L}(M)$ — убывающая функция отступа,

$M_{ij}(w)$ — новое понятие отступа для пар объектов.

Алгоритм SG для максимизации AUC

Возьмём для простоты линейный классификатор:

$$g(x, w) = \langle x, w \rangle, \qquad M_{ij}(w) = \langle x_j - x_i, w \rangle, \qquad y_i < y_j.$$

Вход: выборка X^ℓ, темп обучения h, темп забывания λ;
Выход: вектор весов w;

инициализировать веса w_j, $j = 0, \dots, n$;
инициализировать оценку: $\bar{Q} := \frac{1}{\ell_+ \ell_-} \sum\limits_{i,j} [y_i < y_j] \mathscr{L}(M_{ij}(w))$;

повторять

выбрать пару объектов (i, j): $y_i < y_j$, случайным образом;
вычислить потерю: $\varepsilon_{ij} := \mathscr{L}(M_{ij}(w))$;
сделать градиентный шаг: $w := w - h \mathscr{L}'(M_{ij}(w))(x_j - x_i)$;
оценить функционал: $\bar{Q} := (1 - \lambda)\bar{Q} + \lambda \varepsilon_{ij}$;

пока значение $\bar{Q}$ и/или веса w не сойдутся;

Логарифм правдоподобия, log-loss

Вероятностная модель классификации, $y_i \in \{-1, +1\}$:

$$g(x, w) = \mathsf{P}(y = +1 | x, w).$$

Проблема: ROC и AUC инвариантны относительно монотонных преобразований дискриминантной функции $g(x, w)$.

Критерий логарифма правдоподобия (log-loss):

$$L(w) = \sum_{i=1}^{\ell} [y_i = +1] \ln g(x, w) + [y_i = -1] \ln\bigl(1 - g(x, w)\bigr) \to \max_w$$

Вероятностная модель многоклассовой классификации:

$$a(x) = \arg\max_{y \in Y} \mathsf{P}(y | x, w);$$

$$L(w) = \sum_{i=1}^{\ell} \ln \mathsf{P}(y_i | x_i, w) \to \max_w$$

Оценки качества двухклассовой классификации

В информационном поиске:

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

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

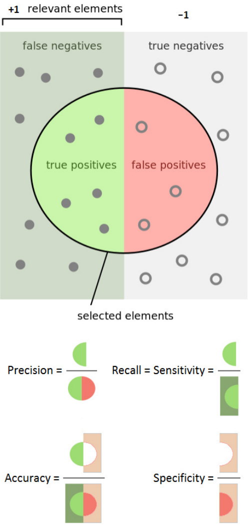
Precision — доля релевантных среди найденных
Recall — доля найденных среди релевантных

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

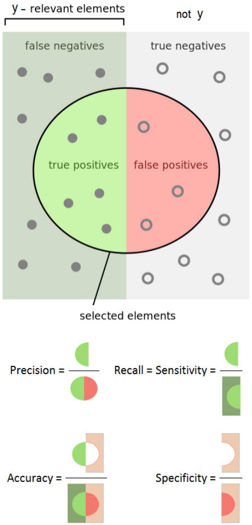
Для каждого класса $y \in Y$:
TP_y — верные положительные
FP_y — ложные положительные
FN_y — ложные отрицательные

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

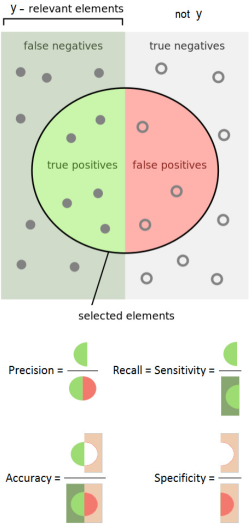
Для каждого класса $y \in Y$:
TP_y — верные положительные
FP_y — ложные положительные
FN_y — ложные отрицательные

Точность и полнота с макроусреднением:

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

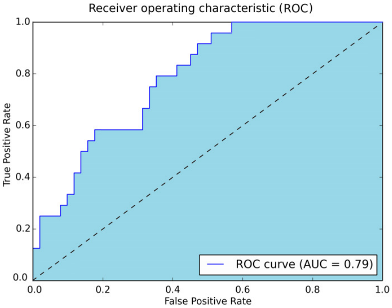
$$\text{Recall: } R = \frac{1}{|Y|} \sum_y \frac{\mathrm{TP}_y}{\mathrm{TP}_y + \mathrm{FN}_y};$$

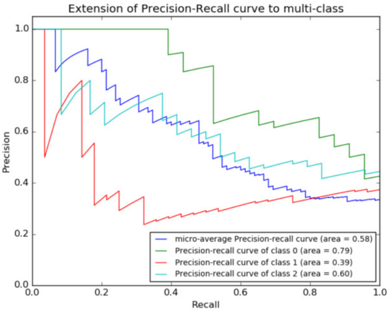
Макроусреднение чувствительно
к ошибкам на малочисленных классах



Кривые ROC и Precision-Recall

Модель классификации: $a(x) = \mathrm{sign}\bigl(\langle x, w\rangle - w_0\bigr)$
Каждая точка кривой соответствует значению порога w_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.

Задачи выбора модели и метода обучения

Дано: X — пространство объектов; Y — множество ответов;
$X^\ell = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка, $y_i = y^*(x_i)$;
$A_t = \{a\colon X \to Y\}$ — модели алгоритмов, $t \in T$;
$\mu_t\colon (X \times Y)^\ell \to A_t$ — методы обучения, $t \in T$.

Найти: метод μ_t с наилучшей *обобщающей способностью*.

Частные случаи:

- выбор лучшей модели A_t (model selection);
- выбор метода обучения μ_t для заданной модели A (в частности, оптимизация *гиперпараметров*);
- отбор признаков (feature selection):
  $F = \bigl\{f_j\colon X \to D_j\colon j = 1, \dots, n\bigr\}$ — множество признаков;
  метод обучения μ_J использует только признаки $J \subseteq F$.

Как оценить качество обучения по прецедентам?

$\mathscr{L}(a, x)$ — функция потерь алгоритма a на объекте x;

$Q(a, X^\ell) = \frac{1}{\ell} \sum\limits_{i=1}^{\ell} \mathscr{L}(a, x_i)$ — *функционал качества* a на X^ℓ.

Внутренний критерий оценивает качество на обучении X^ℓ:

$$Q_\mu(X^\ell) = Q\bigl(\mu(X^\ell), X^\ell\bigr).$$

Недостаток: эта оценка смещена, т.к. μ минимизирует её же.

Внешний критерий оценивает качество «вне обучения», например, по отложенной (hold-out) контрольной выборке X^k:

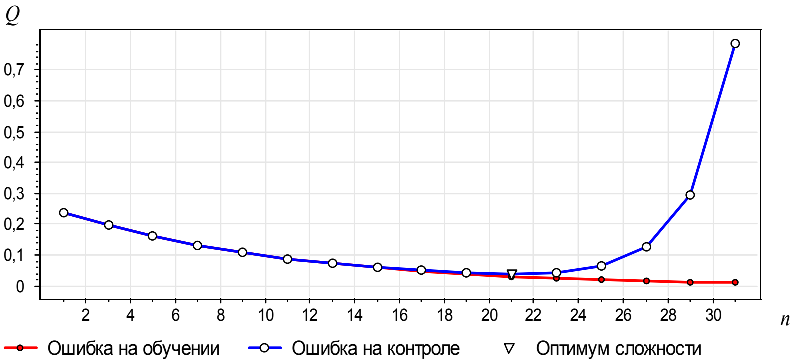
$$Q_\mu(X^\ell, X^k) = Q\bigl(\mu(X^\ell), X^k\bigr).$$

Недостаток: эта оценка зависит от разбиения $X^L = X^\ell \sqcup X^k$.

Основное отличие внешних критериев от внутренних

Внутренний критерий монотонно убывает с ростом сложности модели (например, числа признаков).

Внешний критерий имеет характерный минимум, соответствующий оптимальной сложности модели.



Кросс-проверка (cross-validation, CV)

Усреднение оценок hold-out по заданному N — множеству разбиений $X^L = X^\ell_n \sqcup X^k_n$, $n = 1, \dots, N$:

$$\mathrm{CV}(\mu, X^L) = \frac{1}{|N|} \sum_{n \in N} Q_\mu(X^\ell_n, X^k_n).$$

Частные случаи — разные способы задания N.

1. Случайное множество разбиений.

2. *Полная кросс-проверка* (complete cross-validation, CCV): N — множество всех $C^k_{\ell+k}$ разбиений.

Недостаток: оценка CCV вычислительно слишком сложна. Используются либо малые k, либо комбинаторные оценки CCV.

Скользящий контроль и поблочная кросс-проверка

3. Скользящий контроль (leave one out CV): $k = 1$,

$$\mathrm{LOO}(\mu, X^L) = \frac{1}{L}\sum_{i=1}^{L} Q_\mu(X^L \backslash \{x_i\}, \{x_i\}).$$

Недостатки LOO: ресурсоёмкость, высокая дисперсия.

4. Кросс-проверка по q блокам (q-fold CV): случайное разбиение $X^L = X_1^{\ell_1} \sqcup \dots \sqcup X_q^{\ell_q}$ на q блоков (почти) равной длины,

$$\mathrm{CV}_q(\mu, X^L) = \frac{1}{q}\sum_{n=1}^{q} Q_\mu\bigl(X^L \backslash X_n^{\ell_n}, X_n^{\ell_n}\bigr).$$

Недостатки q-fold CV:
— оценка существенно зависит от разбиения на блоки;
— каждый объект лишь один раз участвует в контроле.

Многократная поблочная кросс-проверка

5. Контроль t раз по q блокам ($t{\times}q$-fold CV)
— стандарт «де факто» для тестирования методов обучения.

Выборка X^L разбивается t раз случайным образом на q блоков
$X^L = X^{\ell_1}_{s1} \sqcup \dots \sqcup X^{\ell_q}_{sq}, \quad s = 1, \dots, t, \quad \ell_1 + \dots + \ell_q = L;$

$$\mathrm{CV}_{t\times q}(\mu, X^L) = \frac{1}{t}\sum_{s=1}^{t}\frac{1}{q}\sum_{n=1}^{q} Q_\mu\bigl(X^L \setminus X^{\ell_n}_{sn}, X^{\ell_n}_{sn}\bigr).$$

Преимущества $t{\times}q$-fold CV:
- — увеличением t можно улучшать точность оценки (компромисс между точностью и временем вычислений);
- — каждый объект участвует в контроле ровно t раз;
- — оценивание доверительных интервалов (95% при $t = 40$).

Критерии непротиворечивости моделей

Идея: Если модель верна, то алгоритмы, настроенные по разным частям данных, не должны противоречить друг другу.

1. По одному случайному разбиению $X^\ell \sqcup X^k = X^L$, $\ell = k$:

$$D_1(\mu, X^L) = \frac{1}{L}\sum_{i=1}^{L} \bigl|\mu(X^\ell)(x_i) - \mu(X^k)(x_i)\bigr|.$$

2. Аналог $\mathrm{CV}_{t\times 2}$: по t разбиениям $X^L = X^\ell_s \sqcup X^k_s$, $s = 1, \ldots, t$:

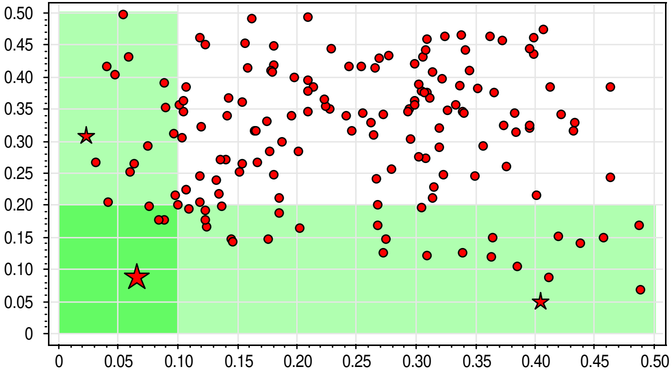
$$D_t(\mu, X^L) = \frac{1}{t}\sum_{s=1}^{t} \frac{1}{L}\sum_{i=1}^{L} \bigl|\mu(X^\ell_s)(x_i) - \mu(X^k_s)(x_i)\bigr|.$$

Недостатки:

— длина обучения сокращается в 2 раза;
— трудоёмкость возрастает в $2t$ раз.

Многокритериальный выбор модели

Модель, немного неоптимальная по обоим критериям, может оказаться лучше, чем модель, оптимальная по одному критерию, но сильно не оптимальная по другому.



Критерии регуляризации

Регуляризатор — аддитивная добавка к внутреннему критерию, обычно штраф за сложность (complexity penalty) модели A:

$$Q_{\text{рег}}(\mu, X^\ell) = Q_\mu(X^\ell) + \text{штраф}(A),$$

Линейные модели: $A = \bigl\{a(x) = \operatorname{sign}\langle w, x\rangle\bigr\}$ — классификация,
$A = \bigl\{a(x) = \langle w, x\rangle\bigr\}$ — регрессия.

L_2-регуляризация (ридж-регрессия):

$$\text{штраф}(w) = \tau\|w\|_2^2 = \tau\sum_{j=1}^n w_j^2.$$

L_1-регуляризация (LASSO):

$$\text{штраф}(w) = \tau\|w\|_1 = \tau\sum_{j=1}^n |w_j|.$$

L_0-регуляризация (AIC, BIC):

$$\text{штраф}(w) = \tau\|w\|_0 = \tau\sum_{j=1}^n \bigl[w_j \neq 0\bigr].$$

Разновидности L_0-регуляризации

Информационный критерий Акаике (Akaike Information Criterion):

$$\mathrm{AIC}(\mu, x) = Q_\mu(X^\ell) + \frac{2\hat{\sigma}^2}{\ell}|J|,$$

где $\hat{\sigma}^2$ — оценка дисперсии ошибки $\mathsf{D}\big(y_i - a(x_i)\big)$,
J — подмножество используемых признаков.

Байесовский информационный критерий (Bayes Inform. Criterion):

$$\mathrm{BIC}(\mu, X^\ell) = \frac{\ell}{\hat{\sigma}^2}\left(Q_\mu(X^\ell) + \frac{\hat{\sigma}^2 \ln \ell}{\ell}|J|\right).$$

Оценка Вапника-Червоненкиса (VC-bound):

$$\mathrm{VC}(\mu, X^\ell) = Q_\mu(X^\ell) + \sqrt{\frac{h}{\ell}\ln\frac{2e\ell}{h} + \frac{1}{\ell}\ln\frac{9}{4\eta}},$$

h — VC-размерность; для линейных, опять-таки, $h = |J|$;
η — уровень значимости; обычно $\eta = 0.05$.

Связь регуляризации с оценками обобщающей способности

Идея обращения верхних оценок вероятности переобучения.

1. Получить верхнюю оценку *вероятности переобучения*, справедливую для любой выборки X^L, широкого класса моделей A и методов обучения μ:

$$\mathsf{P}\Big[Q_\mu(X^\ell, X^k) - Q_\mu(X^\ell) \geqslant \varepsilon\Big] \leqslant \eta(\varepsilon, A).$$

2. Тогда для любой X^L, любых A и μ и любого $\eta \in (0,1)$ с вероятностью не менее $(1-\eta)$ справедлива оценка

$$Q_\mu(X^\ell, X^k) \leqslant Q_\mu(X^\ell) + \varepsilon(\eta, A),$$

где $\varepsilon(\eta, A)$ — функция штрафа на A, обратная к $\eta(\varepsilon, A)$, не зависящая от скрытой контрольной выборки X^k.

3. Оптимизировать метод обучения: $Q_\mu(X^\ell) + \varepsilon(\eta, A) \to \min\limits_\mu$.

Бинарная функция потерь. Матрица ошибок

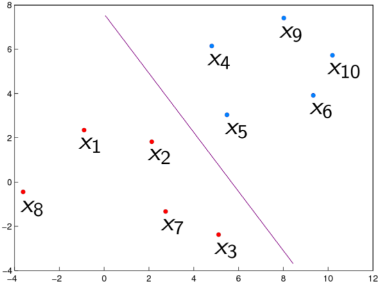
$X^L = \{x_1, \ldots, x_L\}$ — конечное *генеральное множество* объектов;
$A = \{a_1, \ldots, a_D\}$ — конечное семейство *алгоритмов*;
$\mathscr{L}(a,x) \equiv I(a,x) = [$алгоритм a ошибается на объекте $x]$;

$L{\times}D$-*матрица ошибок* с попарно различными столбцами:

	a_1	a_2	a_3	a_4	a_5	a_6	$\cdots$	a_D	
x_1	1	1	0	0	0	1	$\cdots$	1	X^ℓ — наблюдаемая
$\ldots$	0	0	0	0	1	1	$\cdots$	1	(обучающая) выборка
x_ℓ	0	0	1	0	0	0	$\cdots$	0	длины ℓ
$x_{\ell+1}$	0	0	0	1	1	1	$\cdots$	0	X^k — скрытая
$\ldots$	0	0	0	1	0	0	$\cdots$	1	(контрольная) выборка
x_L	0	1	1	1	1	1	$\cdots$	0	длины $k = L - \ell$

$n(a,X) = \sum\limits_{x \in X} I(a,x)$ — *число ошибок* $a \in A$ на выборке $X \subset X^L$;
$\nu(a,X) = n(a,X)/|X|$ — *частота ошибок* a на выборке X;

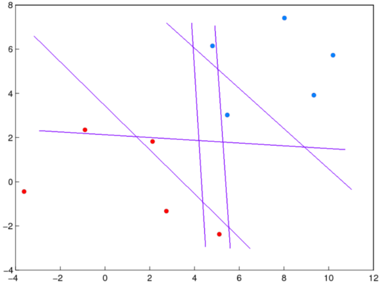
Пример. Матрица ошибок линейных классификаторов



1 вектор с 0 ошибками

x_1	0
x_2	0
x_3	0
x_4	0
x_5	0
x_6	0
x_7	0
x_8	0
x_9	0
x_{10}	0

Пример. Матрица ошибок линейных классификаторов

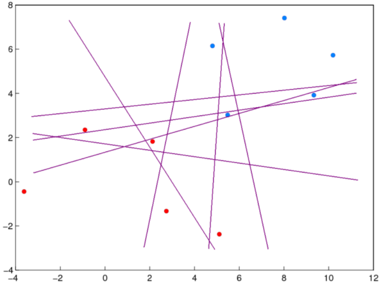


1 вектор с 0 ошибками

5 векторов с 1 ошибкой

x_1	0	1	0	0	0	0
x_2	0	0	1	0	0	0
x_3	0	0	0	1	0	0
x_4	0	0	0	0	1	0
x_5	0	0	0	0	0	1
x_6	0	0	0	0	0	0
x_7	0	0	0	0	0	0
x_8	0	0	0	0	0	0
x_9	0	0	0	0	0	0
x_{10}	0	0	0	0	0	0

Пример. Матрица ошибок линейных классификаторов



1 вектор с 0 ошибками
5 векторов с 1 ошибкой
8 векторов с 2 ошибками
и т. д...

x_1	0	1	0	0	0	0	1	0	0	0	0	1	1	0	...
x_2	0	0	1	0	0	0	1	1	0	0	0	0	0	0	...
x_3	0	0	0	1	0	0	0	1	1	0	0	0	0	1	...
x_4	0	0	0	0	1	0	0	0	1	1	0	0	0	0	...
x_5	0	0	0	0	0	1	0	0	0	1	1	1	0	0	...
x_6	0	0	0	0	0	0	0	0	0	0	1	0	1	0	...
x_7	0	0	0	0	0	0	0	0	0	0	0	0	0	1	...
x_8	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
x_9	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...
x_{10}	0	0	0	0	0	0	0	0	0	0	0	0	0	0	...

Задача оценивания вероятности переобучения

Основное вероятностное предположение:
все разбиения $X^\ell \sqcup X^k = X^L$ равновероятны
(слабый вариант гипотезы независимости выборки X^L).

Переобученность — разность частот ошибок на X^k и на X^ℓ:

$$\delta(\mu, X^\ell) = \nu\bigl(\mu(X^\ell), X^k\bigr) - \nu\bigl(\mu(X^\ell), X^\ell\bigr).$$

Переобучение — это событие $\delta(\mu, X^\ell) \geqslant \varepsilon$.

Основная задача — оценить *вероятность переобучения*:

$$R_\varepsilon(\mu, X^L) = \mathsf{P}\bigl[\delta(\mu, X^\ell) \geqslant \varepsilon\bigr].$$

Простейший, но важный частный случай

Пусть $A = \{a\}$ — одноэлементное множество, $m = n(a, X^L)$.

Тогда вероятность переобучения есть вероятность большого отклонения частот ошибок в двух подвыборках:

$$R_\varepsilon(a, X^L) = \mathsf{P}\bigl[\nu\bigl(a, X^k\bigr) - \nu\bigl(a, X^\ell\bigr) \geqslant \varepsilon\bigr].$$

Теорема

Для любого X^L, любого $\varepsilon \in [0, 1]$

$$R_\varepsilon(a, X^L) = \mathcal{H}_L^{\ell, m}\left(\tfrac{\ell}{L}(m - \varepsilon k)\right),$$

где $\mathcal{H}_L^{\ell, m}(z) = \sum\limits_{s=0}^{\lfloor z \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}$ — функция гипергеометрического распределения.

Доказательство

1. Обозначим $s = n(a, X^\ell)$.

2. «Школьная» задача по теории вероятностей:
в урне L шаров, m из них чёрные; извлекаем ℓ шаров наугад.
Какова вероятность того, что s из них чёрные?

$$\mathsf{P}\bigl[n(a, X^\ell) = s\bigr] = C_m^s C_{L-m}^{\ell-s} / C_L^\ell.$$

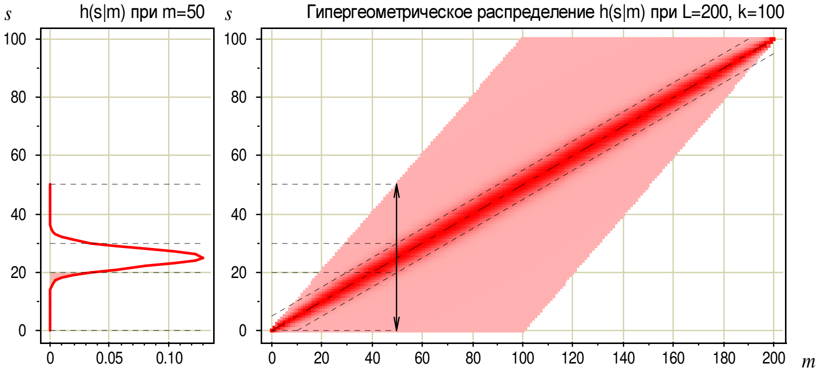
3. Распишем R_ε, подставив $\nu(a, X^k) = \frac{m-s}{k}$, $\nu(a, X^\ell) = \frac{s}{\ell}$:

$$R_\varepsilon(a, X^L) = \mathsf{P}\bigl[\nu(a, X^k) - \nu(a, X^\ell) \geqslant \varepsilon\bigr] =$$

$$= \sum_{s=0}^{\ell} \underbrace{\Bigl[\frac{m-s}{k} - \frac{s}{\ell} \geqslant \varepsilon\Bigr]}_{s \leqslant \frac{\ell}{L}(m-\varepsilon k)} \underbrace{\mathsf{P}\bigl[n(a, X^\ell) = s\bigr]}_{C_m^s C_{L-m}^{\ell-s} / C_L^\ell} =$$

$$= \mathcal{H}_L^{\ell, m}\left(\tfrac{\ell}{L}(m - \varepsilon k)\right). \qquad \blacksquare$$

Гипергеометрическое распределение $h(s|m) = C_m^s C_{L-m}^{\ell-s} / C_L^\ell$



Предсказание числа $m = n(a, X^L)$ по числу $s = n(a, X^\ell)$ возможно благодаря узости гипергеометрического пика, причём при $\ell, k \to \infty$ он сужается, и $\nu(a, X^\ell) \to \nu(a, X^k)$ (явление *концентрации вероятности*, закон больших чисел).

Принцип равномерной сходимости частот

Рассмотрим случай, когда A произвольное, конечное.

1. Вероятность переобучения оценим сверху вероятностью большого *равномерного отклонения* частот: для любых X^L, μ

$$
\begin{aligned}
R_\varepsilon(\mu, X^L) &= \mathsf{P}\bigl[\delta(\mu, X^\ell) \geqslant \varepsilon\bigr] \leqslant \\
&\leqslant \mathsf{P}\Bigl[\max_{a\in A} \delta(a, X^\ell) \geqslant \varepsilon\Bigr] = \widetilde{R}_\varepsilon(A, X^L).
\end{aligned}
$$

2. Оценим вероятность объединения событий суммой их вероятностей (неравенство Буля, union bound):

$$
\begin{aligned}
\widetilde{R}_\varepsilon(A, X^L) &= \mathsf{P}\max_{a\in A}\bigl[\delta(a, X^\ell) \geqslant \varepsilon\bigr] \leqslant \\
&\leqslant \mathsf{P}\sum_{a\in A}\bigl[\delta(a, X^\ell) \geqslant \varepsilon\bigr] = \sum_{a\in A} \underbrace{\mathsf{P}\bigl[\delta(a, X^\ell) \geqslant \varepsilon\bigr]}_{R_\varepsilon(a, X^L)}.
\end{aligned}
$$

Основная теорема Вапника–Червоненкиса

Таким образом, мы доказали важную теорему:

Теорема

Для любых X^L, μ, конечного A и $\varepsilon \in [0,1]$

$$\widetilde{R}_\varepsilon(A, X^L) \leqslant \sum_{a \in A} \mathcal{H}_L^{\ell,m}\left(\tfrac{\ell}{L}(m - \varepsilon k)\right),$$

где $m = n(a, X^L)$.

Следствие (Вапник и Червоненкис, 1968)

Для любых X^L, μ, конечного A и $\varepsilon \in [0,1]$

$$\widetilde{R}_\varepsilon(A, X^L) \leqslant |A| \cdot \max_m \mathcal{H}_L^{\ell,m}\left(\tfrac{\ell}{L}(m - \varepsilon k)\right) \leqslant$$

$$\leqslant |A| \cdot \tfrac{3}{2} \exp\left(-\varepsilon^2 \ell\right), \quad \text{при } \ell = k.$$

Обобщение на случай бесконечных семейств A

Функция роста $\Delta^A(L)$ семейства A — это максимальное по X^L число различных векторов ошибок $\vec{a} = \bigl(I(a,x_1),\dots,I(a,x_L)\bigr)$.
В оценке надо заменить $|A|$ на функцию роста $\Delta^A(L)$.

Ёмкость (размерность Вапника-Червоненкиса) семейства A — это максимальная длина выборки h, для которой $\Delta^A(h) = 2^h$.

Теорема

Если такое h существует, то $\Delta^A(L) \leqslant C_L^0 + \dots + C_L^h \leqslant \frac{3}{2}\frac{L^h}{h!}$.

Теорема

Ёмкость семейства линейных классификаторов на два класса

$$a(x) = \operatorname{sign}(w_1x^1 + \dots + w_nx^n), \quad x = (x^1,\dots,x^n) \in X.$$

равна размерности пространства параметров, $\mathrm{VCdim}(A) = n$.

Обращение оценки Вапника-Червоненкиса (при $\ell = k$)

1. Оценка: $\mathsf{P}\Bigl[\max\limits_{a\in A}\bigl(\nu(a,X^k) - \nu(a,X^\ell)\bigr) \geqslant \varepsilon\Bigr] \leqslant \Delta\,\frac{3}{2}\exp\bigl(-\ell\varepsilon^2\bigr).$

Тогда для любого $a \in A$ с вероятностью не менее $(1-\eta)$

$$\nu(a,X^k) \leqslant \underbrace{\nu(a,X^\ell)}_{\text{эмпирический риск}} + \underbrace{\sqrt{\frac{1}{\ell}\ln\Delta + \frac{1}{\ell}\ln\frac{3}{2\eta}}}_{\text{штраф за сложность}}.$$

2. Оценка: $\mathsf{P}\Bigl[\max\limits_{a\in A}\bigl(\nu(a,X^k) - \nu(a,X^\ell)\bigr) \geqslant \varepsilon\Bigr] \leqslant \frac{3}{2}\frac{L^h}{h!}\cdot\frac{3}{2}\exp\bigl(-\ell\varepsilon^2\bigr).$

Тогда для любого $a \in A$ с вероятностью не менее $(1-\eta)$

$$\nu(a,X^k) \leqslant \underbrace{\nu(a,X^\ell)}_{\text{эмпирический риск}} + \underbrace{\sqrt{\frac{h}{\ell}\ln\frac{2e\ell}{h} + \frac{1}{\ell}\ln\frac{9}{4\eta}}}_{\text{штраф за сложность}}.$$

Метод структурной минимизации риска (СМР)

Дано: система вложенных подсемейств возрастающей ёмкости

$$A_0 \subset A_1 \subset \cdots \subset A_h \subset \cdots$$

Найти: оптимальную ёмкость h^*, такую, что

$$\nu(a, X^k) \leqslant \underbrace{\min_{a \in A_h} \nu(a, X^\ell)}_{\text{минимизация эмпирического риска}} + \underbrace{\sqrt{\frac{h}{\ell} \ln \frac{2e\ell}{h} + \frac{1}{\ell} \ln \frac{9}{4\eta}}}_{\text{штраф за сложность}} \to \min_h$$

Недостатки СМР:

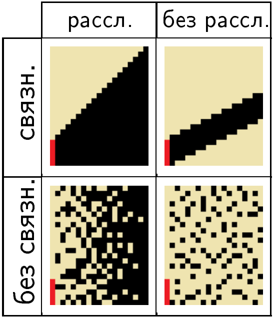
— верхняя оценка R_ε очень сильно завышена
— h^* может оказаться заниженной из-за завышенности R_ε
— на практике эмпирический CV предпочтительнее этих оценок

Зависит ли переобучение от содержимого матрицы ошибок?

Согласно VC-теории, только размер матрицы ошибок $L \times |A|$ влияет на переобучение. Но так ли это?

Эксперимент: сравним R_ε у четырёх матриц ошибок:

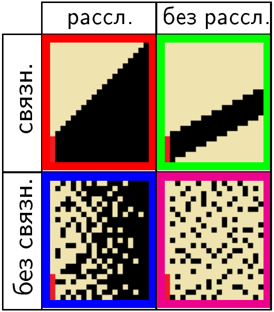
- лучший алгоритм одинаковый
- есть/нет *расслоение* — когда каждый следующий алгоритм допускает на одну ошибку больше, чем предыдущий
- есть/нет *связность* — когда каждый следующий алгоритм лишь на одном объекте отличается от предыдущего

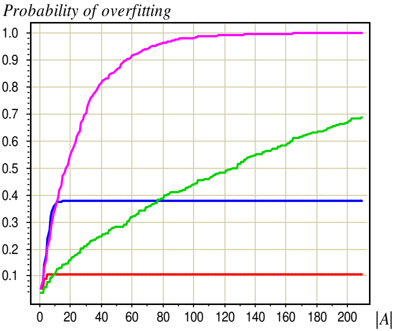


Vorontsov K. V. Splitting and similarity phenomena in the sets of classifiers and their effect on the probability of overfitting. PRIA, 2009.

Эксперимент с разрушением монотонной цепи

$\ell = k = 100, \ \ \varepsilon = 0.05, \ \ N = 1000$ разбиений Монте-Карло.

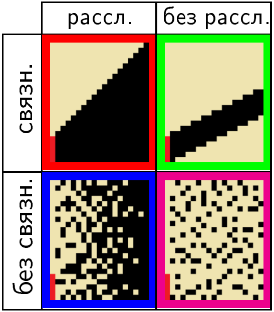


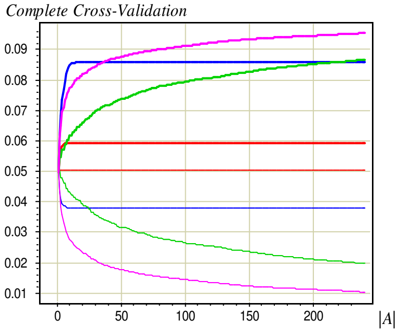


- *связность* замедляет темп роста кривой $R_\varepsilon(|A|)$
- *расслоение* понижает уровень горизонтальной асимптоты
- огромные семейства с P&C могут почти не переобучаться
- VC-оценка линейно мажорирует худшую из этих кривых

Эксперимент с разрушением монотонной цепи

$\ell = k = 100$, $N = 1000$ разбиений Монте-Карло.

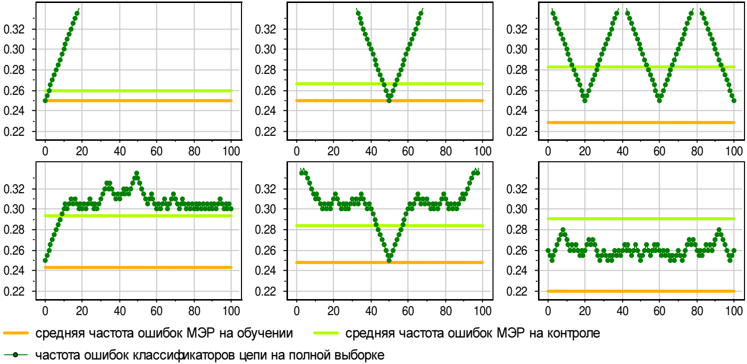




- справа: оценки CCV $= \hat{\mathsf{E}}\nu\bigl(\mu(X^\ell), X^k\bigr)$ и $\hat{\mathsf{E}}\nu\bigl(\mu(X^\ell), X^\ell\bigr)$
- без P&C даже 10 алгоритмов могут сильно переобучаться
- без учёта эффектов расслоения и связности получение точных оценок вероятности переобучения невозможно

Эксперимент. Переобучение цепей с различным расслоением

Условия эксперимента: $L = 100$, $\ell = 50$, $m = 25$, $\varepsilon = 0.05$, метод Монте-Карло по $N = 100000$ случайных разбиений.



- при оптимизации одного скалярного параметра переобучение незначительно только если есть расслоение

Резюме. Оценки обобщающей способности и выбор модели

- Для выбора модели используются внешние критерии:
  — *эмпирические* — на основе разбиений train $\sqcup$ test;
  — *аналитические* — через регуляризацию на основе теоретических верхних оценок вероятности переобучения.
- Классические VC-оценки приводят к L_0-регуляризации.
- Завышенность VC-оценок может приводить в методе СМР к занижению сложности (переупрощению) моделей.
- Два эффекта совместно уменьшают переобучение:
  — *расслоение* — метод обучения с высокой вероятностью выбирает алгоритмы $a = \mu(X^\ell)$ из нижних слоёв A;
  — *связность* — метод обучения часто выбирает схожие алгоритмы благодаря непрерывности модели по параметрам.
- При отсутствии этих эффектов вероятность переобучения может приближаться к 1 уже при $|A|$ порядка десятка.
- Жёсткие верхние оценки (tight bounds) переобучения выводятся в теории COLT (Computational Learning Theory).