Комбинаторные оценки переобучения пороговых классификаторов

Ишкина Шаура Хабировна

Вычислительный центр им. А. А. Дородницына РАН

ММРО-17

19 – 25 сентября 2015, г. Светлогорск, Калининградская область

План

1. **Комбинаторная теория переобучения**
   - Задача оценивания вероятности переобучения
2. **Переобучение пороговых классификаторов**
   - Случай непрерывного и дискретного признаков
   - Алгоритм вычисления вероятности переобучения
3. **Эксперименты**
   - Проверка алгоритма
   - Задача медицинской диагностики
   - Выводы и открытые проблемы

Основные понятия

$\mathbb{X} = \{x_1, \ldots, x_L\}$ — конечное *генеральное множество* объектов;
$A = \{a_1, \ldots, a_D\}$ — конечное множество *классификаторов*;
$I(a, x) = [$классификатор 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 — 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	$\bar{X}$ — скрытая
$\ldots$	0	0	0	1	0	0	$\cdots$	1	(контрольная) выборка
x_L	0	1	1	1	1	1	$\cdots$	0	длины $k = L - \ell$

Вероятностная аксиома:

Пусть все разбиения $X \sqcup \bar{X} = \mathbb{X}$ равновероятны.

Задача оценивания вероятности переобучения

$n(a, X) = \sum_{x \in X} I(a, x)$ — *число ошибок* $a \in A$ на выборке $X \subset \mathbb{X}$;

$\nu(a, X) = \frac{1}{|X|} n(a, X)$ — *частота ошибок* a на выборке X;

Опр. *Метод обучения* $\mu \colon 2^{\mathbb{X}} \to A$ произвольной выборке $X \subset \mathbb{X}$ ставит в соответствие некоторый классификатор $a \in A$.

Опр. Метод *минимизации эмпирического риска*:

$$\mu X \in A(X) = \operatorname{Arg} \min_{a \in A} n(a, X).$$

$\delta(\mu, X) = \nu(\mu X, \bar{X}) - \nu(\mu X, X)$ — *переобученность* μ на X.

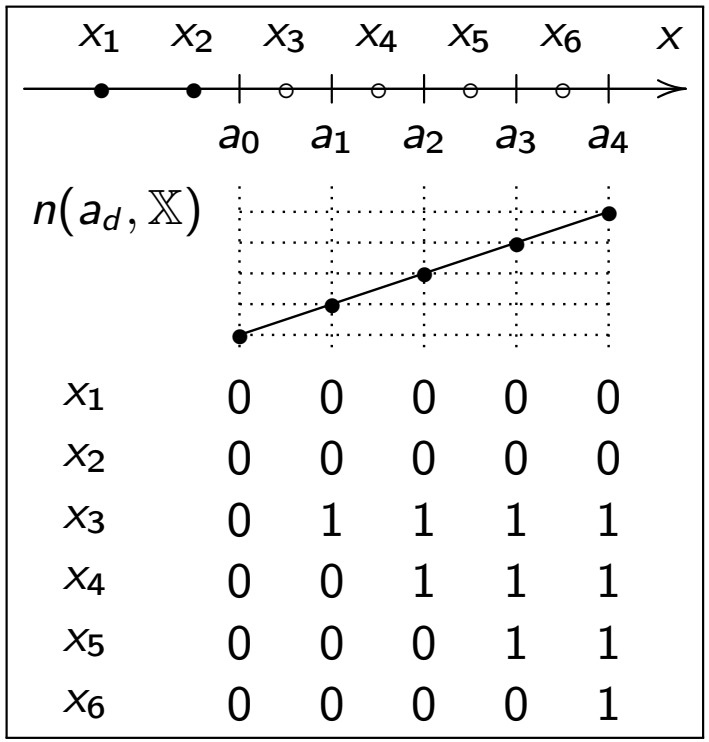
Вероятность переобучения:

$$Q_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\big[\delta(\mu, X) \geqslant \varepsilon\big] = \frac{1}{C_L^\ell} \sum_{X \subset \mathbb{X}} \big[\delta(\mu, X) \geqslant \varepsilon\big].$$

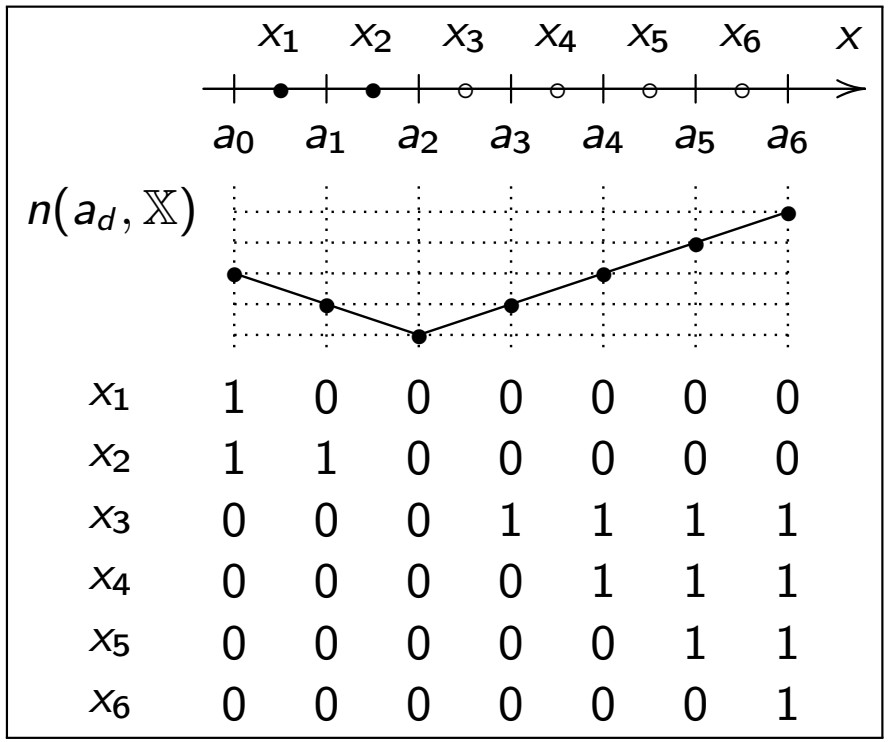
Пороговые классификаторы. Пример

Пусть $x = (x^1, \ldots, x^n) \in \mathbb{X}$, $\mathbb{Y} = \{0, 1\}$. Пусть $x^1 \in \mathbb{R}$.
Рассмотрим семейство

$$A = \{a_\theta(x) = [x^1 \geqslant \theta], \theta \in \mathbb{R}\}.$$



x_1	0	0	0	0	0
x_2	0	0	0	0	0
x_3	0	1	1	1	1
x_4	0	0	1	1	1
x_5	0	0	0	1	1
x_6	0	0	0	0	1



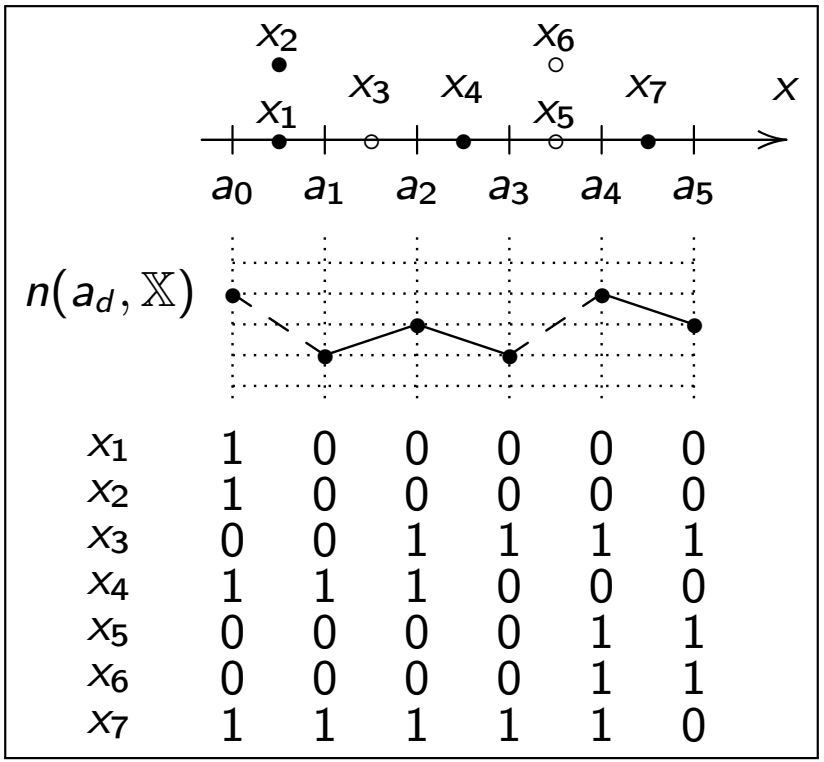
x_1	1	0	0	0	0	0	0
x_2	1	1	0	0	0	0	0
x_3	0	0	0	1	1	1	1
x_4	0	0	0	0	1	1	1
x_5	0	0	0	0	0	1	1
x_6	0	0	0	0	0	0	1

Прямая цепь

Прямая цепь $A = \{a_0, \ldots, a_D\}$ – семейство классификаторов:

$$\mathbb{G}_d = \{x \in \mathbb{X} \mid I(a_{d-1}, x) \neq I(a_d, x)\}, \; d = 1, \ldots, D.$$

Обозначим $\mathbb{G} = \mathbb{G}_1 \sqcup \ldots \sqcup \mathbb{G}_D = \{x \in \mathbb{X} \mid I(a_0, x) \neq I(a_D, x)\}$

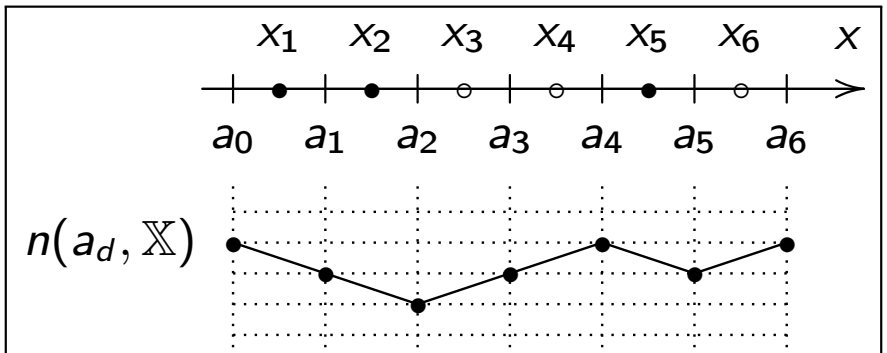


x_1	1	0	0	0	0	0
x_2	1	0	0	0	0	0
x_3	0	0	1	1	1	1
x_4	1	1	1	0	0	0
x_5	0	0	0	0	1	1
x_6	0	0	0	0	1	1
x_7	1	1	1	1	1	0

Непрерывная прямая цепь

Непрерывная прямая цепь $A = \{a_0, \ldots, a_D\}$ – прямая цепь,

$$\forall d \in \{1, \ldots, D\} \quad |\mathbb{G}_d| = 1.$$



x_1	1	0	0	0	0	0	0
x_2	1	1	0	0	0	0	0
x_3	0	0	0	1	1	1	1
x_4	0	0	0	0	1	1	1
x_5	1	1	1	1	1	0	0
x_6	0	0	0	0	0	0	1

Алгоритм вычисления вероятности переобучения

Результат: алгоритм вычисления вероятности переобучения со сложностью

$$O\left(G^6 + G\ell^3 + L\right) = O(L^6),$$

где $G = |\mathbb{G}| = \rho(a_0, a_D)$.
Для непрерывной цепи параметр G равен длине D.

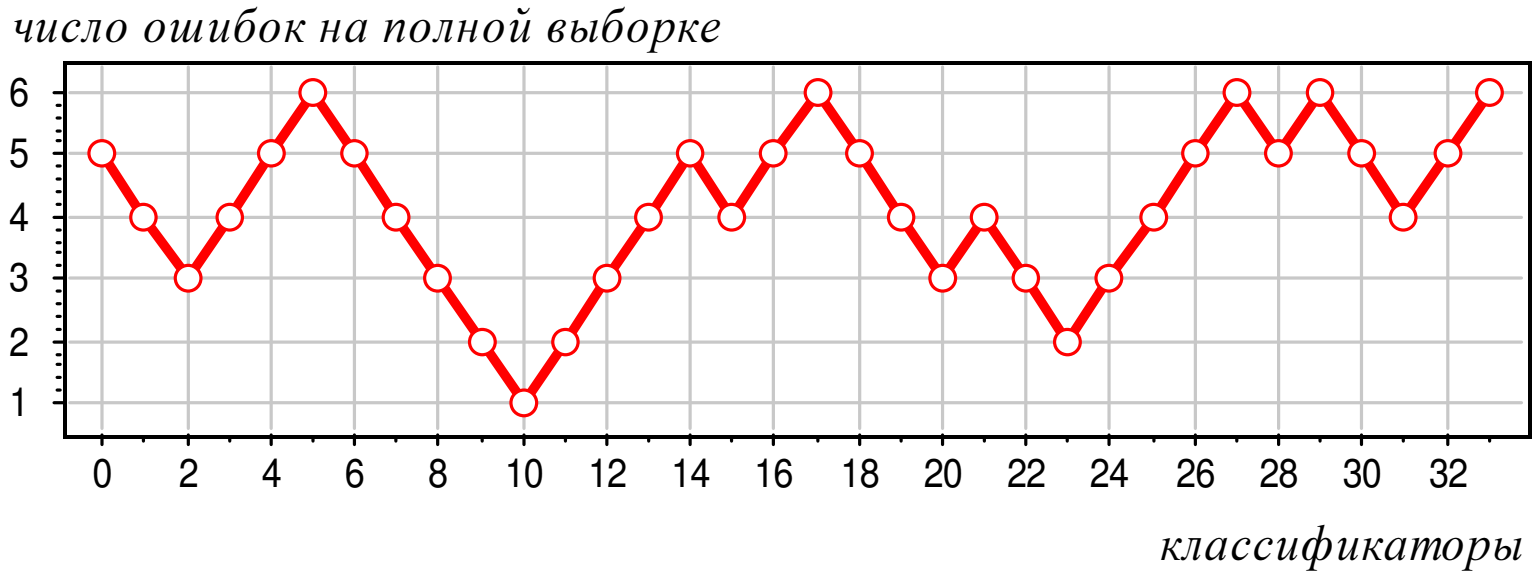


Схема работы алгоритма

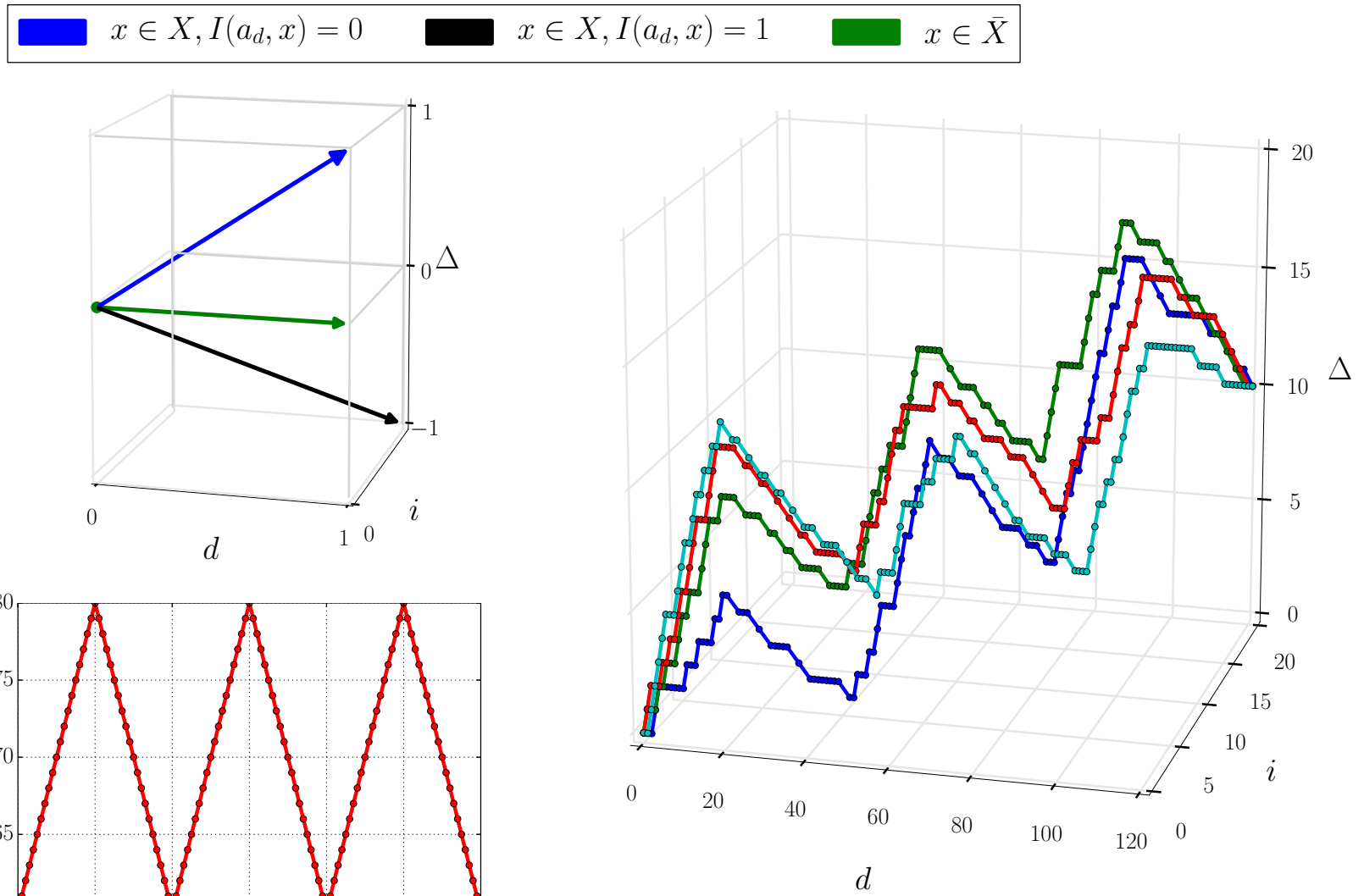
$$Q_\varepsilon = P[\delta(\mu X, X) \geqslant \varepsilon] = \sum_{d=0}^{D} P[\mu X = a_d][\delta(a_d, X) \geqslant \varepsilon]$$

$[\mu X = a_d]$ зависит только от выбора разбиения $\{X \cap \mathbb{G}, \bar{X} \cap \mathbb{G}\}$.
Зададим параметры $t = |X \cap \mathbb{G}|$, $e = n(a_d, X \cap \mathbb{G})$,

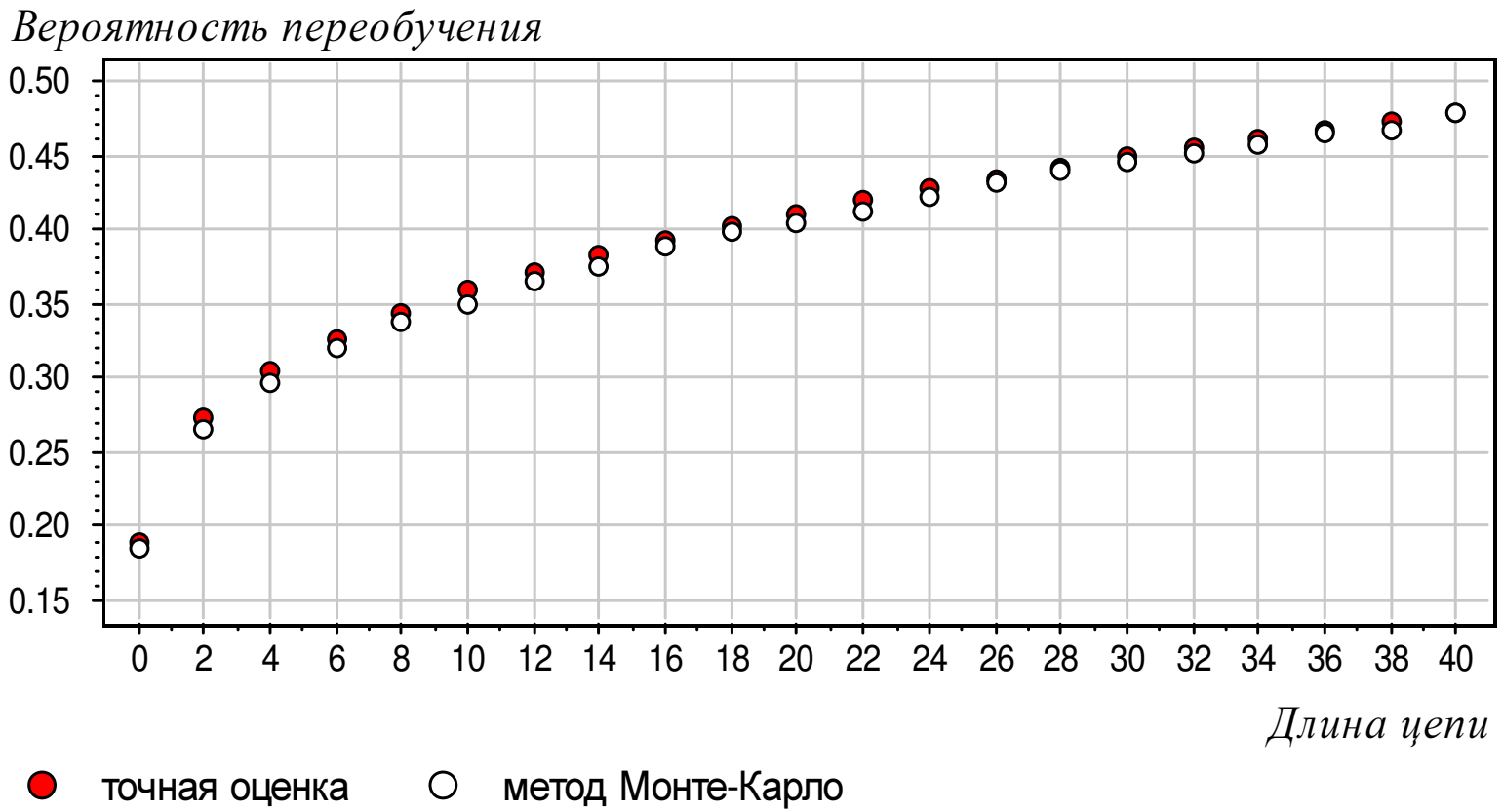
$$Q_\varepsilon = \sum_d \sum_t \sum_e G_d(t, e) N_d(t, e, \varepsilon),$$

где $G_d(t, e)$ – количество разбиений множества $\mathbb{G}$, таких, что $[\mu X = a_d]$, множители $N_d(t, e, \varepsilon)$ вычисляются аналитически.

Разбиение как блуждание по трехмерной сетке



Проверка алгоритма вычисления вероятности переобучения



Условия эксперимента: Случайная цепь;
$L = 200$, $\ell = 100$, $m = 40$, $\varepsilon = 0.05$; метод Монте–Карло по 10^5 случайных разбиений.

Постановка задачи медицинской диагностики

Дано: Выборка $\mathbb{X}$ пациентов S – символьных последовательностей в 6-буквенном алфавите. Метки классов: 1 – болен, 0 – здоров.
Критерий качества: AUC (Area Under Curve) на контрольной выборке, вычисляемый 10×10-блочной кросс-валидацией.

Признаки – частоты p_w триграмм w, т.е. $6^3 = 216$ признаков.

Наивный байесовский классификатор (NB)

$$a(S) = \left[\ln \frac{\pi_1(S)}{\pi_0(S)} \geqslant \beta\right] = \left[\sum_w \gamma_w p_w(S) \geqslant \beta\right],$$

Для некоторых болезней дает AUC $95 - 99\%$.

Отбор признаков

Старая схема: top–K по информационному критерию, K – параметр. Веса γ_w вычисляются по аналитическим формулам.

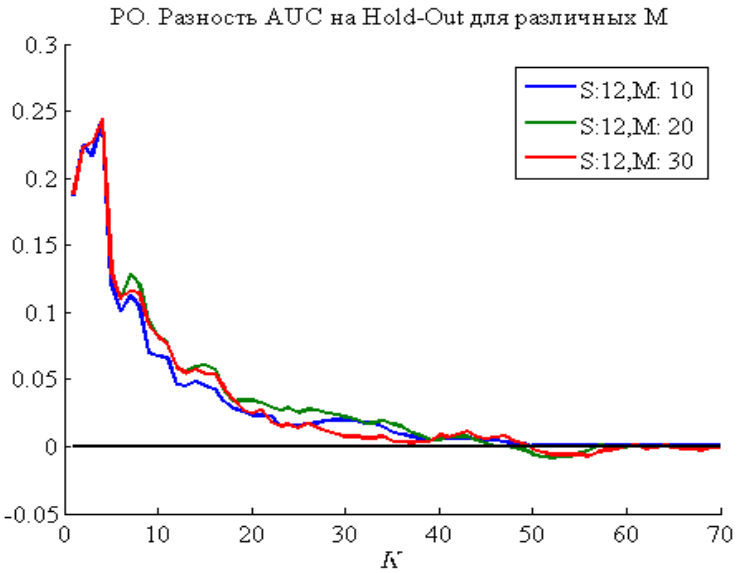
Новая схема: Жадный отбор из более широкого множества кандидатов, top–$(K^* + M)$ по информационному критерию. K^* – оптимальное значение, полученное по старой схеме.

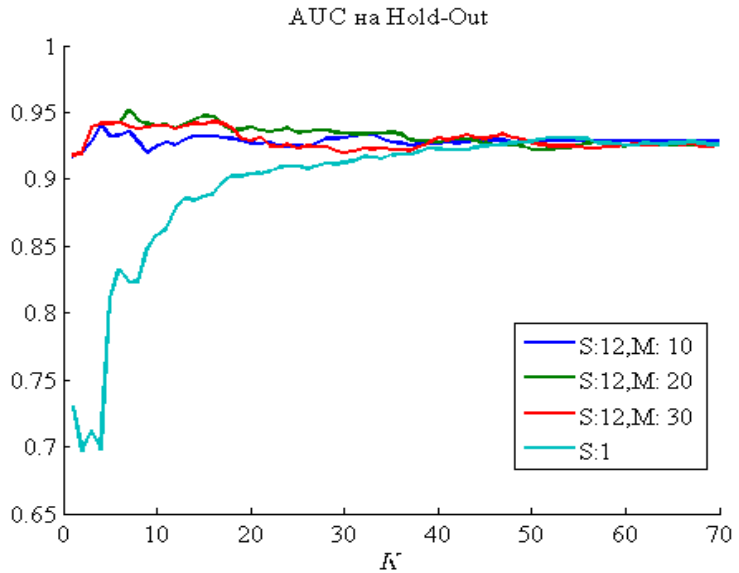
для всех $j = 1, \ldots, K^* + M$

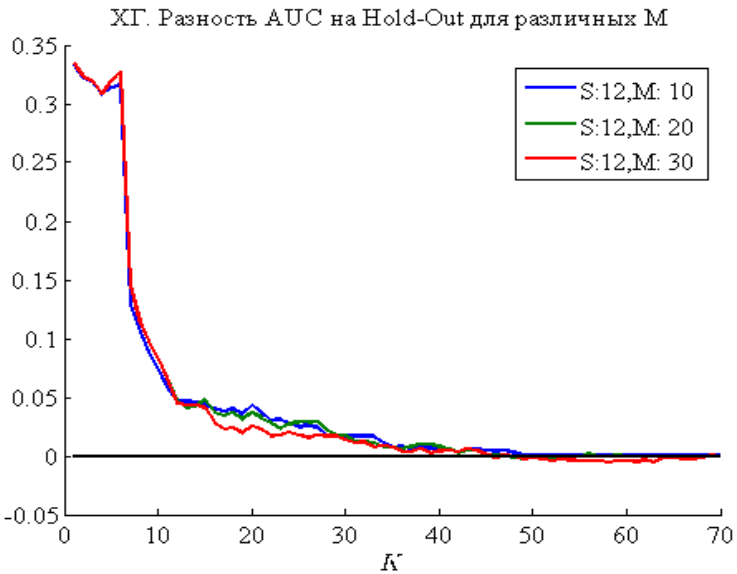
Из оставшихся $K^* + M - j + 1$ кандидатов выбирается тот, который максимально улучшает эмпирический риск на контрольной выборке

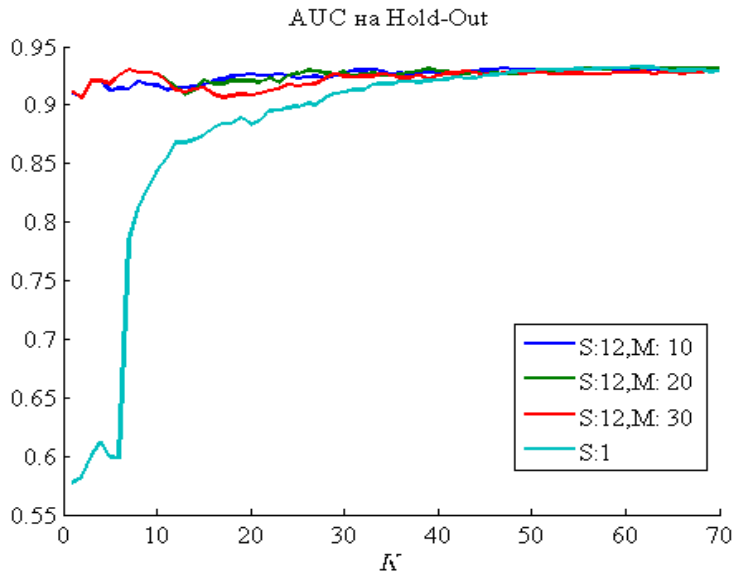
Веса γ_w вычисляются по аналитическим формулам.

Сравнение методов отбора









Выводы

Основные результаты

- Разработан алгоритм вычисления вероятности переобучения произвольной прямой цепи, полиномиальный по длине выборки
- Проведены вычислительные эксперименты с применением на практике подхода, основанного на вычислении переобученности

Открытые проблемы

- Применение в логических алгоритмах классификации
- Сравнение с существующими оценками переобучения в задаче отбора признаков
- Оптимизация весов γ_w признаков с помощью оценок переобучения