ЛЕКЦИЯ 20

РАЗМЕРНОСТЬ ВАПНИКА-ЧЕРВОНЕНКИСА

1. Ранжированное пространство

Перейдем к изучению новой темы. Рассмотрим следующую задачу. Пусть на плоскости расположено произвольным образом какое-то множество X, состоящее из n точек (см. рис. 20.1).

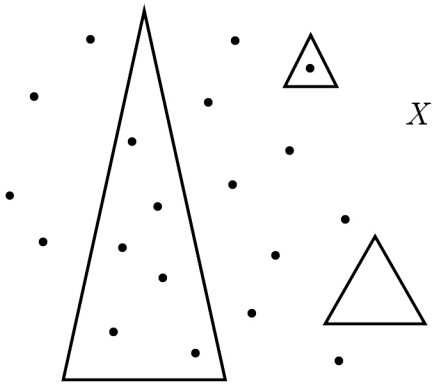

Рис. 20.1

Начнем пересекать данное множество со всевозможными треугольниками на плоскости. В результате пересечения каждый раз будет получаться некое подмножество исходного множества X.

Зафиксируем число ϵ:

$$\epsilon > 0.$$

Рассмотрим только те подмножества, в каждом из которых не меньше, чем ϵn точек.

Обозначим стандартным образом совокупность, состоящую из таких подмножеств:

$$\mathcal{M} = \{M_1, \dots, M_s\}.$$

Очевидно, что s зависит от того, как расположено исходное множество X на плоскости.

Определение 33: *Назовем* ϵ***-сетью*** *для множества* X *любую систему общих представителей для совокупности* $\mathcal{M}$. ♣

Следующую теорему Вапник и Червоненкис доказывали несколько в иных терминах и в более общей формулировке, тем не менее, будем рассматривать ее в таком виде:

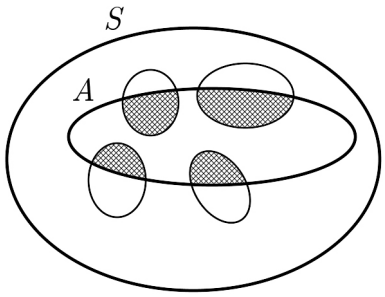

Рис. 20.2

Теорема 64 (1971, Вапник, Червоненкис)

$\forall \epsilon > 0\ \forall X,\ |X| = n:\ \exists$ ϵ-сеть, размер которой не превосходит

величины $\dfrac{1000}{\epsilon} \log_2 \dfrac{1000}{\epsilon}$. ∗

Данная теорема не будет доказываться, потому что далее будет доказан более общий результат.

Определение 34: *Пусть имеется упорядоченная пара* (S, R), *где* S — *произвольное множество, не обязательно конечное, а* R — *произвольная совокупность его подмножеств, также не обязательно конечная. Тогда такая пара* (S, R) *называется* ***ранжированным множеством***. ♣

Определение 35: *Рассмотрим какое-то подмножество* A:

$$A \subset S.$$

Тогда ***проекцией*** *на множество* A *системы областей* R *называется следующее (см. рис. 20.2):*

$$P_{r_A}(R) = \Big\{ r \cap A:\ r \in R \Big\} \to \Big(A, P_{r_A}(R) \Big).$$

Определение 36: *Множество* A ***дробится*** *областями из множества* R, *если:*

$$\big| P_{r_A}(R) \big| = 2^{|A|}.$$

Определение 37: ***Размерность Вапника – Червоненкиса*** *пространства* S:

$$VC(S, R) = \max \Big\{ m:\ \exists A \subset S:\ |A| = m \text{ и } A \text{ — дробится} \Big\}.$$

Замечание Если максимум нигде не достигается, то в таком случае говорится, что размерность Вапника – Червоненкиса равна бесконечности.

Пример такого пространства — любое бесконечное множество в качестве S, а в качестве R — множество всех его подмножеств.

Более конкретный пример — множество натуральных чисел в качестве S и все наборы натуральных числе в качестве областей. $*$

Рассмотрим бесконечное множество с конечной размерностью:

Пример 14 Пусть:

$$S = \mathbb{R}^n, \quad R = H,$$

где H — множество всех открытых полупространств (открытость полупространства здесь означает, что не будут рассматриваться его границы).

Нужно посчитать, чему тогда равна размерность Вапника – Червоненкиса для такого пространства:

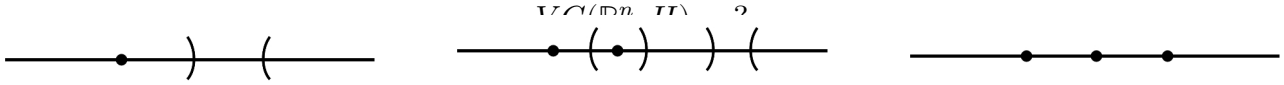

Рис. 20.3

Отметим, что полученный результат используется в статистике.

Посчитаем размерность Вапника – Червоненкиса в случае прямой.

На рисунке **??** видно, что точка на прямой дробится открытыми лучами — один открытый луч на рисунке содержит точку, а другой — не содержит.

То есть, любое одноэлементное множество на прямой — дробится.

Также и любое двухэлементное множество на прямой — дробится (см. рис. **??**).

Но никакое трехэлементное множество на прямой не может дробиться (см. рис. **??**), так как центральную точку никак нельзя отделить.

Таким образом, получен результат для прямой:

$$VC(\mathbb{R}^1, H) = 2.$$

Несложно показать аналогично, что для плоскости результат будет следующим:

$$VC(\mathbb{R}^2, H) = 3,$$

то есть три точки на плоскости — дробятся, а четыре — уже нет.

На рисунке **??** показано дробление трех точек на плоскости (в случае, если они не лежат на одной прямой).

На рисунке **??** приведена картина для четырех точек в лучшей ситуации — когда они образуют выпуклый четырехугольник. Видно, что ни одно из двух выделенных множеств отдробить нельзя.

На рисунке **??** показано еще одно возможное взаимное расположение четырех точек — когда одна точка лежит внутри трех остальных. Тогда очевидно, что никак нельзя отдробить центральную точку.

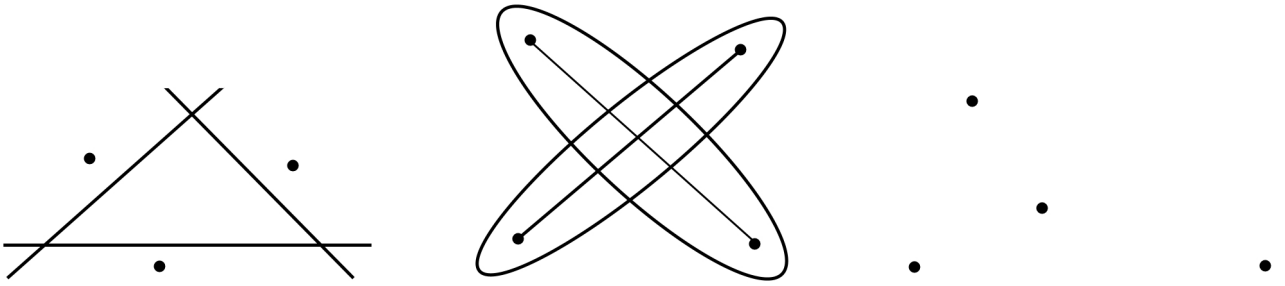

Рис. 20.4

Если же хотя бы три точки из четырех лежат на одной прямой, то это тем более невозможно. Тогда получаем, что размерность Вапника – Червоненкиса для плоскости найдена верно.

В общем случае можно записать:

$$VC(\mathbb{R}^n, H) \geqslant +1.$$

Теорема 65 (Радон) Пусть:

$$X \subset \mathbb{R}^n : \ |X| \geqslant n+2.$$

Тогда:

$$\exists X_1, X_2 : \ X = X_1 \sqcup X_2,$$

и выпуклая оболочка множества X_1 в пересечении с выпуклой оболочкой множества X_2 не дает пустое множество. $*$

Доказательство теоремы 65 остается в качестве несложного упражнения.

Докажем некоторые вспомогательные утверждения для формулировки общего результата — обобщения теоремы 64.

Лемма 10 Пусть (S, R) имеет размерность Вапника – Червоненкиса, равную d, и:

$$|S| = n.$$

Тогда:

$$|R| \leqslant g(n,d) = \sum_{k=0}^{d} C_n^k.$$

Док-во: Очевидно, что:

$$g(n,d) = g(n,d-1) + g(n-1,d).$$

Это выполняется за счет свойств треугольника Паскаля.

Значит, доказываем лемму двумерной индукцией по n и d.

База двумерной индукции — это проверка нужного неравенства отдельно при $n = 0$ и $d = 0$.

Задача Проверить указанным способом базу двумерной индукции в лемме 10. *

Разберемся с шагом индукции. Рассмотрим (S, R) и по нему построим два новых ранжированных пространства:

$$S_1 = S_2 = S \setminus \{x\},$$

где x — произвольный элемент S,

$$R_1 = \Big\{r \setminus \{x\} : \ r \in R\Big\},$$

$$R_2 = \Big\{r \in R : \ x \notin R, \ r \cup \{x\} \in R\Big\}.$$

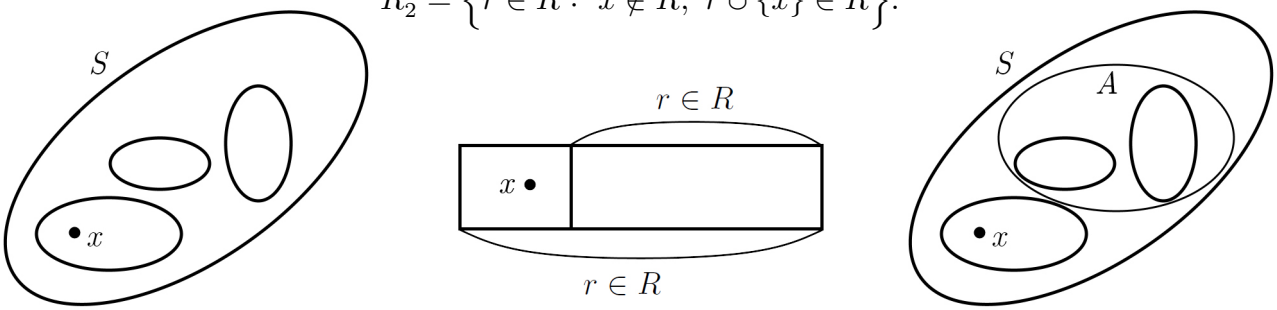

Рис. 20.5

На рисунке **??** показано устройство R_1. Заметим, что удаление элемента x из области, не содержащей его, ничего не меняет.

На рисунке **??** показано удаление элемента x в такой ситуации. При его удалении из большей области получаем меньшую, но и при его удалении из меньшей области получаем ее же. Это поясняет устройство R_2.

Тогда количество исходных областей:

$$|R| = |R_1| + |R_2|.$$

Осталось понять, каковы размерности:

$$VC(S_1, R_1), \quad VC(S_2, V_2).$$

Очевидно, что:

$$VC(S_1, R_1) \leqslant d,$$

так как у исходного пространства эта размерность равна d.

Далее утверждается, что:

$$VC(S_2, R_2) \leqslant d - 1.$$

Предположим, что:

$$VC(S_2, R_2) = d.$$

Следовательно:

$$\exists A \subset S_2 : \ |A| = d \text{ и } A \text{ дрбоится } R_2.$$

Но тогда (см. рис. **??**):

$$A \cap \{x\} \subset S \text{ — дробится } R.$$

Тогда, пользуясь вторым комбинаторным тождеством:

$$|R| = |R_1| + |R_2| \leqslant g(n-1, d-1).$$

Лемма доказана. ■

Имеется очевидное следствие:

Следствия: Пусть:

$$VC(S,R) = d,$$
$$A \subset S: \ |A| = n.$$

Тогда:

$$|P_{r_A}| \leqslant g(n,d),$$

то есть рассматривается пространство:

$$\Big(A, P_{r_A}(R)\Big).$$

Определение 38: *Пусть:*

$$h \geqslant 2, \quad h \in N.$$

Тогда ***измельчением*** *по определению называется следующее:*

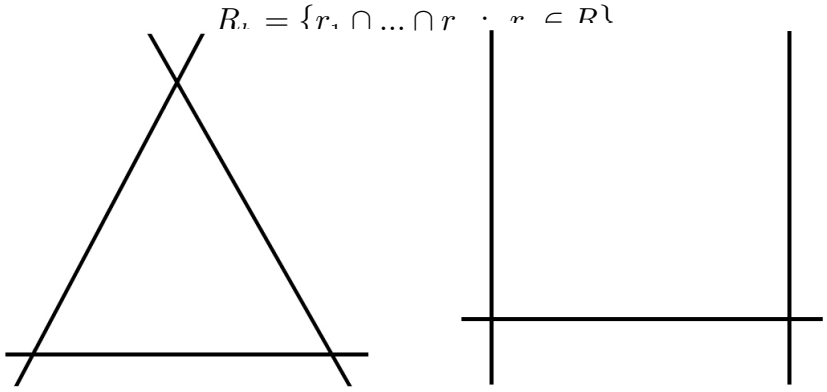

Рис. 20.6

Тогда H_3 — точно содержит множество всех треугольников на плоскости T_3 (см. рис. **??**), однако, бывают и «лоханки» (см. рис. **??**).

Тогда запишем в новых терминах:

Лемма 11 Пусть:

$$h \geqslant 2, \quad d \geqslant 2,$$
$$VC(S,R) = d.$$

Тогда:

$$VC(S, R_h) \leqslant 2dh \log_2 dh.$$

Док-во: Докажем конечность $VC(S, R_h)$, а не точную верхнюю оценку.

Пусть:

$$A \subset S, \quad |A| = n.$$

С одной стороны, если A — дробится R_h, то, по определению:

$$\left|P_{r_A}(R_h)\right| = 2^n.$$

С другой стороны:

$$\left|P_{r_A}(R_h)\right| \leqslant \left|P_{r_A}(R)\right|^h \leqslant \left(g(n,d)\right)^h.$$

На рисунке 20.7 взяты всевозможные пары из пересечений, затем тройки и так далее. Видно, что пересечение пересечений в итоге дает степень h.

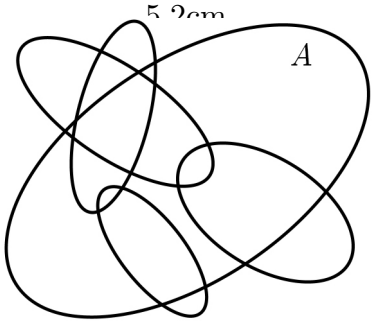

Рис. 20.7

Задача Доказать, что:

$$g(n,d) \leqslant n^d.$$

Воспользовавшись результатом упражнения 12, получим, что:

$$\left|P_{r_A}(R_h)\right| \leqslant \left(g(n,d)\right)^h \leqslant n^{dh}.$$

Следовательно, если:

$$2^n > n^{dh},$$

то никакое A не может дробиться. Значит, начиная с какого-то n нужное неравенство выполняется.

Остается лишь показать, что именно для:

$$n = 2dh \log_2 dh$$

выполняется исходное неравенство, а конечность размерности $VC(S, R_h)$ — уже доказана. ■

Следствия:

$$VC(\mathbb{R}^2, T_3) \leqslant VC(\mathbb{R}^2, H_3) \leqslant 2 \cdot 3 \cdot 3 \cdot \log_2(3 \cdot 3) \leqslant 60,$$

поскольку:

$$h = 3, \quad d = VC(\mathbb{R}^2, H) = 3.$$

2. Обобщение результата

Рассмотрим определение более общего характера:

Определение 39: *Пусть (S, R) — ранжированное пространство. Пусть также:*

$$A \subseteq S \quad и \quad \epsilon > 0.$$

Рассмотрим те и только те множества $r \cap A$ из проекции $|P_{r_A}(R)|$, у которых:

$$\left|r \cap A\right| \geqslant 2|A|,$$

где A — конечно.

Тогда любая система общих представителей для совокупности всех таких $r \cap A$ называется ***ϵ-сетью.*** ♣

Теорема 66 (Вапник, Червоненкис) Пусть:

$$VC(S, R) = d,$$

а $A \subset S$ — это любое конечное подмножество.

Пусть также:

$$\epsilon > 0.$$

Тогда существует ϵ-сеть размера:

$$m \leqslant \left\lceil \frac{8d}{\epsilon} \log_2 \frac{8d}{\epsilon} \right\rceil.$$

Замечание Если:

$$VC(S, R) = \infty,$$

то не выполняется ограничение из теоремы 66 на размер ϵ-сети. ∗

Утверждение 18 Если:

$$VC(S, R) = \infty,$$

то размер ϵ-сети линейно зависит от размера исходного множества A. ∗

Док-во: Так как размерность — бесконечна, то:

$$\forall n \ \exists A : \ |A| = n \text{ и } A \text{ — дробится.}$$

Следовательно, совокупность, для которой системой общих представителей является ϵ-сеть, состоит из всех не менее, чем (ϵn)-элементных подмножеств данного множества A.

Таким образом, любая ϵ-сеть имеет размер не меньше, чем $(n - \epsilon n + 1)$, что является линейной зависимостью. ■

На следующей лекции будет доказана вероятностным методом теорема 66.