Просеминар кафедры ММП, ВМК МГУ

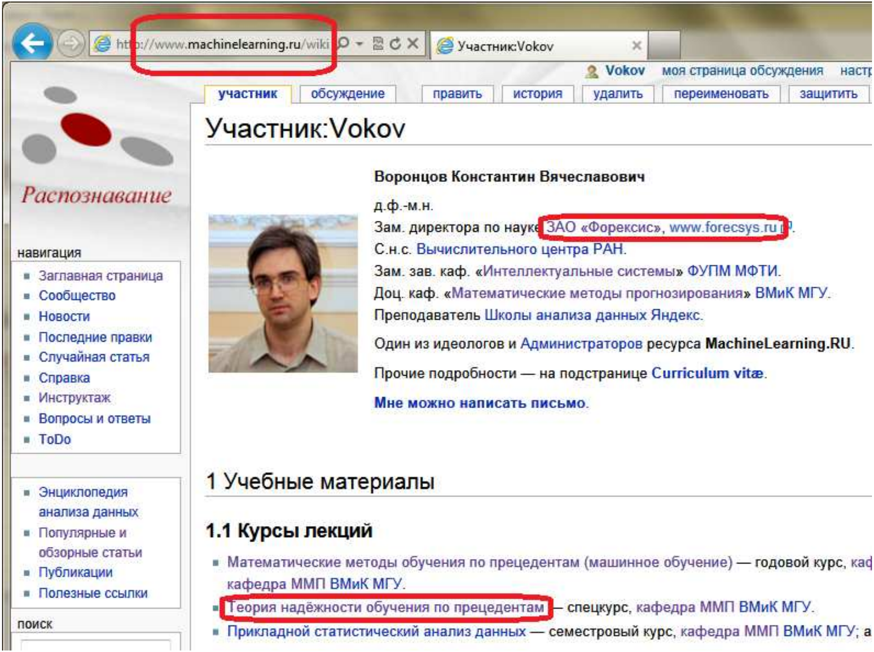
. . . .

Комбинаторная теория переобучения и её применения

Воронцов Константин Вячеславович

27 февраля 2012

Познакомимся



План семинара

1. **Обучение по прецедентам и проблема переобучения**
   - Задачи обучения по прецедентам
   - Методы обучения по прецедентам
   - Проблема переобучения

2. **Комбинаторная теория переобучения**
   - Основные понятия и классические оценки
   - Эксперименты
   - Комбинаторные оценки переобучения

3. **Результаты, открытые проблемы, планы**
   - Логические алгоритмы классификации
   - Отбор эталонов в методе ближайшего соседа
   - Открытые проблемы и планы

Основные определения и обозначения

$\mathbb{X}$ — объекты; $\mathbb{Y}$ — ответы (классы);
$y^*\colon \mathbb{X} \to \mathbb{Y}$ — неизвестная зависимость.

Дано: $x_i = (x_i^1, \ldots, x_i^n)$ — *обучающие объекты* с известными ответами $y_i = y^*(x)$, $i = 1, \ldots, \ell$:

$$\begin{pmatrix} x_1^1 & \ldots & x_1^n \\ \multicolumn{3}{c}{\ldots\ldots\ldots\ldots} \\ x_\ell^1 & \ldots & x_\ell^n \end{pmatrix} \xrightarrow{y^*} \begin{pmatrix} y_1 \\ \ldots \\ y_\ell \end{pmatrix}$$

Найти: алгоритм $a\colon \mathbb{X} \to \mathbb{Y}$, способный давать правильные ответы на новых объектах $\tilde{x}_i = (\tilde{x}_i^1, \ldots, \tilde{x}_i^n)$, $i = 1, \ldots, k$:

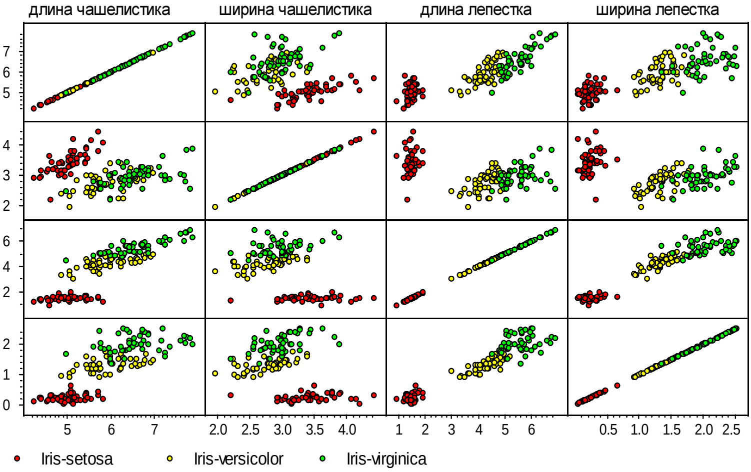
$$\begin{pmatrix} \tilde{x}_1^1 & \ldots & \tilde{x}_1^n \\ \multicolumn{3}{c}{\ldots\ldots\ldots\ldots} \\ \tilde{x}_k^1 & \ldots & \tilde{x}_k^n \end{pmatrix} \xrightarrow{a?} \begin{pmatrix} a(\tilde{x}_1) \\ \ldots \\ a(\tilde{x}_k) \end{pmatrix}$$

Примеры прикладных задач обучения по прецедентам

- Распознавание, классификация, принятие решений ($|\mathbb{Y}| < \infty$):
  - x — пациент; y — долгосрочный результат лечения;
  - x — заёмщик; y — кредит выдать / не выдать;
  - x — курсы акций; y — купить / продать.
  - x — абонент; y — уйдёт / не уйдёт к другому оператору;
  - x — фотопортрет; y — идентификатор личности;
  - x — фрагмент ДНК; y — функция: промотор / ген;
  - x — фрагмент белка; y — тип вторичной структуры;
  - x — текстовое сообщение; y — спам / не спам;
- Регрессия и прогнозирование ($\mathbb{Y} = \mathbb{R}$ или $\mathbb{R}^m$):
  - x — структура химического соединения; y — его свойство;
  - x — параметры технолог. процесса; y — свойство продукции;
  - x — история продаж; y — прогноз потребительского спроса;
  - x — данные о недвижимости; y — продажная стоимость;
  - x — пара $\langle$клиент, товар$\rangle$; y — рейтинг товара.

Пример: задача классификации цветков ириса [Фишер, 1936]

$n = 4$ признака, $|\mathbb{Y}| = 3$ класса, длина выборки $\ell = 150$.



Модель алгоритмов и метод обучения

Модель алгоритмов — параметрическое семейство отображений

$$A = \{g(x,\theta) \mid \theta \in \Theta\},$$

где $g\colon \mathbb{X} \times \Theta \to \mathbb{Y}$ — фиксированная функция,
Θ — множество допустимых значений параметра θ.

В задачах обучения по прецедентам выделяются два этапа:

1. *Метод обучения* $\mu\colon (\mathbb{X} \times \mathbb{Y})^\ell \to A$ по обучающей выборке $X = (x_i, y_i)_{i=1}^\ell$ выбирает из A алгоритм $a = \mu(X)$.
2. Найденный алгоритм a применяется для вычисления прогнозов $\tilde{y}_i = a(\tilde{x}_i)$ на новой выборке $\bar{X} = \{\tilde{x}_1, \dots, \tilde{x}_k\}$.

Примеры методов обучения

из курса «Математические методы распознавания образов»:

- Байесовский классификатор
- Метод ближайших соседей
- Метод потенциальных функций
- Метод опорных векторов
- Логистическая регрессия
- Многослойная нейронная сеть
- Сеть радиальных базисных функций
- Бустинг
- Решающее дерево
- Алгоритм вычисления оценок
- ... ...

Принцип минимизации эмпирического риска

Эмпирический риск — частота ошибок алгоритма a на X:

$$Q(a, X) = \frac{1}{\ell} \sum_{i=1}^{\ell} \bigl[a(x_i) \neq y_i\bigr].$$

Минимизация эмпирического риска — пример метода обучения:

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

Проблема обобщающей способности:

- будет ли алгоритм $a = \mu(X)$ приближать y^* на всём $\mathbb{X}$?
- найдём ли мы «закон природы» или *переобучимся*, т. е. подгоним функцию $g(x, \theta)$ под заданные точки (x_i, y_i)?
- будет ли $Q(a, \bar{X})$ мало́ на новых данных — *контрольной выборке* $\bar{X} = (\tilde{x}_i, \tilde{y}_i)_{i=1}^{k}$, $\tilde{y}_i = y^*(\tilde{x}_i)$?

Пример переобучения. Модельная задача регрессии

Зависимость $y^*(x) = \dfrac{1}{1+25x^2}$ на отрезке $x \in [-2, 2]$.

Признаковое описание $x \mapsto (1, x^1, x^2, \dots, x^n)$.

Алгоритм полиномиальной регрессии

$$a(x, \theta) = \theta_0 + \theta_1 x + \dots + \theta_n x^n \text{ — полином степени } n.$$

Обучение методом наименьших квадратов:

$$Q(\theta, X) = \sum_{i=1}^{\ell} (\theta_0 + \theta_1 x_i + \dots + \theta_n x_i^n - y_i)^2 \to \min_{\theta_0, \dots, \theta_n}.$$

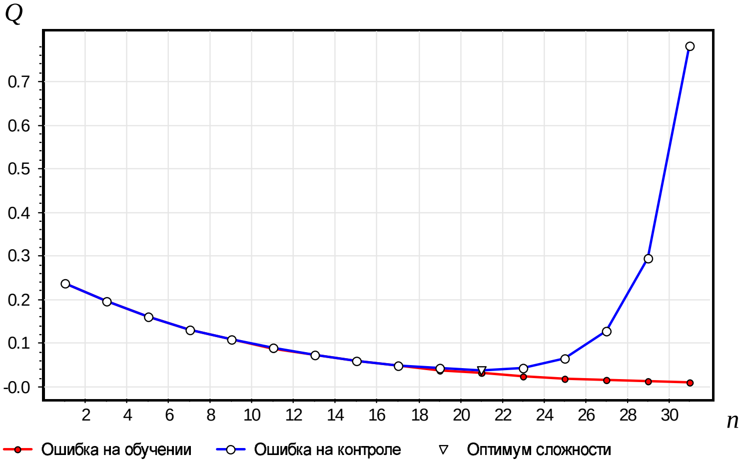
Обучающая выборка: $X = \left\{ x_i = 4\frac{i-1}{\ell-1} - 2 \;\middle|\; i = 1, \dots, \ell \right\}$.

Контрольная выборка: $\bar{X} = \left\{ \tilde{x}_i = 4\frac{i-0.5}{\ell-1} - 2 \;\middle|\; i = 1, \dots, \ell - 1 \right\}$.

Что происходит с $Q(\mu(X), X)$ и $Q(\mu(X), \bar{X})$ при увеличении n?

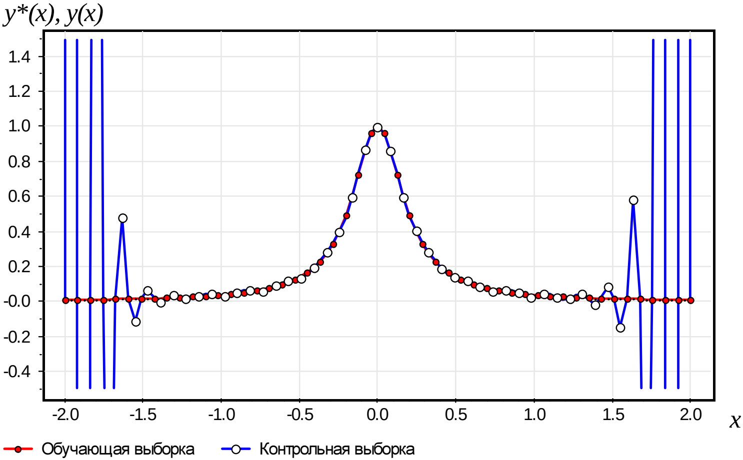
Пример переобучения: эксперимент при $\ell = 50$, $n = 1..31$

Переобучение — это когда $Q(\mu(X), \bar{X}) \gg Q(\mu(X), X)$:



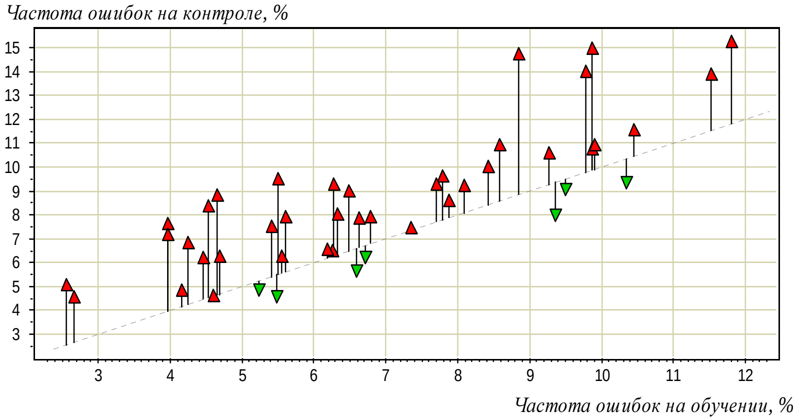
Пример переобучения: эксперимент при $\ell = 50$

Переобучение, «вид изнутри»: что происходит с полиномами слишком высоких степеней (в данном случае $n = 40$)



Пример переобучения. Реальная задача классификации

Задача предсказания отдалённого результата хирургического лечения атеросклероза, $L = 98$. Точки — различные алгоритмы.



Матрица ошибок

$\mathbb{X} = \{x_1, \dots, x_L\}$ — конечное *генеральное множество* объектов;
$A = \{a_1, \dots, a_D\}$ — 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
$\dots$	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
$\dots$	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 $X \subset \mathbb{X}$;

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

Вероятностные определения обобщающей способности

Основная вероятностная аксиома

Все разбиения $X \sqcup \bar{X} = \mathbb{X}$ равновероятны, $|X| = \ell$, $|\bar{X}| = k$.

В этом случае $\mathsf{P} \equiv \mathsf{E} \equiv \frac{1}{C_L^\ell} \sum\limits_{X \subset \mathbb{X}}$ — доля разбиений выборки.

Функционалы обобщающей способности

- ожидаемая частота ошибок на контроле:
$$\mathrm{CCV}(\mu, \mathbb{X}) = \mathsf{E}\, \nu\bigl(\mu(X), \bar{X}\bigr).$$
- вероятность большой частоты ошибок на контроле:
$$R_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\bigl[\nu\bigl(\mu(X), \bar{X}\bigr) \geqslant \varepsilon\bigr].$$
- вероятность переобучения:
$$Q_\varepsilon(\mu, \mathbb{X}) = \mathsf{P}\bigl[\nu\bigl(\mu(X), \bar{X}\bigr) - \nu\bigl(\mu(X), X\bigr) \geqslant \varepsilon\bigr].$$

Теория Вапника–Червоненкиса

Теорема (Вапник, Червоненкис, 1974)

Для любых $\mathbb{X}$, A, μ и $\varepsilon \in [0,1]$, при $\ell = k$

$$Q_\varepsilon(\mu, \mathbb{X}) \leqslant |A| \cdot \tfrac{3}{2} \exp\bigl(-\varepsilon^2 \ell\bigr).$$

Проблема завышенности:

- эта оценка завышена в 10^8–10^{11} раз;
- что приводит к оценкам длины обучения $\ell = 10^6$–10^{10}, когда на самом деле достаточно $\ell = 10^2$–10^3.

Причина завышенности — это оценка «худшего случая»:

- она зависит только от размеров матрицы ошибок $L \times D$;
- не зависит от её содержимого $I(a, x)$, выборки $\mathbb{X}$, метода μ.

Два мысленных эксперимента

1. Пусть в семействе есть один очень хороший алгоритм, $n(a_0, \mathbb{X}) = 0$, и много плохих алгоритмов a: $n(a, \mathbb{X}) \gg 0$. Тогда a_0 почти всегда будет лучшим и на обучающей выборке.

Результат: можно полагать $|A| \approx 1$.

В общем случае надо учитывать *расслоение* семейства A, $A_m = \bigl\{a \in A \colon n(a, \mathbb{X}) = m\bigr\}$, наиболее важны нижние слои.

2. Пусть в семействе есть алгоритм a_0, и все остальные очень похожи на него. Тогда это «почти один и тот же алгоритм».

Результат: можно полагать $|A| \approx 1$.

В общем случае надо учитывать *связность* семейства A, сколько в A вместе с каждым a содержится b: $\|a - b\| = 1$.

Эксперименты с модельными семействами алгоритмов

Физика — экспериментальная, естественная наука, часть естествознания. Математика — это та часть физики, в которой эксперименты дёшевы. [В.И.Арнольд]

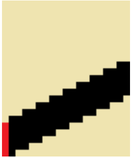
Хотим экспериментально проверить гипотезу о влиянии расслоения и связности на вероятность переобучения.

1. Будем изучать *модельные семейства алгоритмов*, задавая их непосредственно своими матрицами ошибок.
2. Будем оценивать вероятность методом Монте–Карло — как долю разбиений выборки из случайного подмножества N разбиений, $|N|$ порядка 10^3–10^4:

$$\hat{Q}_\varepsilon(\mu, \mathbb{X}) = \frac{1}{|N|} \sum_{(\bar{X}, X) \in N} \Bigl[\nu\bigl(\mu(X), \bar{X}\bigr) - \nu\bigl(\mu(X), X\bigr) \geqslant \varepsilon\Bigr].$$

Эксперимент с четырьмя модельными семействами

Матрицы ошибок: строки — объекты, столбцы — алгоритмы; лучший алгоритм одинаков во всех четырёх семействах.

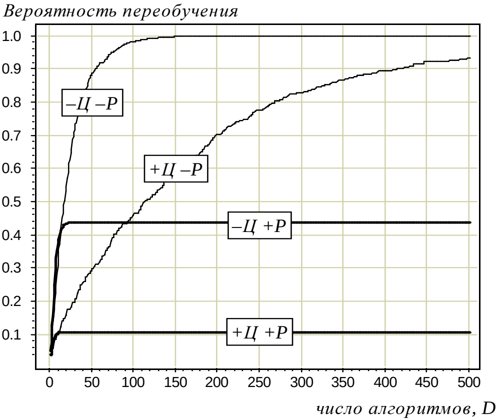
	есть расслоение по числу ошибок	**нет расслоения** по числу ошибок
есть связность, соседние алгоритмы отличаются на одном объекте, образуется *цепь*		
нет связности, соседние алгоритмы существенно различны, *цепь* не образуется		

Результаты эксперимента (при $\ell = k = 100$, $\varepsilon = 0.05$, $|N| = 10^4$)

Условные обозначения:
+Ц — цепь;
−Ц — не цепь;
+Р — с расслоением;
−Р — без расслоения;

Связность замедляет темп роста $Q_\varepsilon(D)$

Расслоение понижает уровень горизонтальной асимптоты $Q_\varepsilon(D)$



Вывод: получение точных оценок вероятности переобучения невозможно без учёта эффектов расслоения и связности.

Граф расслоения–связности множества алгоритмов

Определим бинарные отношения на множестве алгоритмов A:
частичный порядок $a \leqslant b$: $I(a,x) \leqslant I(b,x)$ для всех $x \in \mathbb{X}$;
предшествование $a \prec b$: $a \leqslant b$ и $\|b - a\| = 1$.

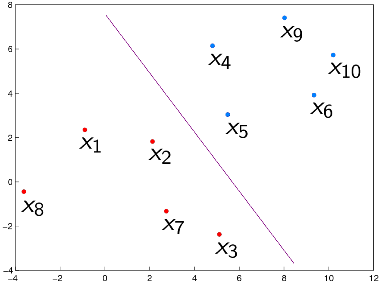
Определение

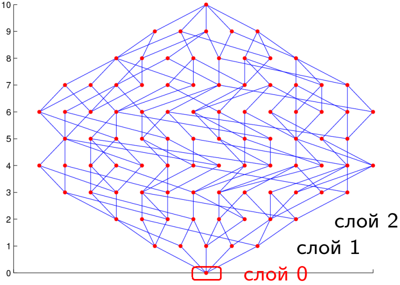
Граф расслоения–связности $\langle A, E \rangle$:
A — множество попарно различных векторов ошибок;
$E = \bigl\{(a,b)\colon a \prec b\bigr\}$.

Свойства графа расслоения–связности:

- это подграф графа Хáссе отношения порядка $\leqslant$ на A;
- каждому ребру (a,b) соответствует объект $x_{ab} \in \mathbb{X}$, такой, что $I(a, x_{ab}) = 0, \quad I(b, x_{ab}) = 1$;
- граф является многодольным со слоями $A_m = \bigl\{a \in A\colon n(a,\mathbb{X}) = m\bigr\}, \quad m = 0, \dots, L$;

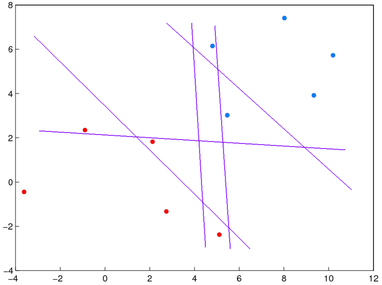
Пример. Семейство линейных алгоритмов классификации

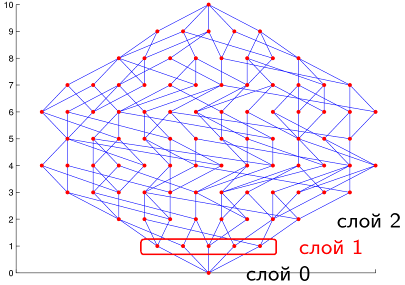




	слой 0
x_1	0
x_2	0
x_3	0
x_4	0
x_5	0
x_6	0
x_7	0
x_8	0
x_9	0
x_{10}	0

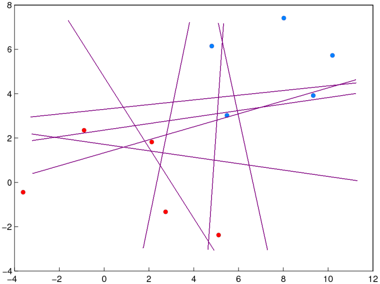
Пример. Семейство линейных алгоритмов классификации

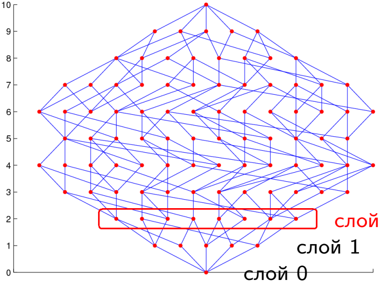




	слой 0	слой 1				
x_1	0	1	0	0	0	0
x_2	0	0	1	0	0	0
x_3	0	0	0	1	0	0
x_4	0	0	0	0	1	0
x_5	0	0	0	0	0	1
x_6	0	0	0	0	0	0
x_7	0	0	0	0	0	0
x_8	0	0	0	0	0	0
x_9	0	0	0	0	0	0
x_{10}	0	0	0	0	0	0

Пример. Семейство линейных алгоритмов классификации





	слой 0	слой 1					слой 2								
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	...

Характеристики расслоения и связности алгоритма $a \in A$

Определение

Верхняя связность $u(a)$ алгоритма a — это число всех рёбер, исходящих из вершины a:

$$u(a) = |X_a|, \quad X_a = \bigl\{x_{ab} \in \mathbb{X} \bigm| a \prec b\bigr\};$$

X_a называется *порождающим множеством* алгоритма a.

Определение

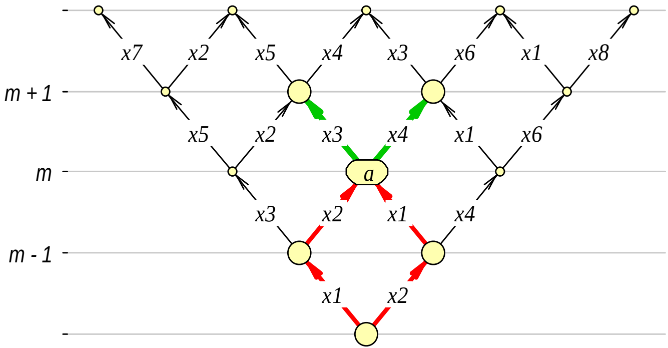
Неполноценность $q(a)$ алгоритма a — это мощность множества объектов, соответствующих всем рёбрам на путях, ведущих в a:

$$q(a) = |X'_a|, \quad X'_a = \bigl\{x \in \mathbb{X} \bigm| \exists b \in A\colon\ b < a,\ I(b,x) < I(a,x)\bigr\};$$

X'_a называется *запрещающим множеством* алгоритма a.

Пример: двумерная сеть алгоритмов

Верхняя связность алгоритма a: $X_a = \{x3, x4\}$, $u(a) = |X_a| = 2$;
Неполноценность алгоритма a: $X'_a = \{x1, x2\}$, $q(a) = |X'_a| = 2$;



Основная лемма: если $\mu X = a$, то $X_a \subseteq X$ и $X'_a \subseteq \bar{X}$.

Верхняя оценка вероятности переобучения

Теорема (Воронцов, Решетняк, Ивахненко, 2010)

Для любого монотонного метода μ, любых $\mathbb{X}$, A и $\varepsilon \in (0,1)$

$$Q_\varepsilon(\mu, \mathbb{X}) \leqslant \sum_{a \in A} \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell} \mathcal{H}_{L-u-q}^{\ell-u,\, m-q}\left(\tfrac{\ell}{L}(m - \varepsilon k)\right),$$

где $u = |X_a|$ — верхняя связность алгоритма a,
$q = |X'_a|$ — неполноценность алгоритма a,
$m = n(a, \mathbb{X})$ — число ошибок алгоритма a,

$$\mathcal{H}_L^{\ell,\, m}(z) = \sum_{s=0}^{\lfloor z \rfloor} \frac{C_m^s C_{L-m}^{\ell-s}}{C_L^\ell}, \quad z = 0, \dots, \ell$$

— функция гипергеометрического распределения:

Следствие: $\mathsf{P}\left[\mu X = a\right] \leqslant C_{L-u-q}^{\ell-u} / C_L^\ell$.

Идея доказательства

1. Пусть μ — произвольный монотонный метод обучения, $\bar{\mu}$ — монотонный пессимистичный метод обучения. Тогда

$$Q_\varepsilon(\mu, \mathbb{X}) \leqslant Q_\varepsilon(\bar{\mu}, \mathbb{X}).$$

2. Если $\bar{\mu}(X) = a$, то $\begin{cases} X_a \subseteq X \text{ в силу пессимистичности } \bar{\mu}, \\ X'_a \subseteq \bar{X} \text{ в силу монотонности } \bar{\mu}. \end{cases}$

3. $$\mathsf{P}\Bigl[\bar{\mu}(X) = a\Bigr] \leqslant \mathsf{P}\bigl[\underbrace{X_a \subseteq X \text{ и } X'_a \subseteq \bar{X}}_{S(a,X)}\bigr] = \frac{C_{L-|X_a|-|X'_a|}^{\ell-|X_a|}}{C_L^\ell} = \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell}.$$

4. По формуле полной вероятности:

$$Q_\varepsilon(\bar{\mu}, \mathbb{X}) = \sum_{a \in A} \underbrace{\mathsf{P}\Bigl[S(a,X)\Bigr]}_{C_{L-u-q}^{\ell-u} / C_L^\ell} \cdot \underbrace{\mathsf{P}\Bigl[\delta(a,X) \geqslant \varepsilon \Bigm| S(a,X)\Bigr]}_{\mathcal{H}_{L-u-q}^{\ell-u,\, m-q}\left(\frac{\ell}{L}(m - \varepsilon k)\right)}. \qquad \blacksquare$$

Свойства оценки $Q_\varepsilon \leqslant \sum\limits_{a\in A} \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell} \mathcal{H}_{L-u-q}^{\ell-u,\,m-q}\left(\frac{\ell}{L}(m-\varepsilon k)\right)$

1. Вклад алгоритма $a \in A$ убывает экспоненциально по $u(a) \Rightarrow$ **связные семейства меньше переобучаются**; по $q(a) \Rightarrow$ **только нижние слои вносят вклад в Q_ε.**
2. При $q = u = 0$ и $\ell = k$ это оценка Вапника-Червоненкиса:
$$Q_\varepsilon \leqslant \sum_{a\in A} \mathcal{H}_L^{\ell,\,m}\left(\tfrac{\ell}{L}(m-\varepsilon k)\right) \leqslant |A| \cdot \tfrac{3}{2}\exp(-\varepsilon\ell^2).$$
3. При $|A| = 1$ это аналог закона больших чисел (утверждается сходимость частот в двух подвыборках):
$$\nu(a,\bar{X}) \xrightarrow{P} \nu(a,X) \quad \text{при} \quad \ell, k \to \infty.$$
4. Оценка обращается в равенство в случае многомерных монотонных сетей алгоритмов [Павел Ботов]
5. Получен критерий точности оценки [Никита Животовский]

Верхние оценки средней частоты ошибок на контроле

Теорема

Для любого монотонного метода μ, любых $\mathbb{X}$ и A

$$\mathrm{CCV}(\mu, \mathbb{X}) \leqslant \sum_{a \in A} \frac{C_{L-u-q}^{\ell-u}}{C_L^\ell} \left(\frac{m}{k} - \frac{(m-q)(\ell-u)}{k(L-u-q)} \right).$$

где $u = |X_a|$ — верхняя связность алгоритма a,
$q = |X'_a|$ — неполноценность алгоритма a,
$m = n(a, \mathbb{X})$ — число ошибок алгоритма a.

Преимущество:
оценка CCV вычисляется намного проще, чем оценки Q_ε и R_ε.

Основные результаты в комбинаторной теории переобучения

1. Оценки Q_ε и новые критерии отбора признаков для пороговых логических закономерностей [Андрей Ивахненко]
2. Точные оценки Q_ε и CCV для многомерных сетей алгоритмов и новые методы обучения деревьев решений [Павел Ботов]
3. Точные оценки CCV для метода k ближайших соседей и новые методы отбора эталонных объектов [Максим Иванов, Анастасия Зухба]
4. Верхние оценки CCV для монотонных классификаторов и новые методы обучения композиций классификаторов [Иван Гуз, Галина Махина]
5. Оценки Q_ε для рандомизированных методов обучения и симметричных семейства на основе теории групп [Александр Фрей, Илья Толстихин]

Понятие логической закономерности

Закономерность класса y — это предикат $r\colon \mathbb{X} \to \{0,1\}$, который *выделяет* ($r(x) = 1$) много объектов класса y:

$$p(r, X) = \sum_{x_i \in X} r(x_i)\bigl[y_i = y\bigr] \;\to\; \max,$$

и как можно меньше объектов всех остальных классов:

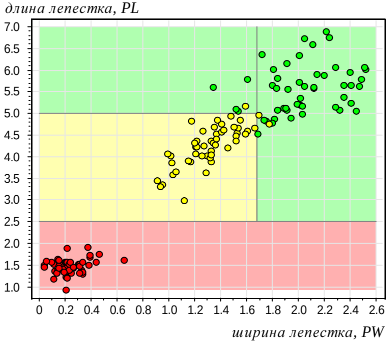
$$n(r, X) = \sum_{x_i \in X} r(x_i)\bigl[y_i \neq y\bigr] \;\to\; \min\,.$$

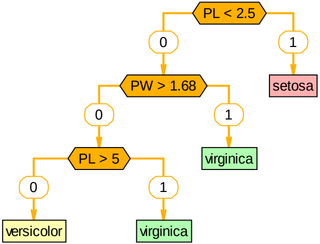
Логическая закономерность — конъюнкция пороговых условий:

$$r(x) = \bigwedge_{j \in J} \bigl[f_j(x) \leqslant \theta_j\bigr],$$

где $f_j(x)$ — числовые признаки, θ_j — пороги, $j = 1, \ldots, n$;
$J \subseteq \{1, \ldots, n\}$ — подмножество признаков, обычно $|J| = 1..5$.

Пример 1. Закономерности в задаче с ирисами Фишера

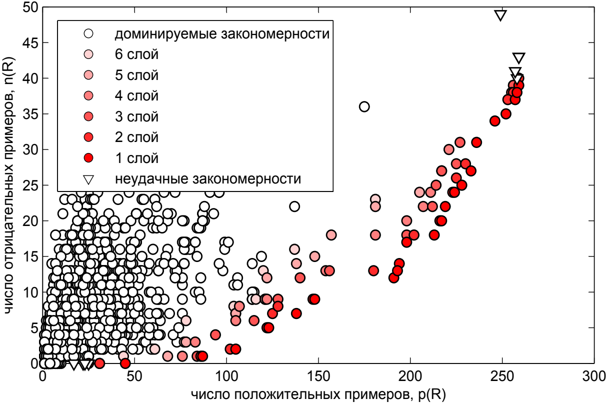




setosa	$r_1(x) = [PL \leqslant 2.5]$
virginica	$r_2(x) = [PL > 2.5] \wedge [PW > 1.68]$
virginica	$r_3(x) = [PL > 5] \wedge [PW \leqslant 1.68]$
versicolor	$r_4(x) = [PL > 2.5] \wedge [PL \leqslant 5] \wedge [PW < 1.68]$

Пример 2. Для классификатора нужно много закономерностей

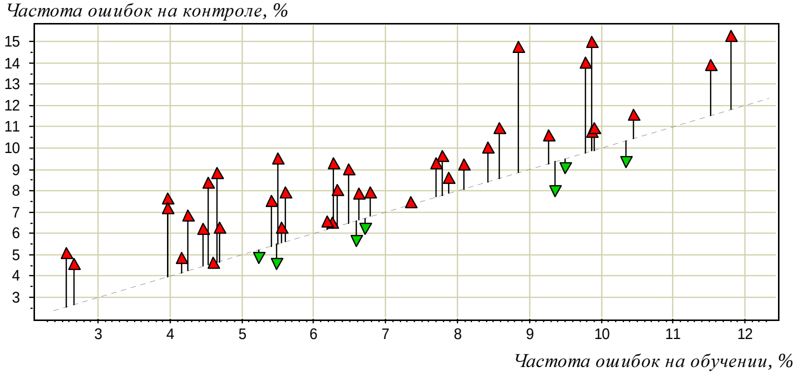
Парето-фронт — множество недоминируемых закономерностей (точка (p, n) недоминируема, если правее и ниже точек нет)



Задача кредитного скоринга german из репозитория UCI.

Пример 3. Проблема переобучения

Как отбросить переобученные закономерности на этапе обучения?



Задача предсказания отдалённого результата хирургического лечения атеросклероза. Точки — найденные закономерности.

Модификация критериев (p, n) с поправкой на переобучение

1. Вычислить оценки расслоения–связности как функции ε:

$$\mathsf{P}\Bigl[\tfrac{1}{k}n(r,\bar X) - \tfrac{1}{\ell}n(r,X) \geqslant \varepsilon\Bigr] \leqslant \eta_n(\varepsilon);$$

$$\mathsf{P}\Bigl[\tfrac{1}{\ell}p(r,X) - \tfrac{1}{k}p(r,\bar X) \geqslant \varepsilon\Bigr] \leqslant \eta_p(\varepsilon);$$

2. Обращение оценок: с вероятностью $1-\eta$

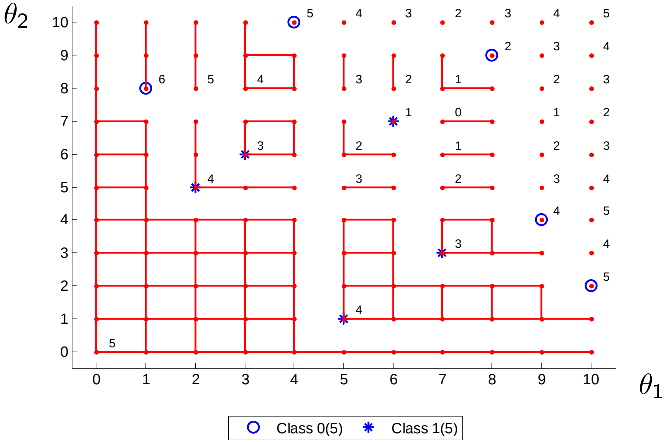
$$n(r,\bar X) \leqslant \underbrace{\tfrac{k}{\ell}n(r,X) + k\varepsilon_n(\eta)}_{\hat n(r,X)};$$

$$p(r,\bar X) \geqslant \underbrace{\tfrac{k}{\ell}p(r,X) - k\varepsilon_p(\eta)}_{\hat p(r,X)}.$$

3. Для поиска закономерностей вместо $(p \to \max, n \to \min)$ использовать модифицированный критерий $(\hat p \to \max, \hat n \to \min)$.

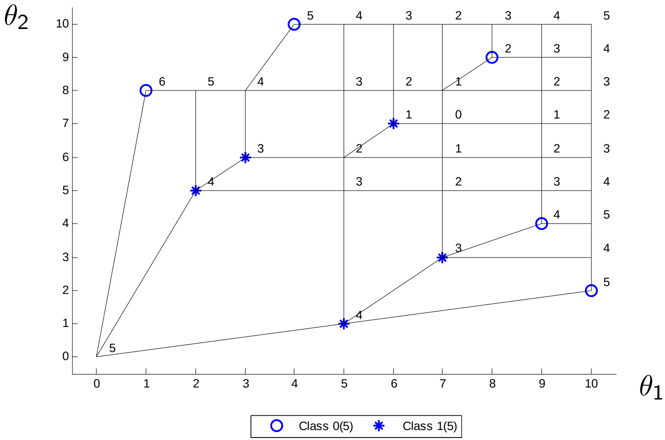
Множество конъюнктивных закономерностей

Пример: разделимая 2-мерная выборка, $L = 10$, два класса.
закономерности: $r(x) = \bigl[f_1(x) \leqslant \theta_1\bigr] \wedge \bigl[f_2(x) \leqslant \theta_2\bigr]$.



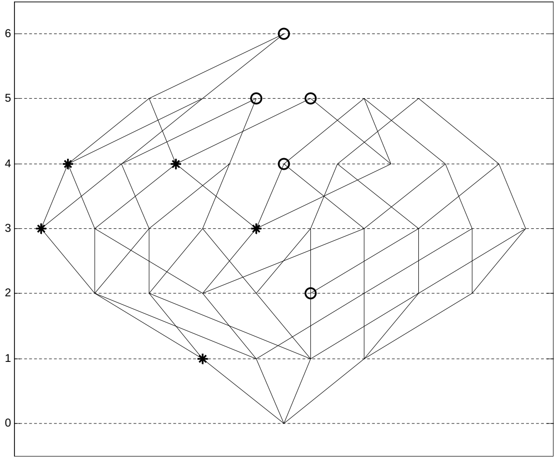
Классы эквивалентности закономерностей

Пример: разделимая 2-мерная выборка, $L = 10$, два класса.
закономерности: $r(x) = \bigl[f_1(x) \leqslant \theta_1\bigr] \wedge \bigl[f_2(x) \leqslant \theta_2\bigr]$.



Классы эквивалентности закономерностей

Пример: граф расслоения–связности, изоморфный графу классов эквивалентности с предыдущего слайда.



Эксперимент на реальных данных

Реальные задачи классификации из репозитория UCI:

задачи	объектов	признаков
australian	690	14
echo cardiogram	74	10
heart disease	294	13
hepatitis	155	19
labor relations	40	16
liver	345	6

Методы обучения композиций логических закономерностей:

- WV (weighted voting) — взвешенное голосование;
- DL (decision list) — решающий список.

Методика тестирования: 10-кратный скользящий контроль.

Результаты эксперимента на реальных данных

методы	задачи					
	austr	echo	heart	hepa	labor	liver
RIPPER-opt	15.5	2.97	19.7	20.7	18.0	32.7
RIPPER+opt	15.2	5.53	20.1	23.2	18.0	31.3
C4.5 (Tree)	14.2	5.51	20.8	18.8	14.7	37.7
C4.5 (Rules)	15.5	6.87	20.0	18.8	14.7	37.5
C5.0	14.0	4.30	21.8	20.1	18.4	31.9
SLIPPER	15.7	4.34	19.4	17.4	12.3	32.2
LR	14.8	4.30	19.9	18.8	14.2	32.0
WV	14.9	4.37	20.1	19.0	14.0	32.3
DL	15.1	4.51	20.5	19.5	14.7	35.8
WV модиф.	14.1	3.2	19.3	18.1	13.4	30.2
DL модиф.	14.4	3.6	19.5	18.6	13.6	32.3

По каждой задаче выделено два лучших результата.

Классификатор ближайшего соседа (NN, nearest neighbor)

Пусть $\rho(x, x')$ — функция расстояния на множестве $\mathbb{X}$.

$$\mu\colon X \mapsto a, \qquad a(x) = y^*\Bigl(\arg\min_{x' \in X} \rho(x, x')\Bigr).$$

Определение (*профиль компактности* выборки $\mathbb{X}$)

доля объектов x_i, у которых m-й сосед x_{im} — в другом классе:

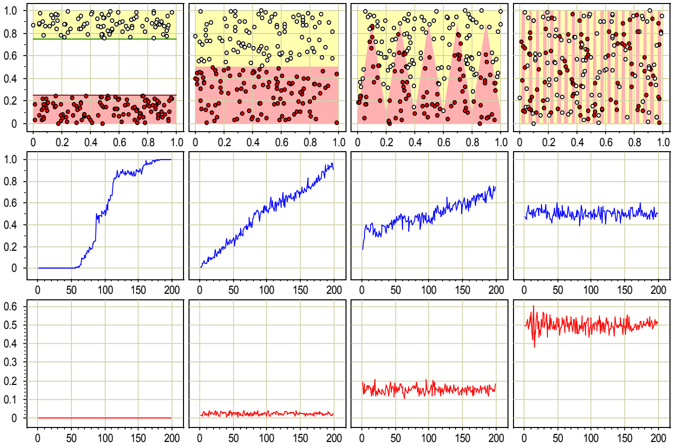
$$K(m, \mathbb{X}) = \frac{1}{L} \sum_{i=1}^{L} \bigl[y^*(x_{im}) \neq y_i\bigr]; \qquad m = 1, \ldots, L-1,$$

Теорема (точная оценка для метода ближайшего соседа)

$$\mathrm{CCV}(\mu, \mathbb{X}) = \sum_{m=1}^{k} K(m, \mathbb{X}) \frac{C_{L-1-m}^{\ell-1}}{C_{L-1}^{\ell}}.$$

Профили компактности для серии модельных задач

средний ряд: профили компактности,
нижний ряд: зависимость CCV от длины контроля $k = |\bar{X}|$.



Задача отбора множества эталонов $\Omega \subseteq \mathbb{X}$

Модификация NN $\quad \mu_{\Omega} \colon X \mapsto a, \quad a(x) = y\bigl(\arg\min\limits_{x' \in \Omega} \rho(x, x')\bigr).$

Определение (*профиль компактности* относительно Ω)

$$K(m, \Omega) = \frac{1}{L} \sum_{i=1}^{L} \bigl[y(x_i) \neq y(x_{im}^{\Omega})\bigr]; \quad m = 1, \ldots, |\Omega|.$$

где x_{im}^{Ω} — m-й сосед объекта x_i из множества Ω;

Теорема (точное выражение CCV относительно Ω)

$$\mathrm{CCV}(\mu_{\Omega}, \mathbb{X}) = \sum_{i=1}^{L} \underbrace{\sum_{m=1}^{k} \bigl[y(x_i) \neq y(x_{im}^{\Omega})\bigr] \frac{C_{L-1-m}^{\ell-1}}{L C_{L-1}^{\ell}}}_{T(x_i, \Omega) \text{ — вклад объекта } x_i \text{ в CCV}}.$$

Жадные алгоритмы отбора эталонов (prototype selection)

Задача: найти Ω: $\mathrm{CCV}(\mu_\Omega, \mathbb{X}) \to \min$.

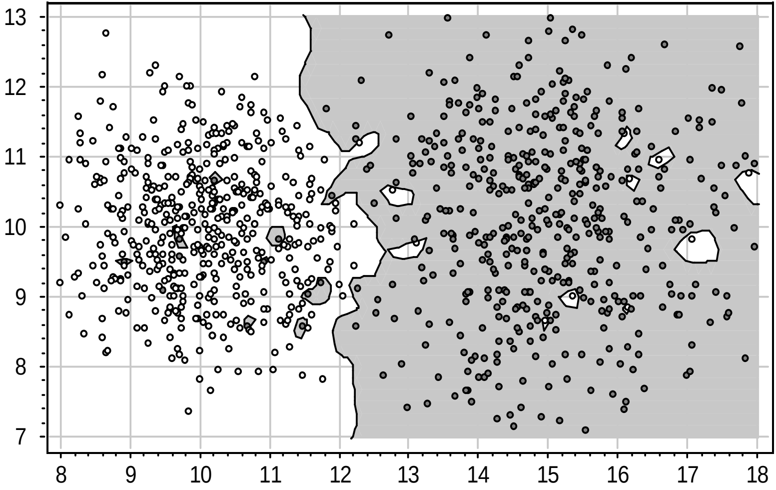
Жадный алгоритм удаления не-эталонов

1: $\Omega := \mathbb{X}$;
2: **повторять**
3: найти $x \in \Omega$: $\mathrm{CCV}(\mu_{\Omega\setminus\{x\}}, \mathbb{X}) \to \min$;
4: $\Omega := \Omega\setminus\{x\}$;
5: **пока** CCV уменьшается или почти не увеличивается;

Жадный алгоритм добавления эталонов

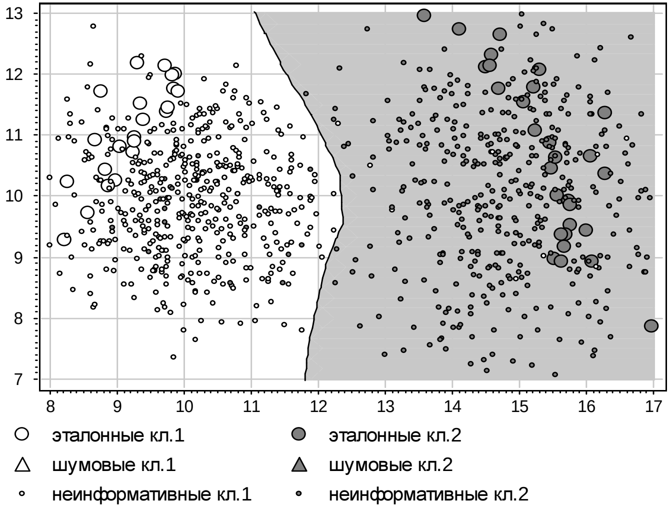
1: $\Omega := \{$по одному объекту от каждого класса$\}$;
2: **повторять**
3: найти $x \in \mathbb{X}\setminus\Omega$: $\mathrm{CCV}(\mu_{\Omega\cup\{x\}}, \mathbb{X}) \to \min$;
4: $\Omega := \Omega \cup \{x\}$;
5: **пока** CCV уменьшается;

Модельные данные

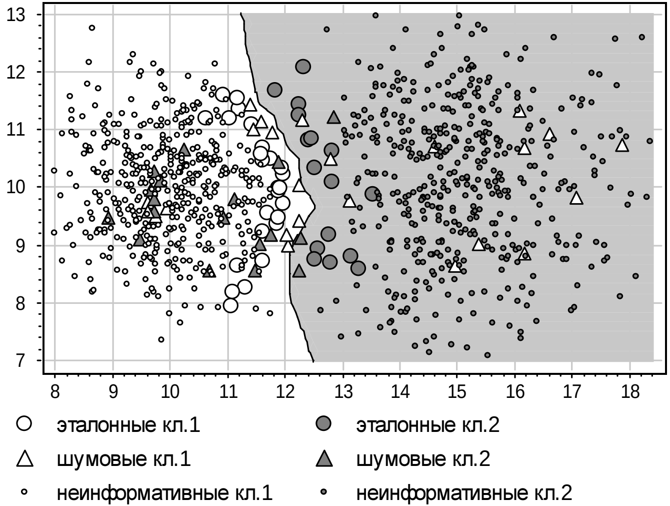


Модельная задача классификации: 1000 объектов, метод NN.

Жадное добавление эталонных объектов

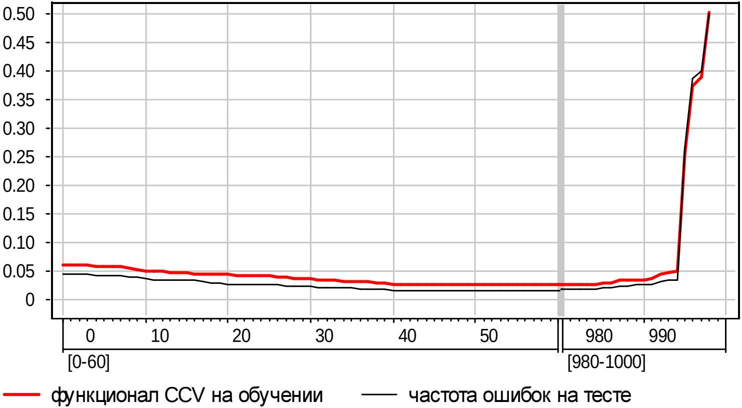


Жадное удаление не-эталонных объектов



Жадное удаление не-эталонных объектов

Зависимость CCV от числа удалённых неэталонных объектов.



Чудо: при отборе эталонов переобучения вообще нет!

Открытые проблемы

1. **Обосновать**, что оценку расслоения–связности можно вычислять не по генеральному множеству $\mathbb{X}$, а по случайной наблюдаемой подвыборке X.
2. **Найти** способ быстрого пересчёта оценок при добавлении в выборку ещё одного объекта, ещё одного признака.
3. **Уточнить** оценки расслоения–связности с учётом конкуренции между алгоритмами с хэмминговым расстоянием, большим 1.
4. **Обобщить** оценки расслоения–связности на случай небинарных функций потерь.
5. **Совершенствовать** методы обучения с помощью комбинаторных оценок обобщающей способности.

Следующий этап — переход от теории к технологии.

Пробные задачи

Задача 1

Для семейства $A = \{a_1, a_2\}$ известны 4 параметра:

$$m_{uv} = \#\{x \in \mathbb{X}\colon I(a_1, x) = u,\ I(a_2, x) = v\}, \quad u, v \in \{0, 1\}.$$

Найти Q_ε и CCV для минимизации эмпирического риска μ.

Задача 2

$X_1, \ldots, X_n$ — множества мощности $H_1, \ldots, H_n$ соответственно.
Найти число способов выбрать подмножество $U \subseteq \bigcup_{i=1}^n X_i$ мощности j так, чтобы $|U \cap X_i| \geqslant 1$ для всех $i = 1, \ldots, n$,

1) если $X_1, \ldots, X_n$ не пересекаются;
2) если $X_1, \ldots, X_n$ могут пересекаться.

Как быстрее всего вычислить это число?

Спецкурс ТНОП

«Теория надёжности обучения по прецедентам»
по понедельникам, 18:00, ауд. 615 (ВМК МГУ)

Страницы на www.MachineLearning.ru:

Теория надёжности обучения по прецедентам (курс лекций, К. В. Воронцов)

> Учебное пособие по курсу ТНОП: Voron-2011-tnop.pdf, 3 МБ (черновая версия).

Расслоение и сходство алгоритмов (виртуальный семинар)

Слабая вероятностная аксиоматика

Участник:Vokov