Комбинаторный подход к выводу точных оценок вероятности переобучения

Константин Воронцов
(«Участник:Vokov» на www.MachineLearning.ru)

Вычислительный Центр им. А. А. Дородницына РАН

Интеллектуализация Обработки Информации, ИОИ-8
18–22 октября 2010, Кипр, г. Пафос

Содержание

1. **Задача оценивания обобщающей способности**
   - Проблема переобучения
   - Слабая вероятностная аксиоматика
2. **Комбинаторные оценки вероятности переобучения**
   - Учёт эффектов расслоения и связности
   - Модельные семейства алгоритмов
   - Некоторые математические техники вывода оценок
3. **Направления исследований и открытые проблемы**
   - Задачи и открытые проблемы
   - Задача оценивания кривой обучения

Задача обучения по прецедентам. Матрица ошибок.

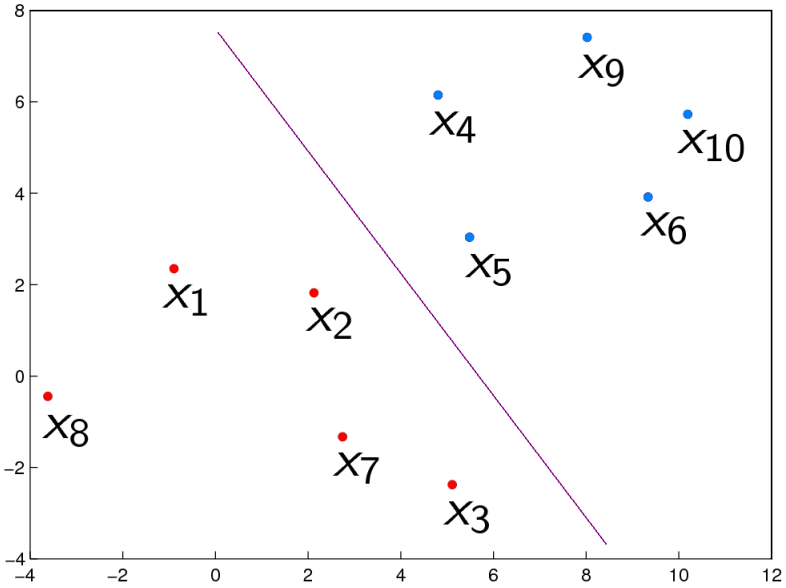
$\mathbb{X}^L = \{x_1, \ldots, x_L\}$ — конечное множество объектов;
$A = \{a_1, \ldots, a_D\}$ — конечное множество алгоритмов;
$I(a, x) = [$алгоритм a ошибается на объекте $x]$;

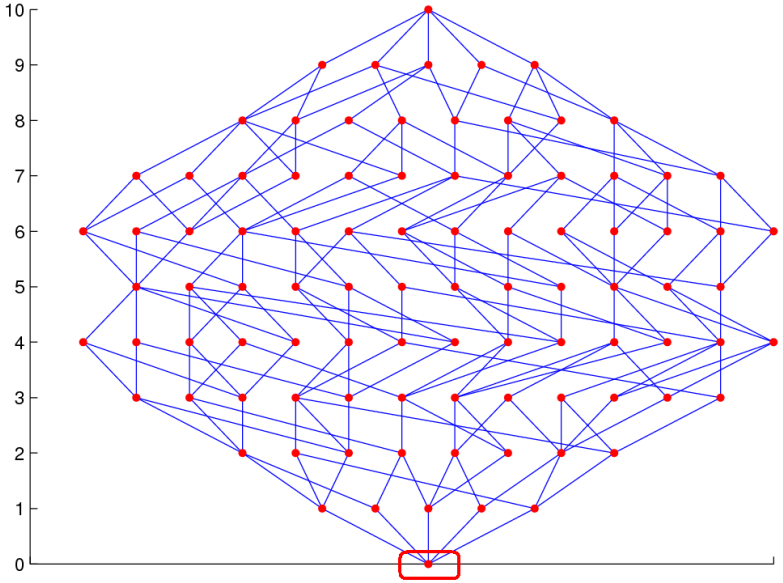
$L \times D$-матрица ошибок с попарно различными столбцами:

	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 — 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 $k = L - \ell$

$n(a, X)$ — *число ошибок* алгоритма a на выборке $X \subset \mathbb{X}^L$;
$\nu(a, X) = n(a, X)/|X|$ — *частота ошибок* a на выборке $X \subset \mathbb{X}^L$;

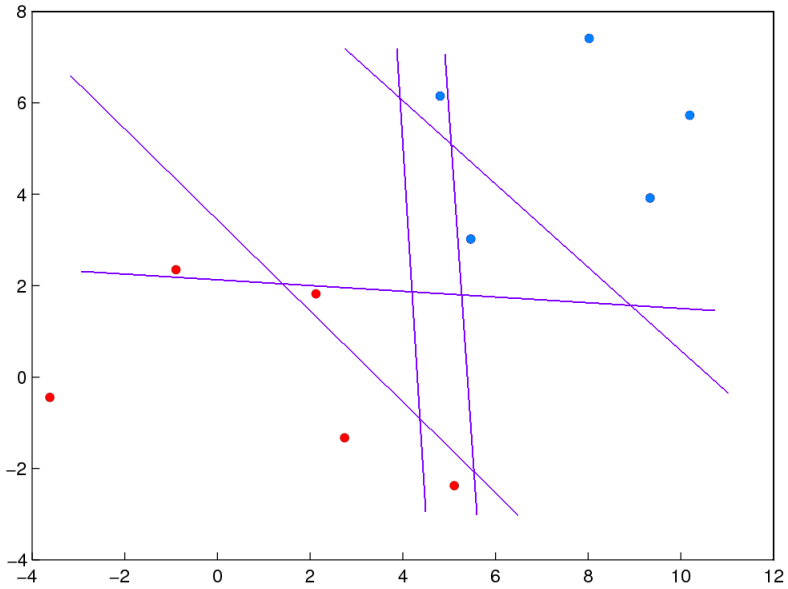
Пример. Матрица ошибок линейных классификаторов

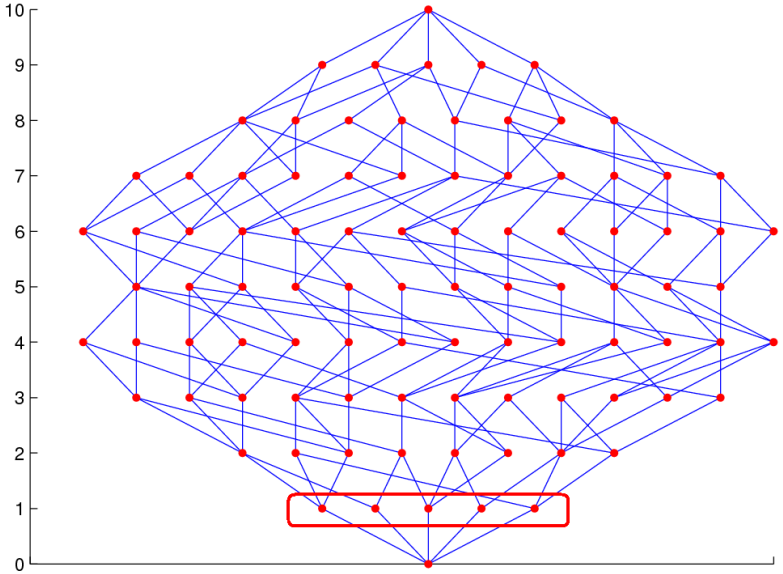




x_1	0
x_2	0
x_3	0
x_4	0
x_5	0
x_6	0
x_7	0
x_8	0
x_9	0
x_{10}	0

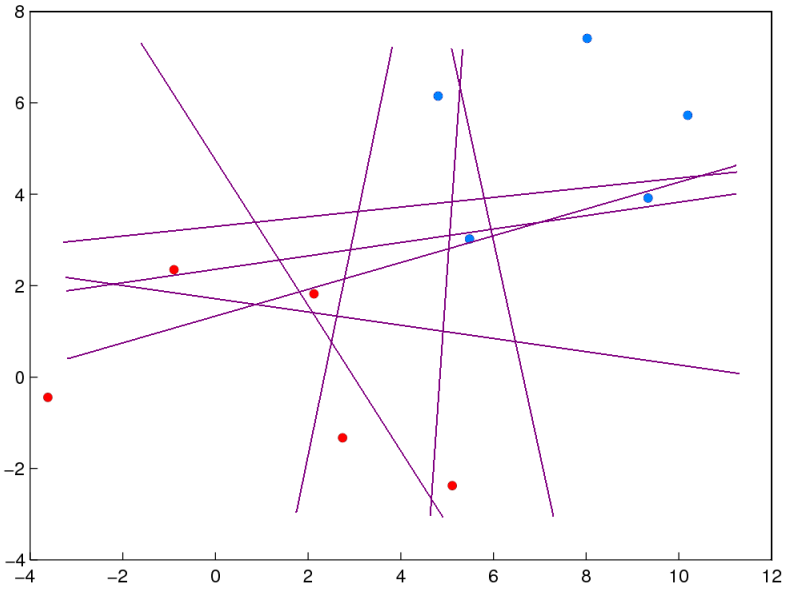
Пример. Матрица ошибок линейных классификаторов

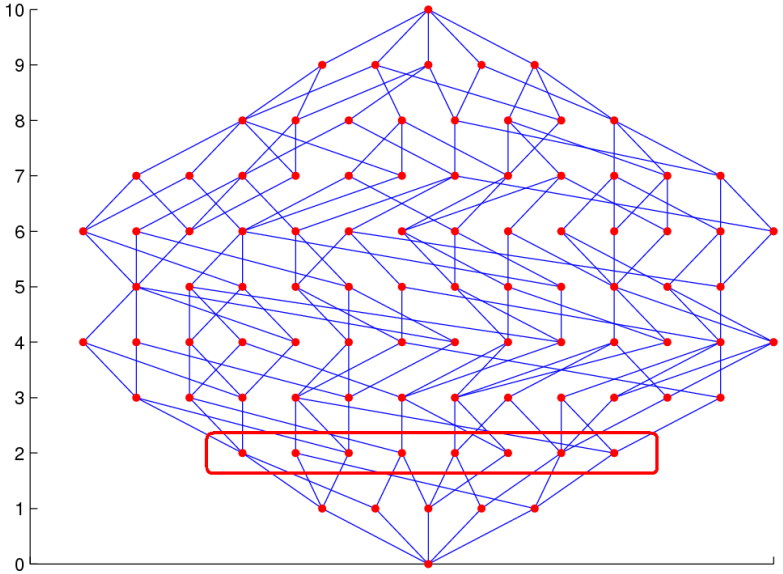




x_1	0	1	0	0	0	0
x_2	0	0	1	0	0	0
x_3	0	0	0	1	0	0
x_4	0	0	0	0	1	0
x_5	0	0	0	0	0	1
x_6	0	0	0	0	0	0
x_7	0	0	0	0	0	0
x_8	0	0	0	0	0	0
x_9	0	0	0	0	0	0
x_{10}	0	0	0	0	0	0

Пример. Матрица ошибок линейных классификаторов





	0-й слой	1-й слой					2-й слой								
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

Опр. *Метод обучения* $\mu\colon 2^{(\mathbb{X}^L)} \to A$ по произвольной выборке $X \subset \mathbb{X}^L$ выбирает некоторый алгоритм $a \in A$.

Опр. *Переобучение* при разбиении $X \sqcup \bar{X} = \mathbb{X}^L$:

$$\delta(\mu, X) \equiv \nu\big(\mu X, \bar{X}\big) - \nu\big(\mu X, X\big) \geqslant \varepsilon.$$

Опр. *Вероятность переобучения*

$$Q_\varepsilon(\mu, \mathbb{X}^L) = \mathsf{P}_X\big[\delta(\mu, X) \geqslant \varepsilon\big].$$

Опр. *Точная оценка*: $Q_\varepsilon(\mu, \mathbb{X}^L) = \eta(\varepsilon)$.

Опр. *Верхняя оценка*: $Q_\varepsilon(\mu, \mathbb{X}^L) \leqslant \eta(\varepsilon)$.

Единственное вероятностное допущение

Итак, $\mathbb{X}^L = \{x_1, \ldots, x_L\}$ — конечное множество объектов.

Аксиома

Все C_L^ℓ *разбиений* $\mathbb{X}^L = X \sqcup \bar{X}$ *равновероятны, где*
X — *наблюдаемая обучающая выборка длины* $\ell = |X|$;
$\bar{X}$ — *скрытая контрольная выборка длины* $k = |\bar{X}| = L - \ell$;

Вероятность понимается только как *доля разбиений выборки*:

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

Это аналог стандартной гипотезы о *независимости* наблюдений.
Теория меры и предельный переход $L \to \infty$ не используются.

Аналог закона больших чисел в слабой аксиоматике

Пусть $|A| = 1$, $\mu X = a$ для всех $X \subset \mathbb{X}^L$.
Обозначим $m = n(a, \mathbb{X}^L)$, $s = n(a, X)$.

Теорема (точная оценка)

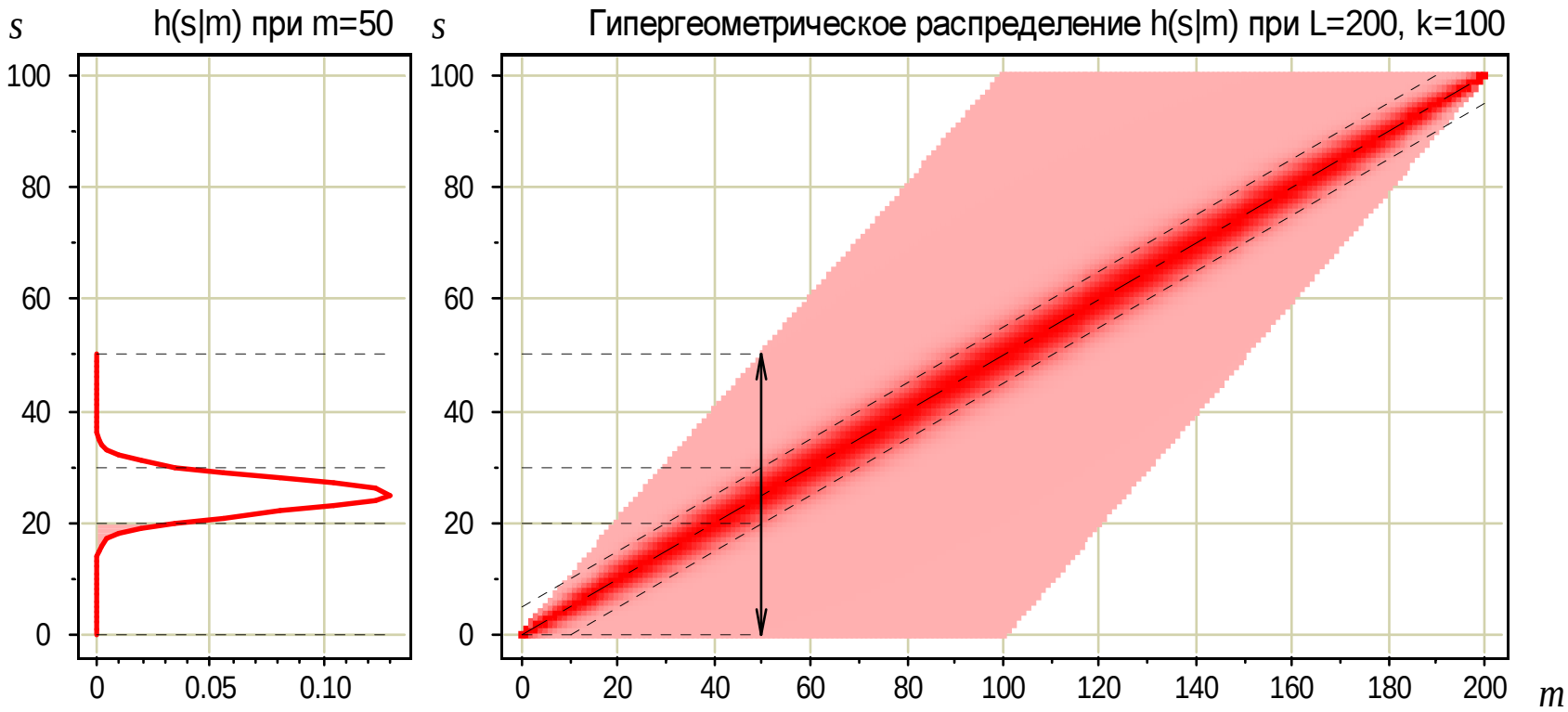
Вероятность большого уклонения частот описывается функцией гипергеометрического распределения (ГГР):

$$Q_\varepsilon(a, \mathbb{X}^L) = H_L^{\ell,m}(s_m(\varepsilon)), \quad s_m(\varepsilon) = \tfrac{\ell}{L}(m - \varepsilon k),$$

где $H_L^{\ell,m}(z) = \sum_{s=0}^{\lfloor z \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}$ *— левый «хвост» ГГР.*

Вывод: основная аксиома обеспечивает возможность предсказания скрытого $n(a, \bar{X})$ по наблюдаемому $n(a, X)$.

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



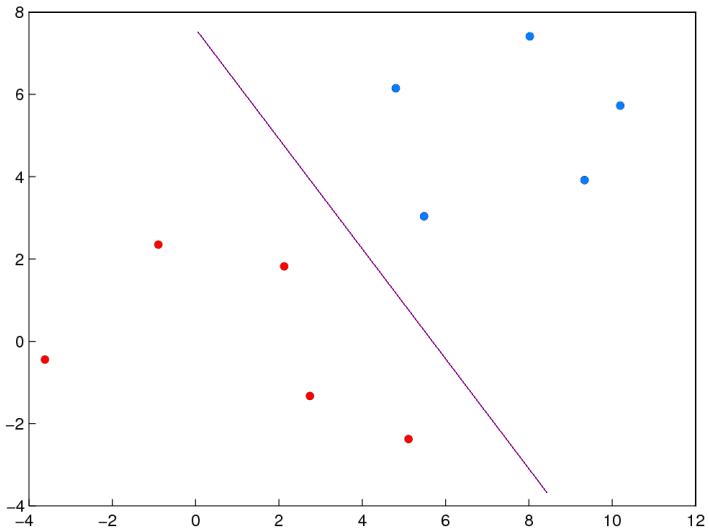
Предсказание $n(a, \bar{X})$ по $n(a, X)$ возможно благодаря узости гипергеометрического пика (*концентрации вероятности*).

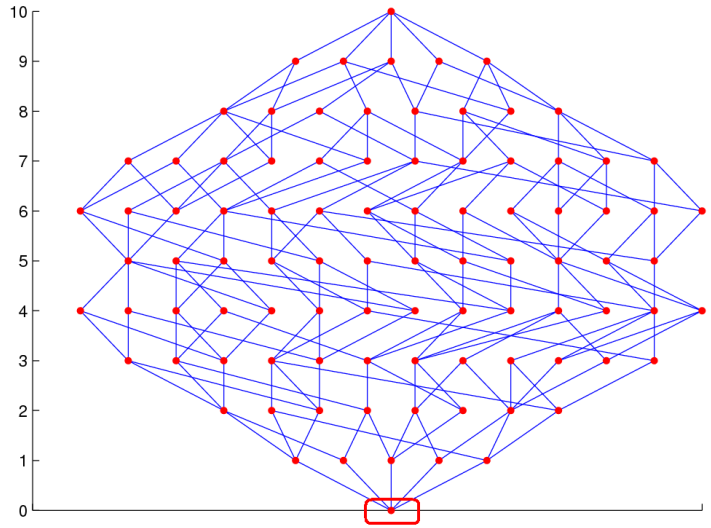
Закон больших чисел: $\nu(a, X) \to \nu(a, \mathbb{X}^L)$ при $\ell \to \infty$.

Ориентированный граф расслоения-связности

Множество *вершин графа* — все алгоритмы $a \in A$.

Множество *рёбер графа* E — все пары вершин (a, a') такие, что $n(a, \mathbb{X}^L) + 1 = n(a', \mathbb{X}^L)$ и $I(a, x_i) \leqslant I(a', x_i)$, $\forall x_i \in \mathbb{X}^L$.





Опр. $A_m = \{a \in A \colon n(a, \mathbb{X}^L) = m\}$ — *m-й слой* множества A.

Опр. $A = A_0 \sqcup \cdots \sqcup A_L$ — *расслоение* множества A.

Опр. $q(a) = \#\{a' \in A \colon (a, a') \in E\}$ — *связность* алгоритма $a \in A$.

Общая оценка расслоения-связности

Опр. *Профиль расслоения* $\Delta_m = |A_m|$.

Опр. *Профиль расслоения–связности* Δ_{mq} — это число алгоритмов в m-м слое со связностью q.

Опр. *Пессимистичная минимизация эмпирического риска* (ПМЭР):

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

Теорема

Если μ — ПМЭР, то для любой $\mathbb{X}^L$ и любого $\varepsilon \in [0, 1]$

$$Q_\varepsilon \leqslant \sum_{m=\lceil \varepsilon k \rceil}^{L} \sum_{q=0}^{L} \Delta_{mq} \cdot \frac{C_{L-q}^{\ell-q}}{C_L^\ell} \cdot H_{L-q}^{\ell-q,m}\left(s_m(\varepsilon)\right).$$

Сравнение с классическими оценками

Оценка для одного алгоритма:

$$Q_\varepsilon = H_L^{\ell,m}\left(s_m(\varepsilon)\right).$$

Оценка Вапника-Червоненкиса (1971):

$$Q_\varepsilon \leqslant \sum_{m=\lceil \varepsilon k \rceil}^{L} \Delta_m \cdot H_L^{\ell,m}\left(s_m(\varepsilon)\right) \leqslant$$
$$\leqslant |A| \cdot \max_{m=0,\ldots,L} H_L^{\ell,m}\left(s_m(\varepsilon)\right).$$

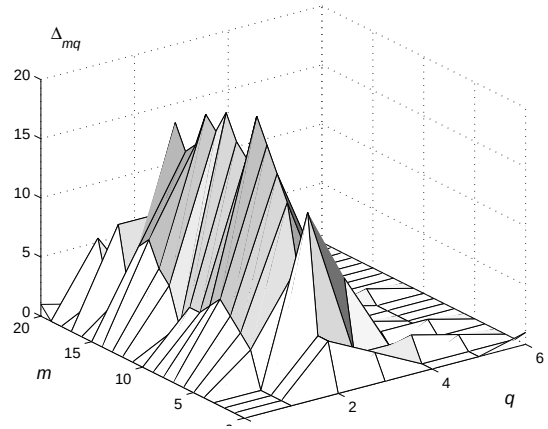
Оценка с учётом расслоения–связности (2010):

$$Q_\varepsilon \leqslant \sum_{m=\lceil \varepsilon k \rceil}^{L} \sum_{q=0}^{L} \Delta_{mq} \cdot \frac{C_{L-q}^{\ell-q}}{C_L^\ell} \cdot H_{L-q}^{\ell-q,m}\left(s_m(\varepsilon)\right).$$

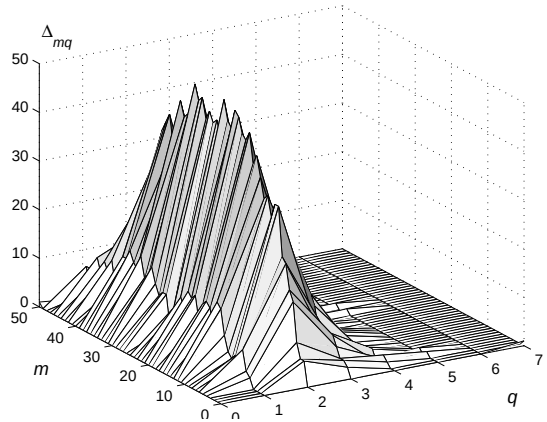
Пример. Профили расслоения-связности Δ_{mq}

линейно разделимые двумерные выборки длины L; линейные классификаторы

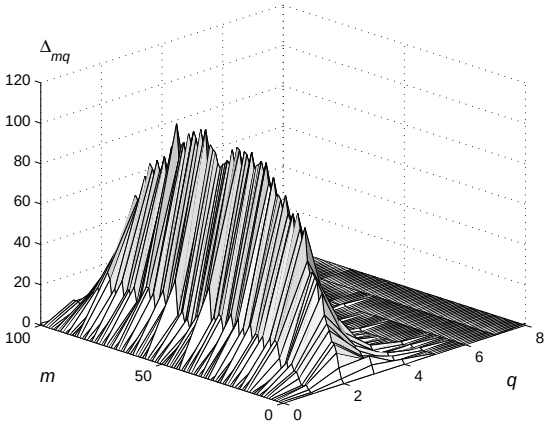
$L = 20$



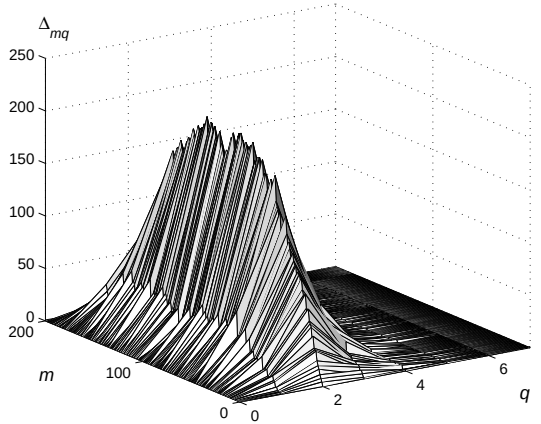
$L = 50$



$L = 100$



$L = 200$



Гипотеза сепарабельности: $\Delta_{mq} \approx \Delta_m \lambda_q$.

Гипотеза размерности: средняя связность $\approx$ размерность пространства

Насколько важно учитывать эффекты расслоения и связности?

Эксперимент с цепочками алгоритмов:

Цепочка с расслоением:

	a_1	a_2	a_3	a_4	a_5	$\ldots$	a_D
x_1	1	1	1	1	1	1	1
x_2	0 $\to$	1	1	1	1	1	1
x_3	0	0 $\to$	1	1	1	1	1
x_4	0	0	0 $\to$	1	1	1	1
x_5	0	0	0	0 $\to$	1	1	1
x_6	0	0	0	0	0 $\to$	1	1

Цепочка без расслоения:

	a_1	a_2	a_3	a_4	a_5	$\ldots$	a_D
x_1	1	1 $\to$	0	0	0	0	0
x_2	0 $\to$	1	1	1 $\to$	0	0	0
x_3	0	0	0 $\to$	1	1	1 $\to$	0
x_4	0	0	0	0	0 $\to$	1	1
x_5	0	0	0	0	0	0	0
x_6	0	0	0	0	0	0	0

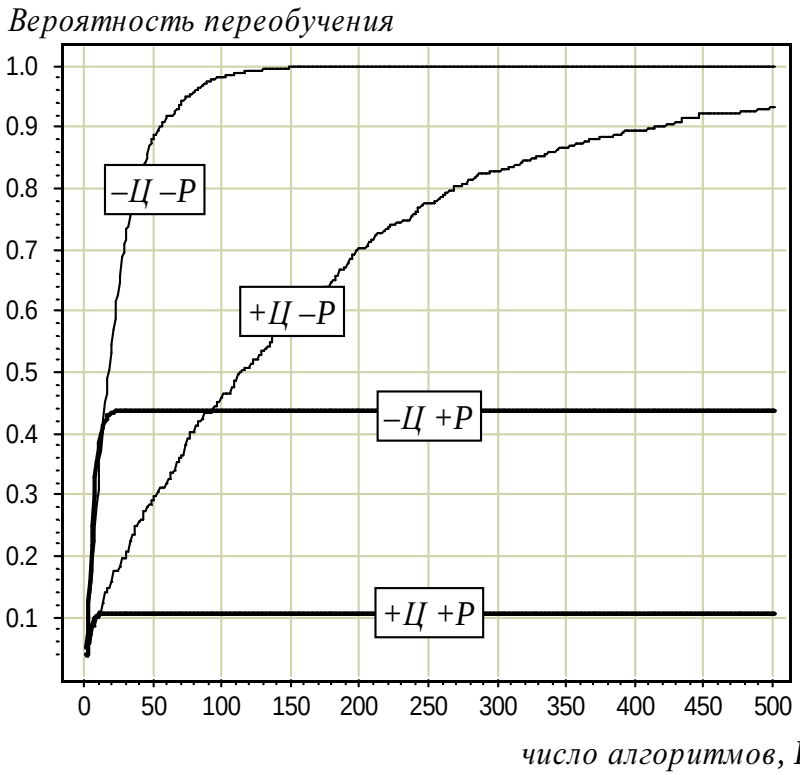
Для каждой цепочки генерируется *не-цепочка*
путём случайной перестановки в каждом столбце.

Итого имеем 4 модельных семейства:

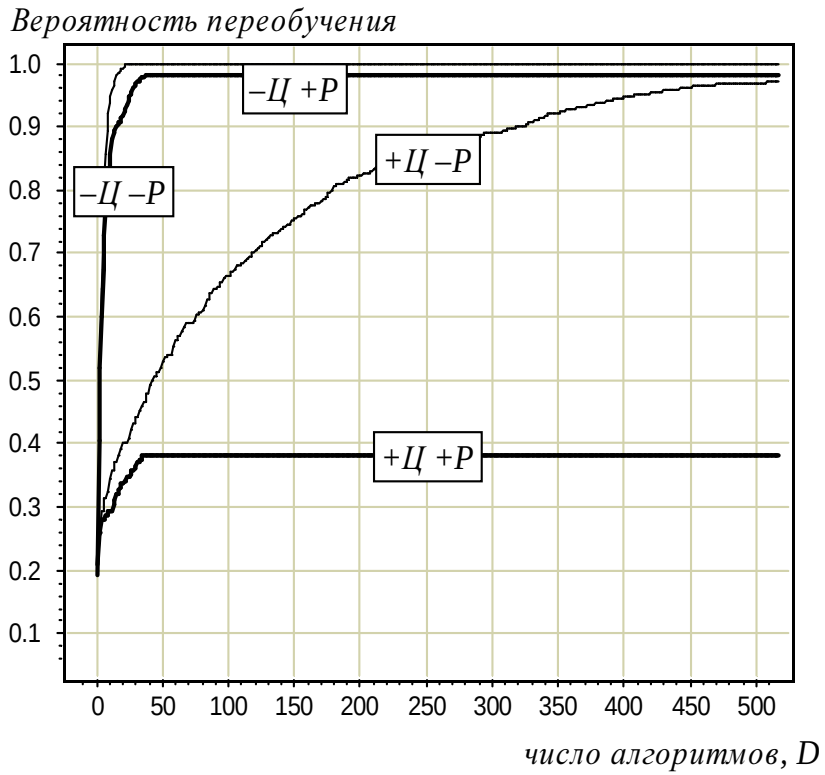
+Ц+Р	+Ц−Р
−Ц+Р	−Ц−Р

Эксперимент: зависимость Q_ε от D при $\ell = k = 100$, $\varepsilon = 0.05$

Простая задача, $n(a_1, \mathbb{X}^L) = 10$



Трудная задача, $n(a_1, \mathbb{X}^L) = 50$

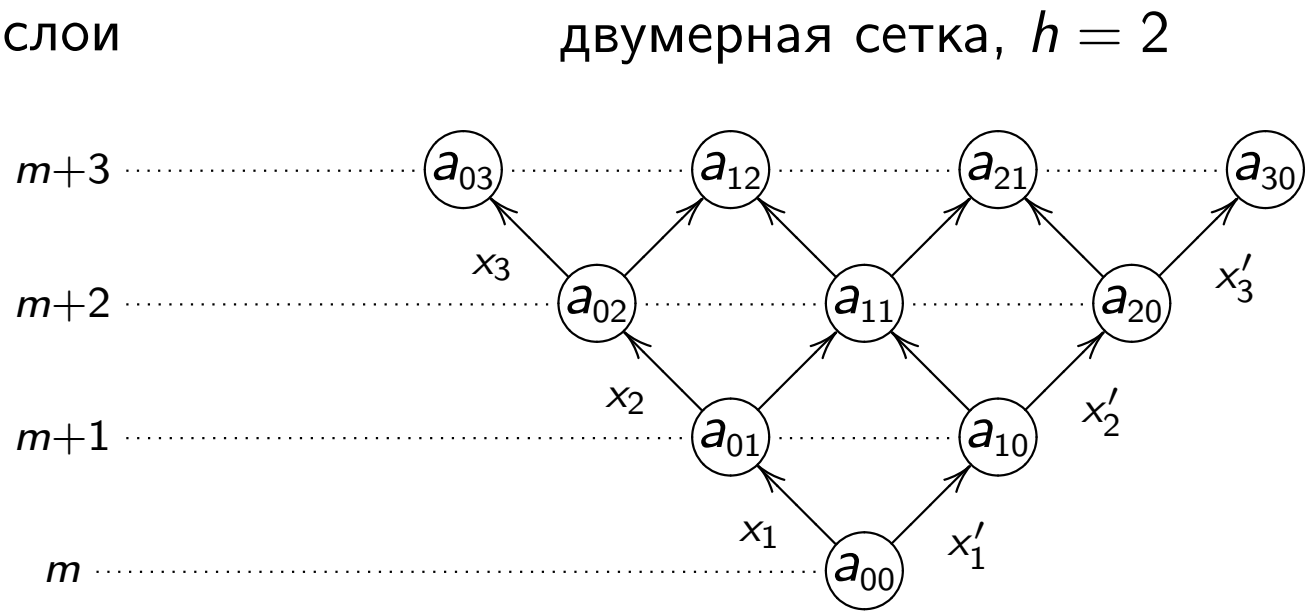


- *Связность* приводит к замедлению роста $Q_\varepsilon(D)$.
- *Расслоение* понижает уровень горизонтальной асимптоты.

Модельные и реальные семейства алгоритмов, для которых уже получены точные комбинаторные оценки переобучения

- монотонная и унимодальная цепочка алгоритмов;
- единичная окрестность лучшего алгоритма;
- слой булева куба;
- интервал булева куба; d его нижних слоёв;
- монотонные и унимодальные h-мерные сетки [П. Ботов];
- пучок h монотонных цепочек [П. Ботов, А. Фрей];
- симметричные семейства алгоритмов, их разреженные подмножества [А. Фрей];
- хэммингов шар, d его нижних слоёв, его разреженные подмножества [И. Толстихин];
- пороговые конъюнкции вещественных признаков [А. Ивахненко].
- метод ближайшего соседа;
- семейство монотонных классификаторов;

Монотонная сетка алгоритмов — модельное семейство, обладающее свойствами расслоения, связности и размерности



Эмпирический факт 1. Q_ε реальных семейств неплохо аппроксимируется Q_ε монотонной сетки при некоторой h.

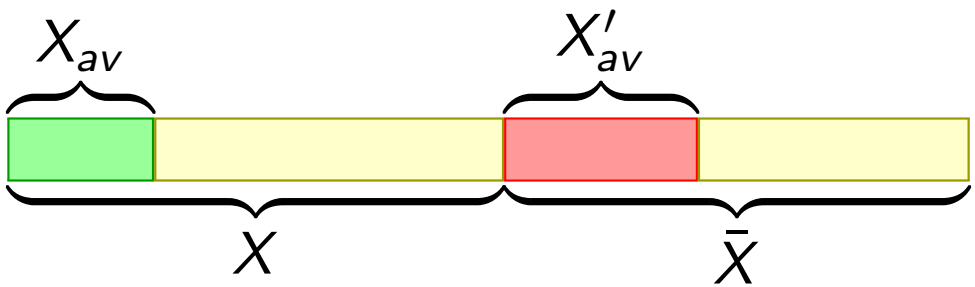
Эмпирический факт 2. Вероятность переобучения Q_ε монотонной сетки растёт практически линейно по h.

Метод порождающих и запрещающих множеств

Теорема

Для каждого $a \in A$ можно указать такой набор пар непересекающихся подмножеств объектов $X_{av}, X'_{av} \subset \mathbb{X}^L$, $v \in V_a$ и такой коэффициент $c_{av} \in \{-1, +1\}$, что

$$[\mu X{=}a] = \sum_{v \in V_a} c_{av} [X_{av} \subseteq X] [X'_{av} \subseteq \bar{X}].$$



Опр. X_{av} — множество объектов, *порождающих* алгоритм a.

Опр. X'_{av} — множество объектов, *запрещающих* алгоритм a.

Метод порождающих и запрещающих множеств

Теорема (точная оценка вероятности переобучения)

Вероятность получить в результате обучения алгоритм с заданным вектором ошибок a:

$$\mathsf{P}\big[\mu X{=}a\big] = \sum_{v \in V_a} c_{av} P_{av}; \quad P_{av} = \frac{C_{L_{av}}^{\ell_{av}}}{C_L^{\ell}}.$$

Вероятность переобучения:

$$Q_\varepsilon = \sum_{a \in A} \sum_{v \in V_a} c_{av} P_{av} H_{L_{av}}^{\ell_{av}, m_{av}} \left(s_{av}(\varepsilon)\right),$$

где $L_{av} = L - |X_{av}| - |X'_{av}|$,
$\ell_{av} = \ell - |X_{av}|$,
$m_{av} = n(a, \mathbb{X}^L) - n(a, X_{av}) - n(a, X'_{av})$,
$s_{av}(\varepsilon) = \frac{\ell}{L}\big(n(a, \mathbb{X}^L) - \varepsilon k\big) - n(a, X_{av})$.

Верхняя оценка вероятности переобучения

Опр. *Объекты, порождающие рёбра графа*, исходящие из a:

$$X_a^+ = \left\{ x_i \in \mathbb{X}^L \;\middle|\; \exists a' \colon (a, a') \in E,\ I(a, x_i) = 0,\ I(a', x_i) = 1 \right\}.$$

Теорема

Пусть μ — ПМЭР.
Если $\mu X = a$, то все $x_i \in X_a^+$ обязаны лежать в обучении:

$$\left[\mu X = a\right] \leqslant \left[X_a^+ \in X\right];$$

вероятность получить алгоритм a в результате обучения:

$$P_a = \mathsf{P}\left[\mu X{=}a\right] \leqslant C_{L-q(a)}^{\ell-q(a)} / C_L^\ell, \quad q(a) = |X_a^+|;$$

вероятность переобучения

$$Q_\varepsilon \leqslant \sum_{a \in A} P_a H_{L-q(a)}^{\ell-q(a),\, m_a}\left(\tfrac{\ell}{L}(m_a - \varepsilon k)\right), \quad m_a = n(a, \mathbb{X}^L).$$

Задачи и открытые проблемы

1. точные оценки переобучения для реальных методов;
2. построение разреженных подмножеств алгоритмов из нижних слоёв в реальных задачах;
3. снятие ограничения, что μ — ПМЭР;
4. обоснование гипотез сепарабельности Δ_{mq} и размерности;
5. получение комбинаторной оценки вероятности переобучения для линейных классификаторов;
6. переход от ненаблюдаемых оценок к наблюдаемым;
7. обобщение понятий расслоения и связности на случай произвольной (не бинарной) функции потерь;
8. онлайновое обучение в условиях нестационарной выборки (в случаях как плавных, так и скачкообразных изменений);

Задача онлайнового обучения (доклад в среду, 15:00)

$\mathbb{X}^L = (x_1, \ldots, x_T)$ — конечная последовательность объектов;
A — множество допустимых предсказаний;

Процесс Online Learning

1: **для всех** $t := 1, \ldots, T$
2: $a_{t+1} := \mu(x_1, \ldots, x_t)$ — предсказание из A;
3: x_{t+1} становится известен;
4: $\mathscr{L}(a_{t+1}, x_{t+1})$ — величина потерь от предсказания;

Задача

Найти кривую обучения — зависимость математического ожидания потери от времени:

$$Q(t) = \mathsf{E}\mathscr{L}(a_{t+1}, x_{t+1}), \quad t = 1, \ldots, T-1.$$

$Q(t)$ характеризует обучаемость предсказывающего алгоритма μ.

Вопросы?

Воронцов Константин Вячеславович
vokov@forecsys.ru

Страницы на www.MachineLearning.ru:

- Участник:Vokov
- Слабая вероятностная аксиоматика
- Расслоение и сходство алгоритмов (виртуальный семинар)