Дискретные структуры

МФТИ, весна 2014

Александр Дайняк

www.dainiak.com

Проекции множеств

Пусть $\mathcal{F}$ — непустое семейство различных подмножеств некоторого множества S.

Пусть $A \subseteq S$.

Проекция семейства $\mathcal{F}$ на множество A — это семейство

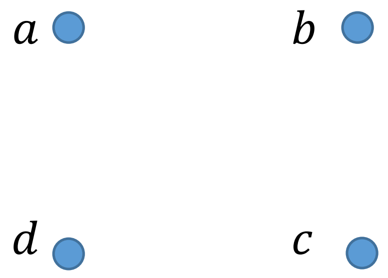
$$\{A \cap X \mid X \in \mathcal{F}\}.$$

Будем обозначать проекцию так: $\mathcal{F}|_A$.

Неформально: мы «фотографируем» A с помощью «объективов» из $\mathcal{F}$, и $\mathcal{F}|_A$ — это всевозможные изображения, которые мы при этом можем увидеть.

Примеры проекций множеств

- Если $A \coloneqq \{1,2,3\}$ и $F \coloneqq \{\emptyset, \{1\}, \{1,2\}, \{3,4\}\}$, то $\mathcal{F}|_A = \{\emptyset, \{1\}, \{1,2\}, \{3\}\}$.
- Если $A \coloneqq \{a, b, c, d\}$ (см. рисунок) и $\mathcal{F} \coloneqq \{\text{открытые полуплоскости в } \mathbb{R}^2\}$, то $\mathcal{F}|_A = 2^A \setminus \{\{a, c\}, \{b, d\}\}$:



Свойства проекций множеств (упражнения)

Если $\mathcal{F}$ — семейство подмножеств S, то

- $\mathcal{F}|_\emptyset = \{\emptyset\}$
- $\mathcal{F}|_S = \mathcal{F}$

Простые свойства:

- $(\mathcal{F}' \cup \mathcal{F}'')|_A = (\mathcal{F}'|_A) \cup (\mathcal{F}''|_A)$
- $(\mathcal{F}' \cap \mathcal{F}'')|_A \subseteq (\mathcal{F}'|_A) \cap (\mathcal{F}''|_A)$
- $(\mathcal{F}|_A)|_B = \mathcal{F}|_{A\cap B}$

Дробление

Будем говорить, что множество A *дробится* семейством $\mathcal{F}$, если $\mathcal{F}|_A = 2^A$.

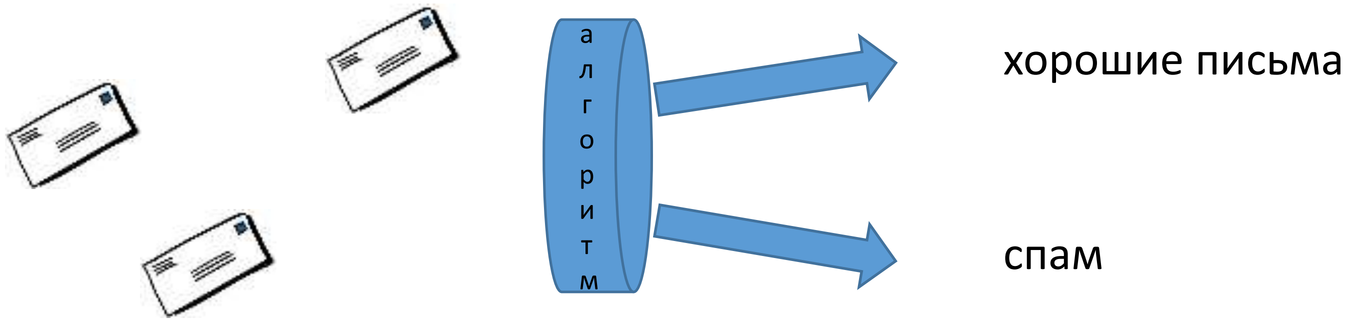
Примеры:

- Множество $\{1,2\}$ дробится семейством $\big\{\{1,3\},\{2,3\},\{1,2\},\{3,4\}\big\}$, но *не* дробится семейством $\big\{\{1,2,3\},\{1\},\{2\}\big\}$.
- Любое множество A дробится семейством $\mathcal{F}$, таким, что $2^A \subseteq \mathcal{F}$.
- Множество из четырёх точек *не* дробится семейством полуплоскостей в $\mathbb{R}^2$.

Связь с задачей классификации

Есть множество объектов $\mathcal{O}$, каждый из которых принадлежит одному из двух классов (надёжные/ненадёжные заёмщики, полезные письма/спам, и т.д.)

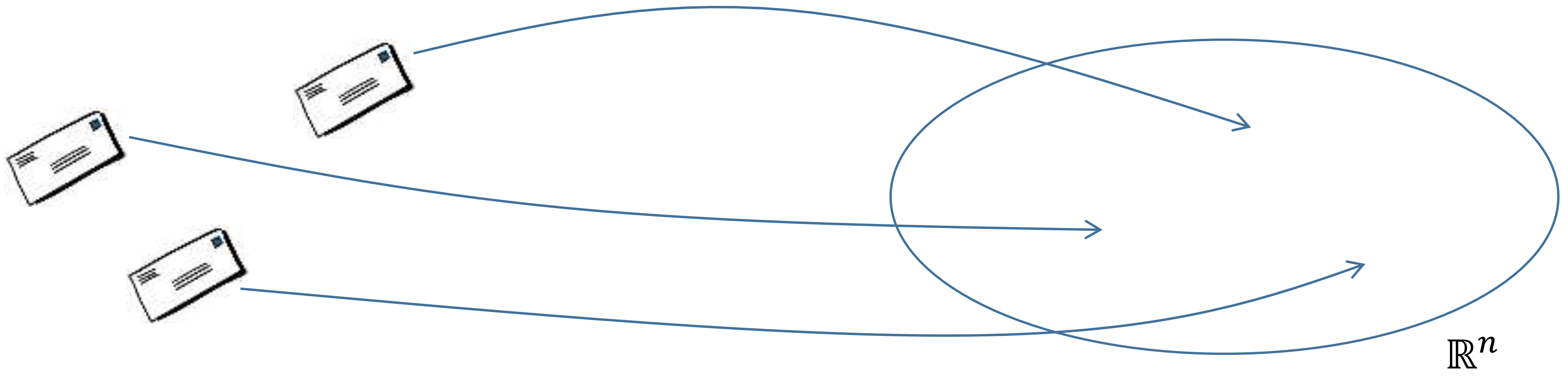
Хочется построить алгоритм классификации объектов, который бы для любого $o \in \mathcal{O}$ выдавал номер класса $\mathrm{class}(o)$, которому принадлежит o.



Связь с задачей классификации

Подход:

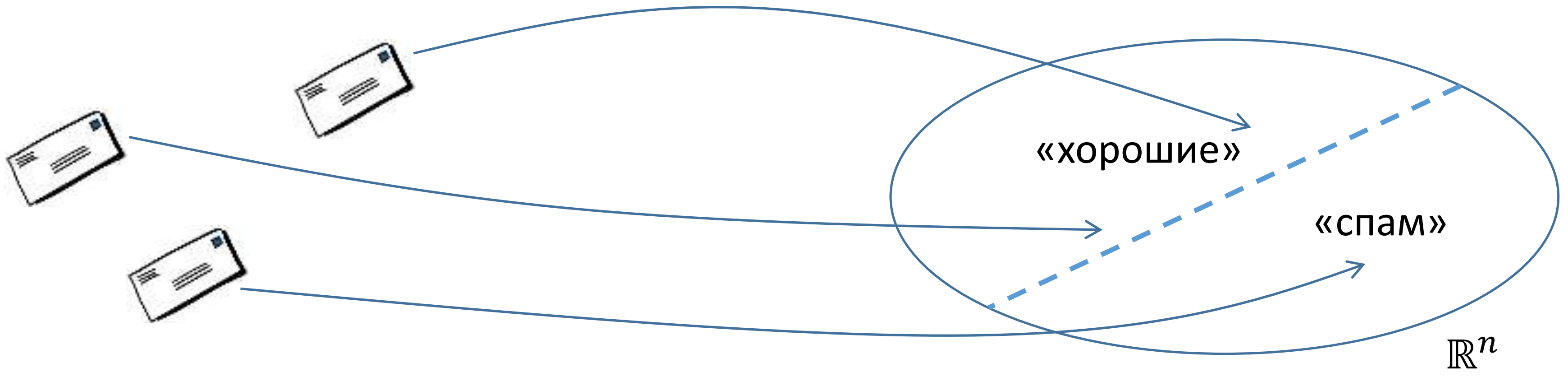
- строим отображение из множества объектов $\mathcal{O}$ в «просто устроенное» пространство — например, в $\mathbb{R}^n$



Связь с задачей классификации

Подход:

- строим отображение из множества объектов $\mathcal{O}$ в «просто устроенное» пространство — например, в $\mathbb{R}^n$,
- учимся классифицировать векторы, сопоставляемые объектам из $\mathcal{O}$, с помощью несложных функций (линейных, пороговых и т.п.) — *классификаторов*.

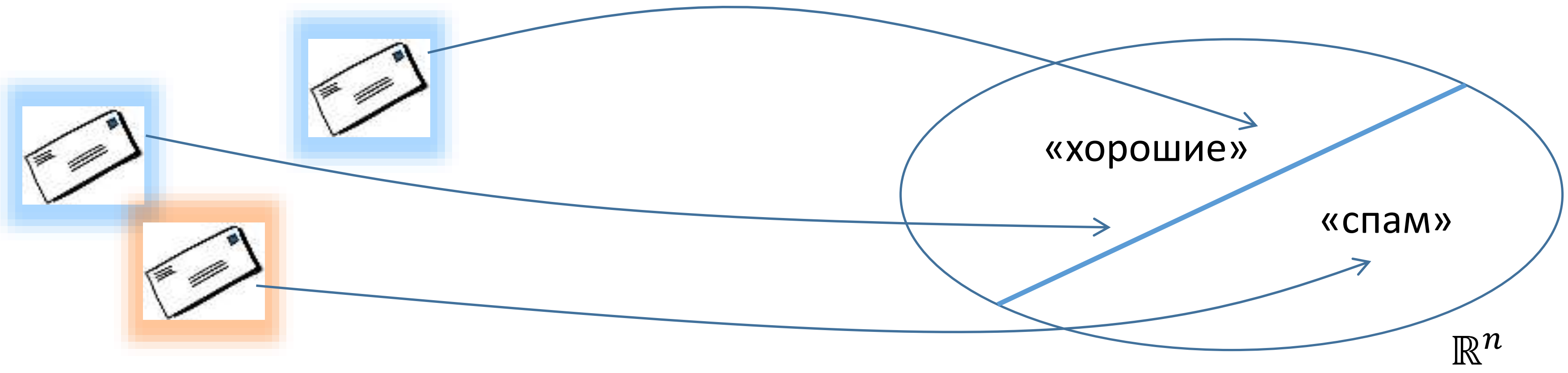


Связь с задачей классификации

Пусть A — множество векторов в $\mathbb{R}^n$, сопоставленных объектам из $\mathcal{O}$.

Пусть $\mathcal{F}$ — семейство всевозможных кусков $\mathbb{R}^n$, которые можно «отсечь» с помощью функций классификации.

Если A дробится с помощью $\mathcal{F}$, это хорошо, т.к. как бы ни были распределены по классам объекты из $\mathcal{O}$, можно построить классификатор, правильно относящий все объекты к своим классам.

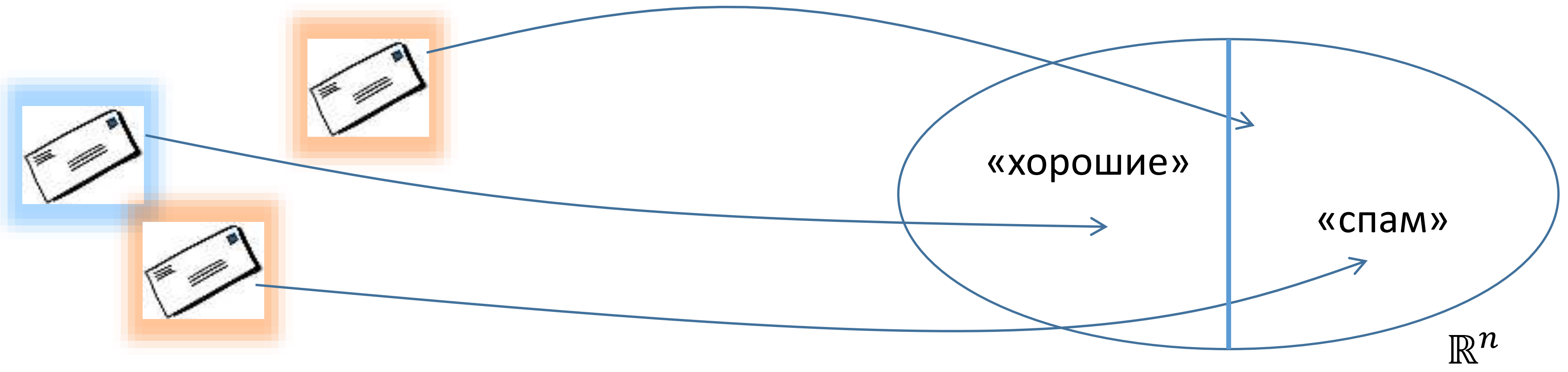


Связь с задачей классификации

Пусть A — множество векторов в $\mathbb{R}^n$, сопоставленных объектам из $\mathcal{O}$.

Пусть $\mathcal{F}$ — семейство всевозможных кусков $\mathbb{R}^n$, которые можно «отсечь» с помощью функций классификации.

Если A дробится с помощью $\mathcal{F}$, это хорошо, т.к. как бы ни были распределены по классам объекты из $\mathcal{O}$, можно построить классификатор, правильно относящий все объекты к своим классам.

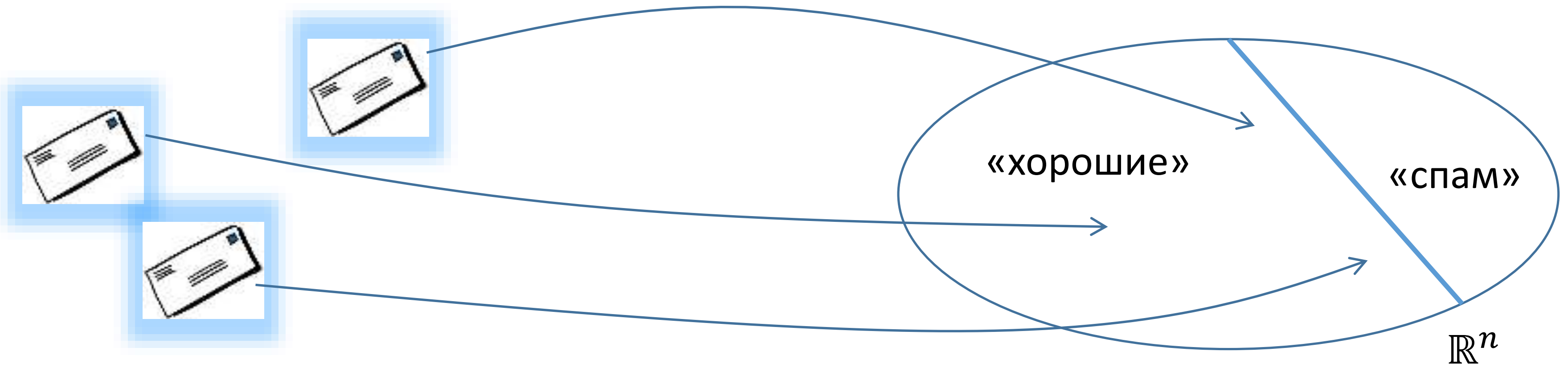


Связь с задачей классификации

Пусть A — множество векторов в $\mathbb{R}^n$, сопоставленных объектам из $\mathcal{O}$.

Пусть $\mathcal{F}$ — семейство всевозможных кусков $\mathbb{R}^n$, которые можно «отсечь» с помощью функций классификации.

Если A дробится с помощью $\mathcal{F}$, это хорошо, т.к. как бы ни были распределены по классам объекты из $\mathcal{O}$, можно построить классификатор, правильно относящий все объекты к своим классам.



Размерность Вапника—Червоненкиса

Размерность Вапника—Червоненкиса (VC-размерность) семейства $\mathcal{F}$ — это

$$\max\{\,|A| \mid A \text{ дробится с помощью } \mathcal{F}\,\}$$

Будем обозначать размерность так: $\operatorname{vc}\mathcal{F}$.

Если $\operatorname{vc}\mathcal{F} = n$, то с помощью $\mathcal{F}$ можно раздробить какое-то (необязательно произвольное!) множество мощности n.

Если максимум не достигается, будем полагать $\operatorname{vc}\mathcal{F} = \infty$.

VC-размерность семейства полуплоскостей в $\mathbb{R}^2$

Пример. Пусть $\mathcal{F}$ — семейство всех открытых полуплоскостей в $\mathbb{R}^2$.

- Никакое множество мощности 4 не дробится с помощью $\mathcal{F}$.
- Есть множества мощности 3, дробящиеся с помощью $\mathcal{F}$: любая тройка неколлинеарных точек плоскости.

Поэтому для такого $\mathcal{F}$ имеем $\mathrm{vc}\,\mathcal{F} = 3$.

В общем случае размерность семейства всех полупространств в $\mathbb{R}^n$ равна $(n+1)$.

Простые свойства VC-размерности

Свойства-упражнения:

- Если $\mathcal{F}' \subseteq \mathcal{F}''$, то $\operatorname{vc}\mathcal{F}' \leq \operatorname{vc}\mathcal{F}''$.

- Для любого множества A и любого семейства $\mathcal{F}$ выполнено
$$\operatorname{vc}(\mathcal{F}|_A) \leq \operatorname{vc}\mathcal{F}\,.$$
(Достаточно доказать, что если какое-то B дробится с помощью $\mathcal{F}|_A$, то оно же дробится и с помощью $\mathcal{F}$.)

Домен

Доменом семейства $\mathcal{F}$ назовём множество

$$\operatorname{dom} \mathcal{F} \coloneqq \bigcup_{X \in \mathcal{F}} X$$

Содержательно, домен (domain) — это набор всех элементов, встречающиеся во множествах из $\mathcal{F}$.

Будем обозначать $\|\mathcal{F}\| \coloneqq |\operatorname{dom} \mathcal{F}|$.

Связь мощности семейств и VC-размерности

Теорема.
Если $\|\mathcal{F}\| = n$ и $\operatorname{vc}\mathcal{F} = d$, то

$$|\mathcal{F}| \leq \sum_{i=0}^{d} \binom{n}{i}$$

Замечание-упражнение.
Оценка теоремы точна: достаточно рассмотреть семейство

$$\mathcal{F} \coloneqq \{X \subseteq \{1, \dots, n\} \mid |X| \leq d\}.$$

$$\|\mathcal{F}\| = n \text{ и } \operatorname{vc}\mathcal{F} = d \quad \Rightarrow \quad |\mathcal{F}| \leq \textstyle\sum_{i=0}^{d} \binom{n}{i}$$

Доказательство индукцией по n, d.

База индукции:

Если $n = 0$, то $\operatorname{dom}\mathcal{F} = \emptyset$, а значит, $\mathcal{F} = \{\emptyset\}$ и $\operatorname{vc}\mathcal{F} = 0$.

Имеем

$$|\mathcal{F}| = 1 = \sum_{i=0}^{0} \binom{0}{i}$$

$$\|\mathcal{F}\| = n \text{ и } \mathrm{vc}\,\mathcal{F} = d \quad \Rightarrow \quad |\mathcal{F}| \leq \textstyle\sum_{i=0}^{d} \binom{n}{i}$$

Доказательство индукцией по n, d.

База индукции:

Если $n > 0$ и $d = 0$, то заметим, что $|\mathcal{F}| \leq 1$.

В противном случае нашлись бы $S_1, S_2 \in \mathcal{F}$, такие, что $S_1 \setminus S_2 \neq \emptyset$.

И тогда для любого $s \in S_1 \setminus S_2$ множество $\{s\}$ дробилось бы
с помощью $\mathcal{F}$, что противоречит условию $d = 0$.

Так что получаем

$$|\mathcal{F}| \leq 1 = \sum_{i=0}^{0} \binom{n}{i}$$

Два вспомогательных семейства

Индуктивный переход:

Пусть $\|\mathcal{F}\| = n > 0,\ \mathrm{vc}\,\mathcal{F} = d > 0$, и пусть для всех семейств с меньшим доменом и/или размерностью утверждение теоремы выполнено.

Зафиксируем произвольный $x \in \mathrm{dom}\,\mathcal{F}$ и рассмотрим семейства $\mathcal{F}'$ и $\mathcal{F}''$:

- $\mathcal{F}' \coloneqq \{\ \ S \subseteq (\mathrm{dom}\,\mathcal{F}) \setminus \{x\} \quad | \quad S \in \mathcal{F} \ \ \text{или} \ \ S \cup \{x\} \in \mathcal{F}\ \ \}$
- $\mathcal{F}'' \coloneqq \{\ \ S \subseteq (\mathrm{dom}\,\mathcal{F}) \setminus \{x\} \quad | \quad S \in \mathcal{F} \ \ \text{и} \ \ S \cup \{x\} \in \mathcal{F}\ \ \}$

Заметим, что $|\mathcal{F}| = |\mathcal{F}'| + |\mathcal{F}''|$.

Параметры вспомогательных семейств

- $\mathcal{F}' \coloneqq \{\ \ S \subseteq (\operatorname{dom} \mathcal{F}) \setminus \{x\} \quad | \quad S \in \mathcal{F} \ \ \text{или} \ \ S \cup \{x\} \in \mathcal{F}\ \ \}$
- $\mathcal{F}'' \coloneqq \{\ \ S \subseteq (\operatorname{dom} \mathcal{F}) \setminus \{x\} \quad | \quad S \in \mathcal{F} \quad \text{и} \ \ S \cup \{x\} \in \mathcal{F}\ \ \}$
- $|\mathcal{F}| = |\mathcal{F}'| + |\mathcal{F}''|$

Имеем $\|\mathcal{F}'\| \leq \|\mathcal{F}\| - 1$ и $\|\mathcal{F}''\| \leq \|\mathcal{F}\| - 1$.

- Если для какого-то A выполнено $\mathcal{F}'|_A = 2^A$, то $\mathcal{F}|_A = 2^A$, поэтому $\operatorname{vc} \mathcal{F}' \leq \operatorname{vc} \mathcal{F}$.
- Если для какого-то A выполнено $\mathcal{F}''|_A = 2^A$, то $\mathcal{F}|_{A \cup \{x\}} = 2^{A \cup \{x\}}$, поэтому $\operatorname{vc} \mathcal{F}'' \leq \operatorname{vc} \mathcal{F} - 1$.

$$\|\mathcal{F}\| = n \text{ и } \operatorname{vc}\mathcal{F} = d \quad \Rightarrow \quad |\mathcal{F}| \leq \sum_{i=0}^{d} \binom{n}{i}$$

- $\mathcal{F}' := \{ \; S \subseteq (\operatorname{dom}\mathcal{F}) \setminus \{x\} \quad | \quad S \in \mathcal{F} \;\text{ или }\; S \cup \{x\} \in \mathcal{F} \; \}$
- $\mathcal{F}'' := \{ \; S \subseteq (\operatorname{dom}\mathcal{F}) \setminus \{x\} \quad | \quad S \in \mathcal{F} \;\text{ и }\; S \cup \{x\} \in \mathcal{F} \; \}$
- $|\mathcal{F}| = |\mathcal{F}'| + |\mathcal{F}''|$
- $\|\mathcal{F}'\| \leq \|\mathcal{F}\| - 1$ и $\|\mathcal{F}''\| \leq \|\mathcal{F}\| - 1$
- $\operatorname{vc}\mathcal{F}' \leq \operatorname{vc}\mathcal{F}$ и $\operatorname{vc}\mathcal{F}'' \leq \operatorname{vc}\mathcal{F} - 1$

Применяя предположение индукции, получаем

$$|\mathcal{F}| \leq \sum_{i=0}^{d-1} \binom{n-1}{i} + \sum_{i=0}^{d} \binom{n-1}{i} = \sum_{i=0}^{d} \binom{n-1}{i-1} + \sum_{i=0}^{d} \binom{n-1}{i} = \sum_{i=0}^{d} \binom{n}{i}$$

Следствие из доказанной теоремы

Теорема.
Если $\|\mathcal{F}\| = n$ и $\operatorname{vc}\mathcal{F} = d$, то $|\mathcal{F}| \le \sum_{i=0}^{d}\binom{n}{i}$.

Следствие.
Если $\operatorname{vc}\mathcal{F} = d$, то для любого множества A выполнено неравенство

$$|\mathcal{F}|_A| \le \sum_{i=0}^{d}\binom{|A|}{i}$$

Доказательство:
Достаточно заметить, что $\operatorname{vc}\mathcal{F}|_A \le \operatorname{vc}\mathcal{F}$.

Измельчения

h-*измельчением* семейства $\mathcal{F}$ называется семейство

$$\mathcal{F}^{(h)} \coloneqq \{X_1 \cap \dots \cap X_h \mid X_1, \dots, X_h \in \mathcal{F}\}$$

Очевидное соотношение:

$$\mathcal{F} = \mathcal{F}^{(1)} \subseteq \mathcal{F}^{(2)} \subseteq \mathcal{F}^{(3)} \subseteq \cdots$$

Мотивация из анализа данных:
вместо одиночных классификаторов рассматриваем конъюнкции нескольких однотипных классификаторов.

Примеры измельчений

- Если $\mathcal{F} := \big\{\{1\}, \{2,3\}, \{1,3\}\big\}$, то $\mathcal{F}^{(2)} = \big\{\emptyset, \{1\}, \{3\}, \{2,3\}, \{1,3\}\big\}$.
- Если $\mathcal{F}$ — семейство всех полуплоскостей в $\mathbb{R}^2$, то $\mathcal{F}^{(3)}$ содержит всевозможные полуплоскости, бесконечные углы, и треугольники (вместе с внутренностью).

VC-размерность измельчений

Теорема.

Для любого $\mathcal{F}$, такого, что $\mathrm{vc}\,\mathcal{F} = d \geq 2$, и для любого $h \geq 2$ выполнено неравенство

$$\mathrm{vc}\,\mathcal{F}^{(h)} \leq 2dh \log_2 dh$$

Доказательство:

Пусть A — произвольное множество, такое, что $|A| = n > 2dh \log_2 dh$.

Имеем

$$\left|\mathcal{F}^{(h)}|_A\right| \leq |\mathcal{F}|_A|^h \leq \left(\sum_{i=0}^{d} \binom{n}{i}\right)^h \leq n^{dh}.$$

Достаточно доказать, что $n^{dh} < 2^n$.

VC-размерность измельчений

Доказываем: $n^{dh} < 2^n$ при $n > 2dh \log_2 dh$.

Пусть $n = \alpha \cdot dh \log_2 dh$, где $\alpha > 2$.

Имеем

$$n^{dh} < 2^n \Leftrightarrow$$
$$\Leftrightarrow (\alpha dh \log_2 dh)^{dh} < (dh)^{\alpha dh} \Leftrightarrow$$
$$\Leftrightarrow \alpha dh \log_2 dh < (dh)^{\alpha} \Leftrightarrow$$
$$\Leftrightarrow \alpha \log_2 dh < (dh)^{\alpha - 1}$$

Последнее неравенство следует из того,
что $\alpha < (dh)^{\alpha - 1.5}$ при $\alpha > 2$ и $dh \geq 4$,
и $\log_2 dh \leq (dh)^{0.5}$ при $dh \geq 4$.

VC-размерность измельчений

Теорема.

Для любого $\mathcal{F}$, такого, что $\operatorname{vc}\mathcal{F} = d \geq 2$, и для любого $h \geq 2$ выполнено неравенство

$$\operatorname{vc}\mathcal{F}^{(h)} \leq 2dh\log_2 dh$$

Следствие.

Пусть $\mathcal{F}_\Delta$ — семейство всех треугольников (с внутренностью) на плоскости. Имеем

$$\operatorname{vc}\mathcal{F}_\Delta \leq 2 \cdot 3 \cdot 3 \cdot \log_2(3 \cdot 3) < 58$$

ε-сети

Пусть задано семейство $\mathcal{F}$ и конечное множество A.

ε-сетью для множества A (относительно $\mathcal{F}$) называется любая с.о.п. для совокупности

$$\{S \in \mathcal{F}|_A \text{ т.ч. } |S| \geq \varepsilon \cdot |A|\}$$

Иначе говоря, ε-сеть для A — это такое подмножество $A' \subseteq A$, что любое множество из $\mathcal{F}$, «залезающее» хотя бы на ε-ю долю элементов A, «зацепит» и хотя бы один элемент из A'.

Существование «экономных» ε-сетей

Теорема. (Хаусслер—Вельцль ’1987)
Если семейство $\mathcal{F}$ таково, что $\mathrm{vc}\,\mathcal{F} = d < \infty$, то для любого $\varepsilon > 0$ и любого A существует ε-сеть размера не больше $\frac{8d}{\varepsilon}\log_2\frac{8d}{\varepsilon}$.

Сила теоремы: оценка в теореме никак не зависит от размера множества A!

Геометрическое следствие теоремы Хаусслера—Вельцля

Теорема. *(Простое следствие теоремы Х-В)*
Пусть $\varepsilon \in (0,1)$ — сколь угодно малое число, и пусть $n = \left\lceil \frac{500}{\varepsilon} \log_2 \frac{500}{\varepsilon} \right\rceil$.

Пусть A — *произвольное* конечное множество точек на плоскости.

Тогда в A можно «испачкать» n точек, так, что *любой* треугольник, содержащий по крайней мере ε-ю долю точек из A, окажется запачканным (т.е. в него попадёт хотя бы одна запачканная точка).

Техническая лемма

Утверждение. Пусть проводятся k испытаний Бернулли с вероятностью успеха в отдельном испытании не менее ε.
Тогда вероятность того, что всего произойдёт не менее $\frac{\varepsilon k}{2}$ успехов, не меньше $1 - \frac{4}{\varepsilon k}$.

Доказательство. Пусть ξ — число успехов в k испытаниях. Имеем

$$\Pr\left\{\xi \geq \frac{\varepsilon k}{2}\right\} \quad = \quad 1 - \Pr\left\{\xi < \frac{\varepsilon k}{2}\right\} \quad \geq \quad 1 - \Pr\left\{|\xi - \mathbb{E}\xi| \geq \frac{\varepsilon k}{2}\right\} \quad \geq$$

$$\geq \quad 1 - \frac{\mathbb{D}\xi}{\left(\frac{\varepsilon k}{2}\right)^2} \quad = \quad 1 - \frac{k \cdot \varepsilon(1 - \varepsilon)}{\left(\frac{\varepsilon k}{2}\right)^2} \quad \geq \quad 1 - \frac{4}{\varepsilon k}.$$

Доказательство теоремы Х—В

Пусть A — произвольное множество.

Положим $k \coloneqq \left\lfloor \frac{8d}{\varepsilon} \log_2 \frac{8d}{\varepsilon} \right\rfloor$.

Выберем случайное подмножество в A мощности k и докажем, что оно с ненулевой вероятностью будет искомой ε-сетью.

Будем случайно выбирать упорядоченный набор $U = (u_1, \ldots, u_k)$, где $a_i \in A$.

Каждая компонента набора выбирается из n возможностей равновероятно. (Допускаются повторения в U.)

Доказательство теоремы Х—В

«Плохое» событие для нас:

$$B := \{ \;\exists S \in \mathcal{F} \quad \text{т. ч.} \quad |S \cap A| \geq \varepsilon n \;\; \text{и} \;\; S \cap U = \emptyset \; \}$$

<u>Расширение «плохого» события:</u>

Случайно выберем набор U длины k
и независимо случайно выберем набор $V := (v_1, \dots, v_k)$.

Положим

$$B' := \left\{\exists S \in \mathcal{F} \;\; \text{т. ч.} \;\; |S \cap A| \geq \varepsilon n \;\; \text{и} \;\; S \cap U = \emptyset \;\; \text{и} \;\; |S \cap V| \geq \frac{\varepsilon k}{2}\right\}$$

Здесь $|S \cap V| := \#\{i \in \{1, \dots, k\} \mid v_i \in S\}$.

Лирическое отступление

Казалось бы, зачем вводить какое-то V, когда этого совершенно не требует формулировка задачи?

Нам нужно как-то столкнуть два враждующих события: что некоторое S *сильно пересеклось* с A, но при этом *никак не пересеклось* со случайным подмножеством A.

Столкнуть здесь означает показать, что вероятность одновременного совпадения этих противоположностей мала.

Мы не знаем свойства S (перед ним квантор $\exists$, оно выбирается из какого-то сложного $\mathcal{F}$), зато у нас есть конечное фиксированное A, и выбором ещё одного случайного $V \subset A$ мы сглаживаем своё незнание структуры S.

Случайность здесь — это вид усреднения, сглаживания острых углов.

Связь вероятностей $\Pr B$ и $\Pr B'$

- $B := \{\exists S \in \mathcal{F} \quad \text{т.ч.} \quad |S \cap A| \geq \varepsilon n \quad \text{и} \quad S \cap U = \emptyset\}$
- $B' := \left\{\exists S \in \mathcal{F} \quad \text{т.ч.} \quad |S \cap A| \geq \varepsilon n \quad \text{и} \quad S \cap U = \emptyset \quad \text{и} \quad |S \cap V| \geq \frac{\varepsilon k}{2}\right\}$

Утверждение.

$$\Pr B' \geq \frac{1}{2} \Pr B$$

Доказательство:

Т.к. B' включает B, то $\frac{\Pr B'}{\Pr B} = \Pr(B' \mid B)$.

Значит, достаточно доказать, что $\Pr(B' \mid B) \geq \frac{1}{2}$.

Связь вероятностей $\Pr B$ и $\Pr B'$

- $B := \{\exists S \in \mathcal{F} \quad \text{т.ч.} \quad |S \cap A| \geq \varepsilon n \quad \text{и} \quad S \cap U = \emptyset\}$
- $B' := \left\{\exists S \in \mathcal{F} \quad \text{т.ч.} \quad |S \cap A| \geq \varepsilon n \quad \text{и} \quad S \cap U = \emptyset \quad \text{и} \quad |S \cap V| \geq \frac{\varepsilon k}{2}\right\}$

Доказываем, что $\Pr(B' \mid B) \geq \frac{1}{2}$.

Пусть мы случайно выбрали U и оказалось, что для некоторого $S_0 \in \mathcal{F}$ выполнилось условие $|S_0 \cap A| \geq \varepsilon n$.

Тогда при случайном выборе V имеем

$$\Pr\left\{|S_0 \cap V| \geq \frac{\varepsilon k}{2}\right\} = \Pr\left\{\text{среди элементов } v_1, \dots, v_k \text{ не менее } \frac{\varepsilon k}{2} \text{ попали в } S_0\right\}$$

Это равно вероятности того, что при k испытаниях Бернулли с вероятностью успеха $\frac{|S_0 \cap A|}{|A|}$ произойдёт не менее $\frac{\varepsilon k}{2}$ успехов.

Связь вероятностей $\Pr B$ и $\Pr B'$

- $B \coloneqq \{\exists S \in \mathcal{F} \;\text{ т.ч. }\; |S \cap A| \geq \varepsilon n \;\text{ и }\; S \cap U = \emptyset\}$
- $B' \coloneqq \left\{\exists S \in \mathcal{F} \;\text{ т.ч. }\; |S \cap A| \geq \varepsilon n \;\text{ и }\; S \cap U = \emptyset \;\text{ и }\; |S \cap V| \geq \frac{\varepsilon k}{2}\right\}$

При случайном выборе U получилось $|S_0 \cap A| \geq \varepsilon n$ для $S_0 \in \mathcal{F}$.

$\Pr(B' \mid B)$ — вероятность того, что при k испытаниях Бернулли с вероятностью успеха $\frac{|S_0 \cap A|}{|A|} \geq \varepsilon$ произойдёт не менее $\frac{\varepsilon k}{2}$ успехов.

$$\Pr(B' \mid B) \quad \geq \quad 1 - \frac{4}{\varepsilon k} \quad = \quad 1 - \frac{4}{\varepsilon \left\lceil \frac{8d}{\varepsilon} \log_2 \frac{8d}{\varepsilon} \right\rceil} \quad > \quad \frac{1}{2}.$$

Смена вероятностной модели

- $B \coloneqq \{\exists S \in \mathcal{F} \quad \text{т. ч.} \quad |S \cap A| \geq \varepsilon n \quad \text{и} \quad S \cap U = \emptyset\}$
- $B' \coloneqq \left\{\exists S \in \mathcal{F} \quad \text{т. ч.} \quad |S \cap A| \geq \varepsilon n \quad \text{и} \quad S \cap U = \emptyset \quad \text{и} \quad |S \cap V| \geq \frac{\varepsilon k}{2}\right\}$
- $\Pr B < 2 \Pr B'$

Осталось теперь показать, что $\Pr B' < \frac{1}{2}$.

Рассмотрим другую вероятностную модель: выбираем случайно набор $T = (t_1, \dots, t_{2k}) \in A^{2k}$, а затем случайно (одним из $\binom{2k}{k}$ способов) делим его на два набора U и V.

Заметим, что вероятности всех событий, зависящих от U и V, в новой модели те же, что и в старой.

Смена вероятностной модели

- Выбираем случайно $T \in A^{2k}$, случайно разбиваем его на U и V.
- $B' \coloneqq \left\{\exists S \in \mathcal{F} \quad \text{т.ч.} \quad |S \cap A| \geq \varepsilon n \quad \text{и} \quad S \cap U = \emptyset \quad \text{и} \quad |S \cap V| \geq \frac{\varepsilon k}{2}\right\}$
- Хотим показать, что $\Pr B' < \frac{1}{2}$.

$$\Pr B' = \sum_{T_0 \in A^{2k}} \Pr(B' \mid T = T_0) \cdot \Pr\{T = T_0\}$$

Достаточно показать, что при любом фиксированном T_0 выполнено

$$\Pr(B' \mid T = T_0) \leq \frac{1}{2}.$$

Оценка $\Pr B'$ при фиксированном T

- Выбираем случайно $T \in A^{2k}$, случайно разбиваем его на U и V.
- $B' := \left\{\exists S \in \mathcal{F} \;\text{ т.ч. }\; |S \cap A| \geq \varepsilon n \;\text{ и }\; S \cap U = \emptyset \;\text{ и }\; |S \cap V| \geq \frac{\varepsilon k}{2}\right\}$
- Хотим показать, что $\Pr(B' \mid T = T_0) \leq \frac{1}{2}$ для любого T_0.

Зафиксируем $T = T_0$.

$B' = \bigcup_{S \in \mathcal{F}} B'_S$, где $B'_S := \left\{|S \cap A| \geq \varepsilon n \;\text{ и }\; S \cap U = \emptyset \;\text{ и }\; |S \cap V| \geq \frac{\varepsilon k}{2}\right\}$.

Т.к. событие B'_S определяется не целым S, а лишь пересечением $S \cap T_0$, то

$$B' \subseteq \bigcup_{S \in \mathcal{F}|_{T_0}} \left\{S \cap U = \emptyset \;\text{ и }\; |S \cap V| \geq \frac{\varepsilon k}{2}\right\}$$

Оценка $\Pr B'$ при фиксированном T

- Выбираем случайно $T \in A^{2k}$, случайно разбиваем его на U и V.
- $B' = \bigcup_{T_0} \bigcup_{S \in \mathcal{F}|_{T_0}} \left\{T = T_0 \quad \text{и} \quad S \cap U = \emptyset \quad \text{и} \quad |S \cap V| \geq \frac{\varepsilon k}{2}\right\}$
- Хотим показать, что $\Pr(B' \mid T = T_0) \leq \frac{1}{2}$ для любого T_0.

Зафиксируем T_0. Событие $\left\{S \cap U = \emptyset \text{ и } |S \cap V| \geq \frac{\varepsilon k}{2}\right\}$ происходит только если $|T_0 \cap S| = t \geq \frac{\varepsilon k}{2}$.

$\Pr\{$при $|T_0 \cap S| = t$ разбиение T_0 на U и V окажется таким, что $S \cap U = \emptyset\} =$

$$= \frac{\binom{2k-t}{k}}{\binom{2k}{k}} = \frac{k(k-1) \cdot \ldots \cdot (k-t+1)}{2k(2k-1) \cdot \ldots \cdot (2k-t+1)} \leq 2^{-t} \leq 2^{-\varepsilon k/2}$$

Оценка $\Pr B'$ при фиксированном T

- Выбираем случайно набор $T \in A^{2k}$, а затем случайно разбиваем его на два набора U и V.
- $B' = \bigcup_{S \in \mathcal{F}|_{T_0}} \left\{ S \cap U = \emptyset \quad \text{и} \quad |S \cap V| \geq \frac{\varepsilon k}{2} \right\}$
- Хотим показать, что $\Pr(B' \mid T = T_0) \leq \frac{1}{2}$ для любого T_0.

Из доказанного следует, что при фиксированном T

$$\Pr B' \quad \leq \quad \left| \mathcal{F}|_{T_0} \right| \cdot 2^{-\varepsilon k/2} \quad \leq \quad 2^{-\varepsilon k/2} \cdot \sum_{i=0}^{d} \binom{2k}{i} \quad < \quad \frac{1}{2}$$

(при $k = \left\lfloor \frac{8d}{\varepsilon} \log_2 \frac{8d}{\varepsilon} \right\rfloor$).