Комбинаторная теория переобучения

К. В. Воронцов
(vokov@forecsys.ru, http://www.ccas.ru/voron)

Вычислительный Центр им. А. А. Дородницына РАН

Математические методы распознавания образов, ММРО-14
21–25 сентября 2009
г. Суздаль

Содержание

Основные понятия и обозначения

Объекты $\mathbb{X} = \{x_1, \ldots, x_L\}$; алгоритмы $A = \{a_1, \ldots, a_D\}$;

$I(a, x) = \bigl[$алгоритм a ошибается на объекте $x\bigr]$;

$\vec{a} = \bigl(I(a, x_i)\bigr)_{i=1}^{L}$ — *вектор ошибок* алгоритма $a \in A$;

	a_1	a_2	a_3	a_4	a_5	a_6	a_7	$\cdots$	a_D	
x_1	1	1	0	0	0	1	1	$\cdots$	1	
x_2	0	0	0	0	1	1	1	$\cdots$	1	X, обучение
x_3	0	0	1	0	0	0	0	$\cdots$	0	
x_4	0	0	0	1	1	1	1	$\cdots$	0	
x_5	0	0	0	1	0	0	0	$\cdots$	1	$\bar{X}$, контроль
x_L	0	1	1	1	1	1	0	$\cdots$	0	

$n(a, X) = \sum\limits_{x_i \in X} I(a, x_i)$ — *число ошибок* a на выборке $X \subset \mathbb{X}$;

Задача: зная только X, выбрать алгоритм с малым $n(a, \mathbb{X})$.

Постановка задачи

Метод обучения μ — минимизация эмпирического риска:

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

Переобученность метода μ при разбиении $X \sqcup \bar{X} = \mathbb{X}$:

$$\delta_\mu(X) = \frac{1}{|\bar{X}|} n\bigl(\mu X, \bar{X}\bigr) - \frac{1}{|X|} n\bigl(\mu X, X\bigr).$$

Аксиома (слабая вероятностная аксиоматика)

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

Основная задача — оценить вероятность переобучения:

$$Q_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\bigl[\delta_\mu(X) \geqslant \varepsilon\bigr] = \frac{1}{C_L^\ell} \sum_{X \sqcup \bar{X}} \bigl[\delta_\mu(X) \geqslant \varepsilon\bigr].$$

Простейший частный случай: один алгоритм

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

Теорема (точная оценка)

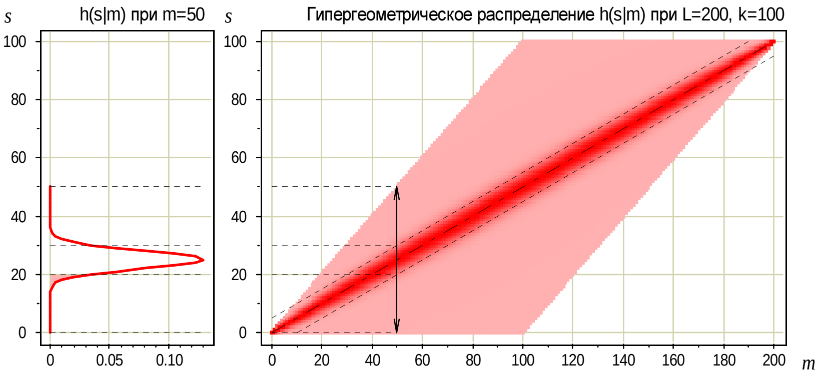
Вероятность большого уклонения частот описывается гипергеометрической функцией распределения:

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

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

Вывод: одно лишь предположение независимости объектов (⇔ равной вероятности разбиений) 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: $\frac{1}{\ell} n(a, X)$, $\frac{1}{k} n(a, \bar{X}) \to \frac{1}{L} n(a, \mathbb{X})$ при $\ell, k \to \infty$.

Теория Вапника-Червоненкиса

Теорема (VC-оценка, 1971)

Для любых $\mathbb{X}$, μ, $\varepsilon \in (0,1)$

$$\mathsf{P}\bigl[\delta_\mu(X) \geqslant \varepsilon\bigr] \leqslant \Delta^A(\mathbb{X}) \max_{m=0..L} H_L^{\ell,m}\left(\tfrac{\ell}{L}(m - \varepsilon k)\right).$$

где $\Delta^A(\mathbb{X}) = \#\bigl\{\vec{a} \bigm| a \in A\bigr\}$ *— коэффициент разнообразия (shatter coefficient) множества алгоритмов* A.

Экспериментальное измерение завышенности VC-оценки:

$\Delta^A(\mathbb{X}) \sim 10^6 - 10^{12}$ — коэффициент разнообразия (КР);
$\hat{\Delta}^A(\mathbb{X}) \sim 10^1 - 10^2$ — эффективный локальный КР (ЭЛКР)

Vorontsov K. V. Combinatorial probability and the tightness of generalization bounds // Pattern Recognition and Image Analysis. MAIK Nauka. No 2, Vol. 18, 2008, Pp. 243–259.

Эксперимент 1: Выявление причин завышенности VC-оценок

Эксперименты по оцениванию отдельных причин завышенности:
— логические алгоритмы классификации (Forecsys LogicPro);
— 7 реальных задач классификации из репозитория UCI;
— удалось раздельно оценить 4 фактора завышенности.

Основные факторы завышенности:

- не учитывается *расслоение* семейства алгоритмов:
  чем выше $m = n(a, \mathbb{X})$, тем меньше $\mathsf{P}[\mu X = a]$
  (фактор завышенности — в 10^2–10^5 раз);
- не учитывается *связность* семейства алгоритмов:
  чем больше схожих алгоритмов, тем меньше разнообразие
  (фактор завышенности — в 10^3–10^4 раз).

Vorontsov K. V. Combinatorial probability and the tightness of generalization bounds // Pattern Recognition and Image Analysis. MAIK Nauka. No 2, Vol. 18, 2008, Pp. 243–259.

Эксперимент 2: Цепочки алгоритмов

Цель эксперимента: на простом примере продемонстрировать влияние связности и расслоения на вероятность переобучения.

Цепочка с расслоением:

	a_0	a_1	a_2	a_3	a_4	a_5	a_6
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
x_7	0	0	0	0	0	0 $\to$	1

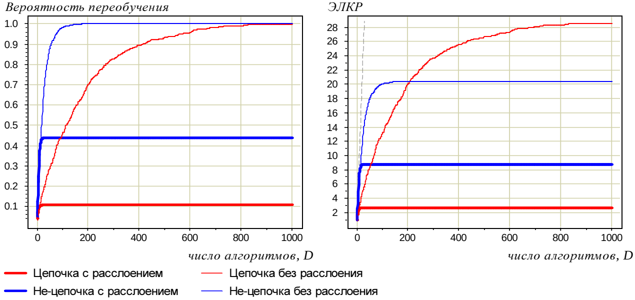
Цепочка без расслоения:

	a_0	a_1	a_2	a_3	a_4	a_5	a_6
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
x_7	0	0	0	0	0	0	0

Для каждой цепочки генерируется *не-цепочка* путём случайной перестановки единиц в каждом столбце — чтобы устранить свойство связности в семействе.

Цепочки и не-цепочки; с расслоением и без

Условия эксперимента: $D{=}1000$, $\ell{=}k{=}100$, $\varepsilon{=}0.05$, $n(a_0, \mathbb{X}){=}10$.



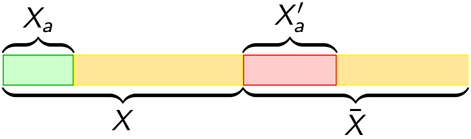
1. *Связность* приводит к замедлению роста $Q_\varepsilon(D)$.
2. *Расслоение* понижает уровень горизонтальной асимптоты.
3. **Без расслоения и связности $Q_\varepsilon = \frac{1}{2}$ при $D \approx 20 \Rightarrow$**
   $\Rightarrow$ реальные семейства и расслоены, и связны.

Гипотеза о порождающих и запрещающих объектах

Гипотеза (1)

Для каждого $a \in A$ можно указать пару непересекающихся подмножеств объектов $X_a \subset \mathbb{X}$, $X'_a \subset \mathbb{X}$ такую, что:

$$(\mu X = a) \;\Leftrightarrow\; (X_a \subseteq X) \text{ и } (X'_a \subseteq \bar{X}), \quad \forall X \subset \mathbb{X}.$$



Опр. X_a — множество объектов, *порождающих* алгоритм a.

Опр. X'_a — множество объектов, *запрещающих* алгоритм a.

Опр. $\mathbb{X} \backslash (X_a \cup X'_a)$ — множество объектов, *нейтральных* для a.

Обозначения и основная лемма

Введём для каждого $a \in A$ следующие обозначения:

$L_a = L - |X_a| - |X'_a|$ — число нейтральных объектов;
$\ell_a = \ell - |X_a|$ — число нейтральных обучающих объектов;

Лемма (о вероятности получения алгоритма)

Если гипотеза (1) справедлива, то вероятность получить в результате обучения алгоритм a равна доле разбиений, при которых объекты из X_a и X'_a остаются на своих местах:

$$P_a = \mathsf{P}\bigl[\mu X{=}a\bigr] = \frac{C_{L_a}^{\ell_a}}{C_L^{\ell}}.$$

Ещё обозначения и основная теорема

$m_a = n(a, \mathbb{X}) - n(a, X_a) - n(a, X'_a)$
— число ошибок алгоритма a на нейтральных объектах;

$s_a(\varepsilon) = \frac{\ell}{L}\bigl(n(a, \mathbb{X}) - \varepsilon k\bigr) - n(a, X_a)$
— наибольшее число ошибок алгоритма a на нейтральных обучающих объектах $X \setminus X_a$, при котором $\delta(a, X) \geqslant \varepsilon$.

Теорема (точная оценка вероятности переобучения)

Если гипотеза (1) справедлива, то

$$\mathsf{P}\bigl[\delta_\mu(X) \geqslant \varepsilon\bigr] = \sum_{a \in A} P_a H_{L_a}^{\ell_a, m_a}\bigl(s_a(\varepsilon)\bigr).$$

Нестрогое доказательство основной теоремы

Для каждого вектора ошибок a из A известна:
1) вероятность получить a в результате обучения:

$$P_a = \mathsf{P}\bigl[\mu X{=}a\bigr] = C_{L_a}^{\ell_a} \;/\; C_L^{\ell};$$

2) условная вероятность большого уклонения, если получен a:

$$\mathsf{P}\bigl[\delta(a,X) \geqslant \varepsilon \mid \mu X{=}a\bigr] = H_{L_a}^{\ell_a, m_a}\bigl(s_a(\varepsilon)\bigr).$$

Тогда, по формуле полной вероятности

$$Q_\varepsilon = \sum_{a\in A} \mathsf{P}\bigl[\mu X{=}a\bigr]\,\mathsf{P}\bigl[\delta(a,X) \geqslant \varepsilon \mid \mu X{=}a\bigr].$$

$$= \sum_{a\in A} \frac{C_{L_a}^{\ell_a}}{C_L^{\ell}} \cdot H_{L_a}^{\ell_a, m_a}\bigl(s_a(\varepsilon)\bigr).$$

Обобщение основной гипотезы

Гипотеза (2)

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

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

Обозначения: для каждого $a \in A$ и каждого $v \in V_a$

$$L_{av} = \ell_{av} + k_{av}; \quad \ell_{av} = \ell - |X_{av}|;$$
$$m_{av} = n(a, \mathbb{X}) - n(a, X_{av}) - n(a, X'_{av});$$
$$s_{av}(\varepsilon) = \tfrac{\ell}{L}\bigl(n(a, \mathbb{X}) - \varepsilon k\bigr) - n(a, X_{av}).$$

Обобщение: лемма и основная теорема

Лемма (о вероятностях получения алгоритмов)

Если гипотеза (2) справедлива, то вероятность получить в результате обучения алгоритм с вектором ошибок a

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

Теорема (точная оценка вероятности переобучения)

Если гипотеза (2) справедлива, то

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

Сильное ли ограничение накладывает гипотеза (2)?

Оказывается, почти не накладывает.
Это общий случай!

Теорема

Пусть векторы ошибок алгоритмов $a_1, \ldots, a_D$ *попарно различны и метод* μ *минимизирует эмпирический риск.*

Тогда справедлива гипотеза (2), причём $c_{av} = 1$.

Доказательство конструктивно, но «тавтологично» — по матрице ошибок строится *система подмножеств* $\mathbb{S}(a) = (X, \bar{X})_{\mu X = a}$.

НО система подмножеств $\mathbb{S}(a)$ может быть не единственна.

Открытая проблема:
Как искать системы подмножеств с наименьшими $|X_{av}|$, $|X'_{av}|$?

Связные семейства и цепочки

Опр. 1. Семейство A *связное*, если $\forall a \in A\ \exists a' \in A$: $\rho(a, a') = 1$.

$\rho(a, a')$ — хэммингово расстояние между векторами ошибок:

$$\rho(a, a') = \sum_{i=1}^{L} \bigl| I(a, x_i) - I(a', x_i) \bigr|, \quad \forall a, a' \in A.$$

Опр. 2. *Цепочка алгоритмов* — последовательность $a_0, a_1, \ldots, a_D$ такая, что $\rho(a_{d-1}, a_d) = 1$ для всех $d = 1, \ldots, D$.

Опр. 3. Цепочка алгоритмов *монотонная*, если $I(a_{d-1}, x) \leqslant I(a_d, x)$ для всех $x \in \mathbb{X}$ и $d = 1, \ldots, D$.

Пример:

	a_1	a_2	a_3	a_4	$\cdots$	a_D
x_1	0	1	1	1	1	1
x_2	0	0	1	1	1	1
x_3	0	0	0	1	1	1
$\cdots$	0	0	0	0	1	1
x_L	0	0	0	0	0	1

Вероятность переобучения монотонной цепочки

Пусть μ — *пессимистичная минимизация эмпирического риска* (выбор алгоритма по принципу «худший из лучших»):

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

Теорема (точная верхняя оценка)

Пусть $a_0, a_1, \ldots, a_D$ — *монотонная цепочка,* $n(a_0,\mathbb{X}) = m$, $k \leqslant D \leqslant L - m$. *Тогда*

$$P_d = \mathsf{P}\bigl[\mu X{=}a\bigr] = \frac{C_{L-d-1}^{\ell-1}}{C_L^{\ell}};$$

$$\mathsf{P}\bigl[\delta_\mu(X) \geqslant \varepsilon\bigr] = \sum_{d=0}^{k} P_d H_{L-d-1}^{\ell-1,m}\left(\tfrac{\ell}{L}(m + d - \varepsilon k)\right).$$

Идея доказательства

Перенумеруем объекты так, чтобы a_d ошибался на $x_1, \ldots, x_d$.

	a_0	a_1	a_2	a_3	$\cdots$	a_D
x_1	0	1	1	1	1	1
x_2	0	0	1	1	1	1
x_3	0	0	0	1	1	1
x_4	0	0	0	0	1	1
$\cdots$	0	0	0	0	0	1
$\cdots$	0	0	0	0	0	0
$\cdots$	0	0	0	0	0	0
$\cdots$	1	1	1	1	1	1
x_L	1	1	1	1	1	1

Тогда:

$$(\mu X = a_d) \;\Leftrightarrow\; (x_{d+1} \in X) \text{ и } (x_1, \ldots, x_d \in \bar{X}), \text{ при } d \leqslant k;$$

$$(\mu X = a_d) \text{ невозможно, при } d > k.$$

Другие модельные семейства

Семейства, для которых к настоящему времени получены точные оценки вероятности переобучения:

- монотонная цепочка;
- унимодальная цепочка (применяется Гипотеза (2));
- единичная окрестность лучшего алгоритма;
- семейства из небольшого числа алгоритмов (2,3,...);
- монотонная и унимодальная h-мерные сетки [П. Ботов];
- некоторые симметричные семейства [А. Фрей];
- конъюнкции пороговых предикатов [А. Ивахненко];
- связные семейства общего вида [Д. Кочедыков].

Постановка задачи рекуррентного вычисления Q_ε

$\mathbb{S}(a) = \langle X_{av}, X'_{av}, c_{av} \rangle_{v \in V_a}$ — информация об алгоритме $a \in A$, необходимая для вычисления вероятности переобучения Q_ε.

Пусть $n(a_0, \mathbb{X}) \leqslant n(a_1, \mathbb{X}) \leqslant \cdots \leqslant n(a_D, \mathbb{X})$.

Пусть μ_d — пессимистичный метод обучения, выбирающий алгоритмы только из подмножества $A_d = \{a_0, \ldots, a_d\}$.

Задача (пересчёт Q_ε при добавлении алгоритма a_d)

Известна *информация* $\mathbb{S}(a_t)$ *относительно метода* μ_{d-1} *для всех алгоритмов* a_t, $t \leqslant d-1$.

Вычислить *информацию* $\mathbb{S}(a_t)$, $t \leqslant d$ *относительно метода* μ_d.

Дополнительное предположение: $n(a_0, \mathbb{X}) = 0$.

Теоремы о рекуррентном вычислении Q_ε

Теорема (об информации $\mathbb{S}(a_d)$)

$$\bigl[\mu_d X{=}a_d\bigr] = \bigl[X'_d \subseteq \bar{X}\bigr], \quad X'_d = \bigl\{x_i \in \mathbb{X}\colon I(a_d, x_i) = 1\bigr\}.$$

Теорема (о коррекции информации $\mathbb{S}(a_t)$, $t < d$)

Для каждого $v \in V_t$ такого, что $X_{tv} \cap X'_d = \varnothing$

1) *если $X'_d \setminus X'_{tv} = \{x_i\}$ — одноэлементное множество, то присоединить x_i к X_{tv};*
2) *если $|X'_d \setminus X'_{tv}| > 1$, то добавить в V_t индекс w, положив $c_{tw} = -c_{tv}$, $X_{tw} = X_{tv}$, $X'_{tw} = X'_{tv} \cup X'_d$;*
3) *если $|X'_d \setminus X'_{tv}| = 0$, то удалить из V_t индекс v.*

Упрощённое рекуррентное вычисление Q_ε

Теорема (О рекуррентной верхней оценке)

*Если иногда пропускать шаг 2) при $c_{tv} = 1$,
то вычисляемое значение Q_ε может только увеличиться.*

Теорема

*Если всегда пропускать шаг 2),
то вычисляемое значение Q_ε может только увеличиться.
При этом условия шага 3) не будут выполняться никогда.*

Рекуррентное вычисление Q_ε может занять время $O(L2^D)$.
Упрощённое рекуррентное вычисление Q_ε занимает $O(LD^2)$.
Его можно сократить до $O(LD)$ и даже до $O(L)$.

Расслоение и связность

Расслоение множества алгоритмов $A = \bigsqcup\limits_{m=0}^{L} A_m$, где

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

Отношение частичного порядка на алгоритмах:

$$a \prec a' \;\Leftrightarrow\; \bigl(I(a,x) \leqslant I(a',x),\ \forall x \in \mathbb{X}\bigr).$$

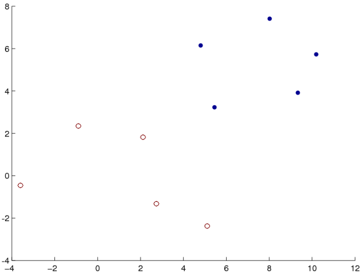
Связность $q(a)$ алгоритма $a \in A$ — число сравнимых с ним алгоритмов в следующем слое:

$$q(a) = \#\Bigl\{a'\colon a \prec a' \text{ и } n(a,\mathbb{X}) + 1 = n(a',\mathbb{X})\Bigr\}.$$

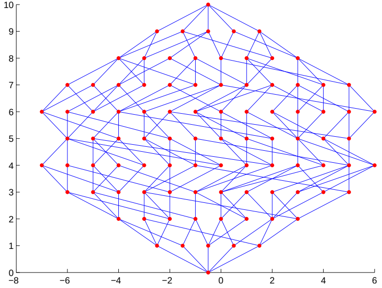
Граф связности множества алгоритмов A:
— вершины — алгоритмы,
— рёбра (a, a') — пары алгоритмов $a \prec a'$ в соседних слоях.

Пример. Граф расслоения и связности

Линейно разделимая выборка длины $L = 10$



Граф связности семейства линейных классификаторов



Оценка Q_ε через профиль расслоения–связности

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

Теорема

Пусть векторы ошибок всех алгоритмов $a \in A$ попарно различны, и в A есть корректный на $\mathbb{X}$ алгоритм. Тогда справедлива верхняя оценка вероятности переобучения

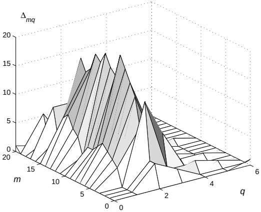
$$Q_\varepsilon \leqslant \sum_{m=\lceil \varepsilon k \rceil}^{L} \sum_{q=0}^{\ell} \Delta_{mq} \frac{C_{L-m-q}^{\ell-q}}{C_L^\ell}.$$

Как выглядят профили Δ_{mq}?

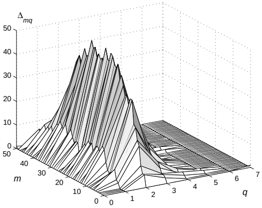
Примеры профилей расслоения-связности

$\mathbb{X}$ — линейно разделимые 2D-выборки; A — линейные классификаторы.

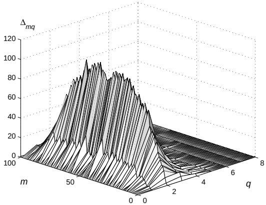
$L = 20$



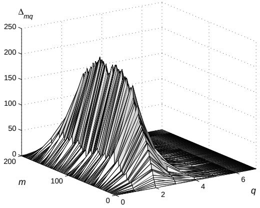
$L = 50$



$L = 100$



$L = 200$



Две эмпирические гипотезы

Гипотеза (о сепарабельности профилей)

$$\Delta_{mq} \lessapprox \Delta_m \lambda_q,$$

где Δ_m — *коэффициент разнообразия m-го слоя,*
λ_q — *доля алгоритмов m-го слоя со связностью q.*

$(\Delta_m,\ m = 0, \ldots, L)$ — *профиль расслоения.*
$(\lambda_q,\ q = 0 \ldots, L)$ — *профиль связности,* $\sum_{q=0}^{L} \lambda_q = 1$.

Гипотеза (о размерности пространства)

Положение пика в профиле λ_q приблизительно равно размерности пространства.

(для семейства линейных классификаторов это так)

Оценка вероятности переобучения

Теорема

В условиях предыдущей теоремы и гипотезы сепарабельности

$$Q_\varepsilon \leqslant \underbrace{\sum_{m=\lceil \varepsilon k \rceil}^{k} \Delta_m \frac{C_{L-m}^{\ell}}{C_L^{\ell}}}_{\text{VC-оценка}} \underbrace{\sum_{q=0}^{L} \lambda_q \left(\frac{\ell}{L-m} \right)^q}_{\text{поправка на связность}}.$$

При известных Δ_m, λ_q вычисления займут $O(L)$.

Задачи и открытые проблемы

1. Обобщить на случай $n(a_0, \mathbb{X}) > 0$.
2. Как оценивать профиль связности реальных семейств?
3. Когда справедлива гипотеза сепарабельности?
4. Когда справедлива гипотеза размерности?
5. Возможно ли оценить характеристики нижних слоёв графа в тех случаях, когда метод обучения строит некоторое подмножество алгоритмов в явном виде?

Задачи и открытые проблемы

www.MachineLearning.ru/wiki
Участник:Vokov

Страницы:

- Слабая вероятностная аксиоматика
- Расслоение и сходство алгоритмов (виртуальный семинар)
- Теория надёжности обучения по прецедентам (спецкурс каф. ММП ВМиК МГУ, К.В.Воронцов)

  Сборник задач и открытых проблем:

  www.MachineLearning.ru/wiki/images/6/6b/Voron09problems.pdf