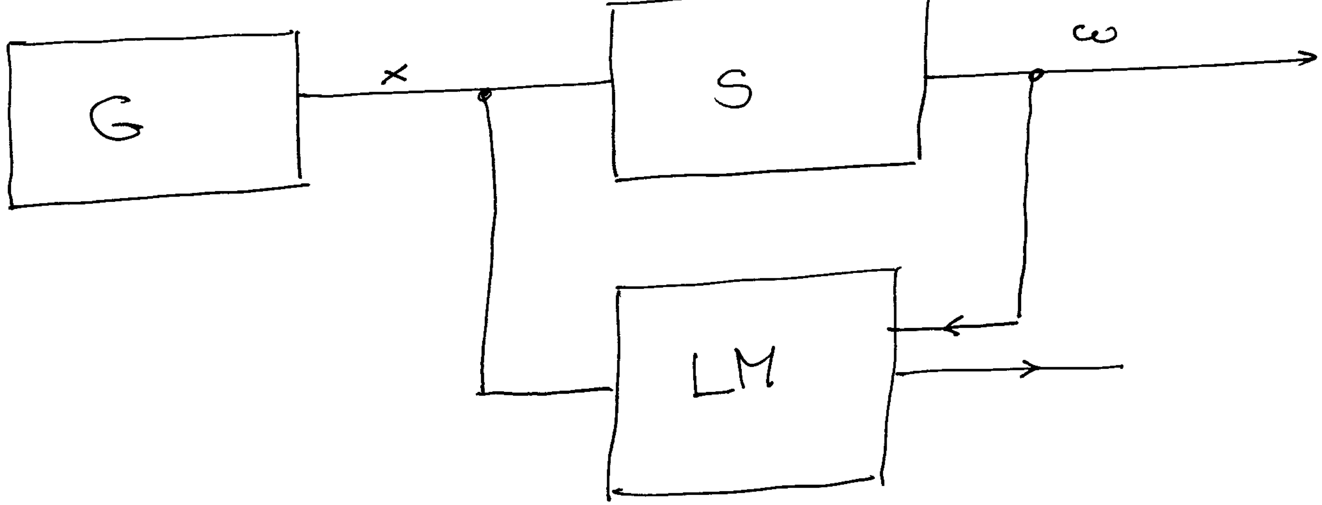
Teoria uczenia się pojęć

(V. Vapnik : Statistical Learning Theory, Wiley 1998)

Materiał wykładu znajduje się w artykule "~~Introduction to~~
"An overview of statistical learning theory" (plik LearningTheory1.pdf)
oraz w pliku PrzykladyVC.pdf)

Generator losowych wektorów (przykładów) $x \in R^n$ pojawiających się zgodnie z ustalonym ale nieznanym rozkładem $P(x)$

Nauczyciel podający decyzję dla x zgodnie z rozkładem $P(\omega|x)$



LM – maszyna ucząca ma do dyspozycji zbiór funkcji $\{\Phi(x,\alpha) : \alpha \in \Lambda\}$

oraz próbkę $(x_1,\omega_1),\ldots,(x_l,\omega_l)$ wygenerowaną zgodnie z nieznanym jej rozkładem $P(x,\omega)$

– ma za zadanie wybrać "najlepszą" funkcję względem ryzyka (teoretycznego) $R(\alpha)$

Funkcja straty

• Rozważmy najprostszy przypadek tzw. "pattern recognition"

$$L(\omega, \omega') = \begin{cases} 1 & \omega \neq \omega' \\ 0 & \omega = \omega'. \end{cases}$$

Wtedy $L(\omega, \phi(x,\alpha))$ ma wartość 1 jeśli wybór funkcji $\phi(x,\alpha)$ daje decyzję różną od ω i 0 w przeciwnym przypadku.

• $z = (x, \omega)$

$$Q(z,\alpha) \overset{ozn}{=} L(\omega, \phi(x,\alpha)), \text{ gdzie } z = (x,\omega)$$

$$Q(z,\alpha) \in \{0, 1\}$$

• Bardziej złożony przypadek, gdy np. $Q(z,\alpha) \in [A,B]$, gdzie $0 \leq A \leq B$, A, B – liczby rzeczywiste – wspomnimy o nim dalej; wyniki mogą być uogólnione na ten przypadek.

Ryzyko empiryczne

próbka o dł. l : $\mathbb{Z} = (z_1, \dots, z_l)$, gdzie $z_i = (x_i, \omega_i)$ dla $i = 1, \dots, l$

$$R_{emp}(\alpha) = \frac{1}{l} \sum_{i=1}^{l} Q(z_i, \alpha) \qquad \alpha \in \Lambda$$

Uwaga. $R_{emp}(\alpha)$ faktycznie zależy od $\mathbb{Z}$ (ale zwyczajowo nie piszemy $R_{emp}(\alpha, \mathbb{Z})$ pamiętając o tym!)

Ryzyko teoretyczne

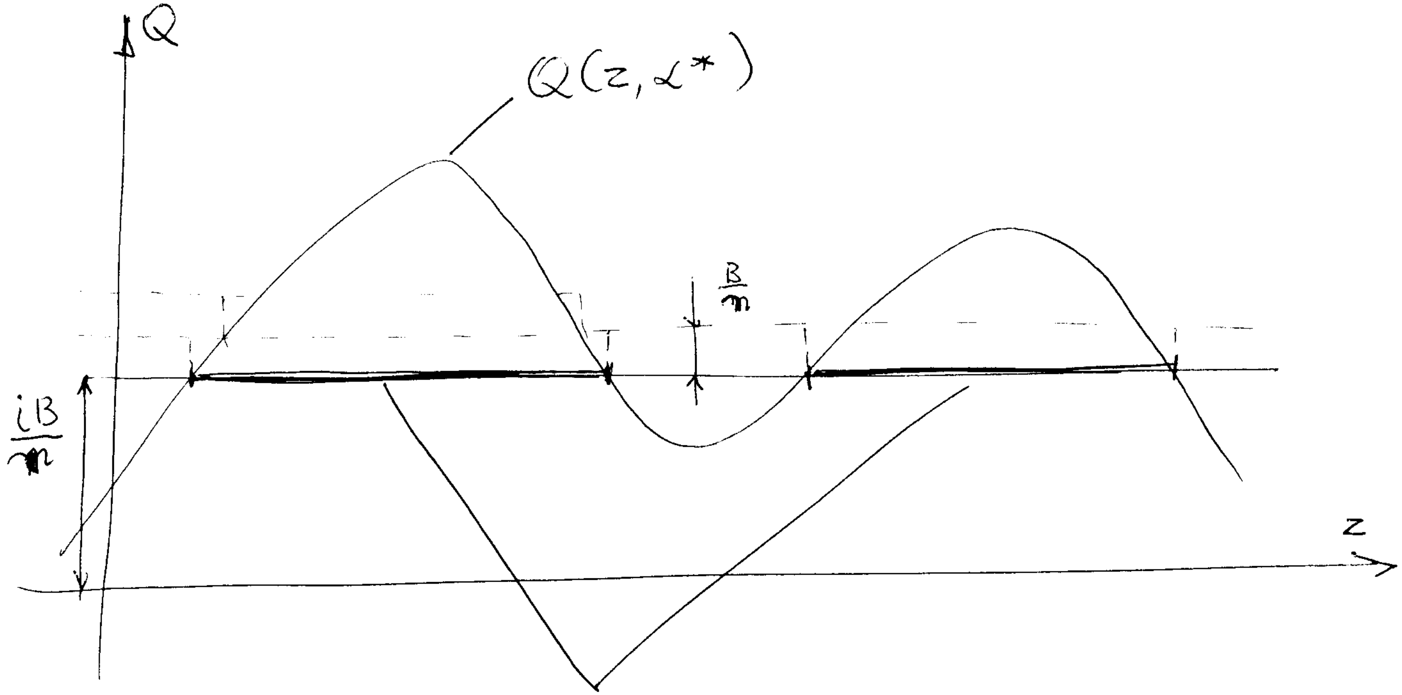
$$R(\alpha) = \int Q(z, \alpha)\, dP(z)$$

całka Lebesque – Stieltjes'a

$$0 \leqslant Q(z, \alpha^*) \leqslant B$$

ustalone

$$R(\alpha^*) = \int Q(z,\alpha^*)\,dP(z) = \lim_{n \to \infty} \sum_{i=1}^{n} \frac{B}{n} \underbrace{P\left\{ Q(z,\alpha^*) > \frac{iB}{n} \right\}}_{P\left\{ z :\, Q(z,\alpha^*) > \frac{iB}{n} \right\}}$$



$$P\left\{ Q(z,\alpha^*) > \frac{iB}{n} \right\}$$

Zasada minimalizacji ryzyka empirycznego

(ERM – empirical risk minimization)

Dla zadanej próbki $Z = (z_1, \ldots, z_\ell)$ wyznaczamy optymalne α_ℓ

takie, że

$$R_{emp}(\alpha_\ell) = \inf_{\alpha \in \Lambda} R_{emp}(\alpha).$$

Czy takie postępowanie gwarantuje, że ze wzrostem ℓ

~~"zbiegają"~~ "zbliżamy" się do $R(\alpha_{opt})$ gdzie

$$R(\alpha_{opt}) = \inf_{\alpha \in \Lambda} R(\alpha) \; ?$$

Niesprzeczność ERM

$\mathbf{Z}=(z_1,\dots,z_\ell)$ α_ℓ – optymalne dla $\mathbf{Z}$ ze względu na $R_{emp}(\alpha)=\frac{1}{\ell}\sum_{i=1}^{\ell}Q(z_i,\alpha)$

$$R_{emp}(\alpha_\ell)=\inf_{\alpha\in\Lambda}R_{emp}(\alpha)$$

$$R(\alpha)=\int Q(z,\alpha)\,dP(z)$$

$$R(\alpha_{opt})=\inf_{\alpha\in\Lambda}R(\alpha)$$

1. $R(\alpha_\ell)\xrightarrow[\ell\to\infty]{P}R(\alpha_{opt})$ tzn. $\forall\varepsilon>0\ \forall\eta>0\ \exists\,\ell_0=\ell(\varepsilon,\eta)\ \forall\,\ell>\ell_0$ $P\{\mathbf{Z}:\ \mathbf{Z}\text{-o dł. }\ell\ \&\ R(\alpha_\ell)-R(\alpha_{opt})<\varepsilon\}>1-\eta$

2. $R_{emp}(\alpha_\ell)\xrightarrow[\ell\to\infty]{P}R(\alpha_{opt})$

Tw. $\{Q(z,\alpha):\alpha\in\Lambda\}$ rodzina funkcji taka, że $A\le\int Q(z,\alpha)\,dP(z)\le B$ dla $\alpha\in\Lambda$.
Warunkiem koniecznym i dostatecznym na to aby ERM było niesprzeczne jest aby dla każdego $\varepsilon>0$

$$\lim_{\ell\to\infty}P\{\mathbf{Z}:\ \mathbf{Z}\text{-próbka o dł. }\ell\ \&\ \sup_{\alpha\in\Lambda}\left(R(\alpha)-R_{emp}(\alpha)\right)>\varepsilon\}=0$$

Problemy

1. ~~Czy~~ Kiedy zbieżność wymieniona w zasadzie niesprzeczności ma miejsce?

2. Jak zagwarantować aby ta zbieżność była "szybka"?

3. Jak uniezależnić się od rozkładu prawdopodobieństwa P w szczególności na szybkość zbieżności?

Z tymi problemami związane są główne twierdzenia tzw. milowe kroki w teorii uczenia się pojęć.

Warunki pozwalające zapewnić (1 – 3) są zależne od ~~różnych~~ wartości różnych rodzajów entropii zbioru funkcji $\{Q(z,\alpha) : \alpha \in \Lambda\}$.

$\{Q(z,\alpha) : \alpha \in \Lambda\}$ $\qquad Q(z,\alpha) = L(\omega, \Phi(x,\alpha)) \in \{0, 1\}$

$z = (x, \omega)$

$(z_1, \dots, z_l)$ – průběha o dl. l $\qquad z_i = (x_i, \omega_i) \quad (i = 1, \dots, l)$

$$N^{\Lambda}(z_1, \dots, z_l) = \text{card}\,\{(Q(z_1,\alpha), \dots, Q(z_l,\alpha)) : \alpha \in \Lambda\} \leq 2^l$$

$$H^{\Lambda}(z_1, \dots, z_l) = \ln N^{\Lambda}(z_1, \dots, z_l)$$

$$\boxed{H^{\Lambda}(l) = \int H^{\Lambda}(z_1, \dots, z_l)\, dP(z_1, \dots, z_l)}$$

entropie VC

Vapnik $\qquad$ Chervonenkis

Tw. (Pierwszy kamień milowy) $\{Q(z,\alpha) : \alpha \in \Lambda\}$; $Q(z,\alpha) \in \{0,1\}$

Następujące warunki są równoważne:

1. $$\lim_{l \to \infty} \frac{H^{\Lambda}(l)}{l} = 0$$

2. $$P\left\{(z_1,\ldots,z_l) : \sup_{\alpha \in \Lambda} \left| \int Q(z,\alpha) P(z) - \frac{1}{l}\sum_{i=1}^{l} Q(z_i,\alpha) \right| > \varepsilon \right\}$$

$$\xrightarrow[l \to \infty]{} 0$$; dla każdego $\varepsilon > 0$

$$\forall \varepsilon, \delta \ \exists\, l(\varepsilon, \delta, \Lambda, P) \ \forall\, l > l(\varepsilon, \delta, \Lambda, P) \quad P\{\ldots\} < \delta$$

Entropia "annealing":

$$H^{\Lambda}_{ann}(\ell) = \ln E N^{\Lambda}(z_1,\dots,z_\ell) =$$

$$= \ln \int N^{\Lambda}(z_1,\dots,z_\ell)\, dP(z_1,\dots,z_\ell)$$

$$H^{\Lambda}(\ell) \leq H^{\Lambda}_{ann}(\ell)$$

z nierówności Jensena: jeśli ψ wklęsła to $\int \psi(\Phi(x))\,dP(x) \leq \psi\left(\int \Phi(x)\,dP(x)\right)$

"Drugi kamień milowy"

Tw. Dla ustalonego P, Λ. Jeśli $\lim\limits_{\ell\to\infty} \frac{H^{\Lambda}(\ell)}{\ell} = 0$ to

$$\forall \varepsilon > 0\ \exists b, c > 0\ \exists \ell(\varepsilon, \Lambda, P)\ \forall \ell > \ell(\varepsilon, \Lambda, P)$$

$$P\left\{(z_1,\dots,z_\ell) : \sup_{\alpha\in\Lambda} \left| \int Q(z,\alpha)\,dP(z) - \frac{1}{\ell}\sum_{i=1}^{\ell} Q(z_i,\alpha) \right| > \varepsilon \right\} < \frac{b}{e^{c\varepsilon^2 \ell}}$$

Uwaga. Tylko warunek konieczny!

Funkcja wzrostu

$$G^{\Lambda}(\ell) = \ln \sup_{(z_1,\ldots,z_\ell)} N^{\Lambda}(z_1,\ldots,z_\ell)$$

$$H^{\Lambda}(\ell) \leqslant H^{\Lambda}_{ann}(\ell) \underset{\text{oczywiste}}{\leqslant} G^{\Lambda}(\ell)$$

(Trzeci kamień milowy)

Tw. Λ - ustalone; $\mathcal{P}_0$

Następujące warunki są równoważne:

1. $$\lim_{\ell \to \infty} \frac{G^{\Lambda}(\ell)}{\ell} = 0$$

2. $\forall \varepsilon > 0 \ \exists b, c > 0 \ \exists \ell(\varepsilon, \Lambda) \ \forall \ell > \ell(\varepsilon, \eta)$

$$\sup_{P \in \mathcal{P}_0} P\left\{(z_1,\ldots,z_\ell) : \sup_{\alpha \in \Lambda} \left| \int Q(z,\alpha)\,dP(z) - \frac{1}{\ell}\sum_{i=1}^{\ell} Q(z_i,\alpha)\right| > \varepsilon \right\} < \frac{b}{e^{c\varepsilon^2 \ell}}$$

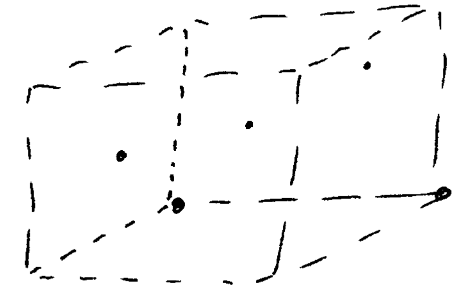
Entropia dla zbioru funkcji o wartościach rzeczywistych

$\{Q(z,\alpha) : \alpha \in \Lambda\}$ Załóżmy, że funkcje te są wspólnie ograniczone tzn. dla pewnych stałych A, B $\quad A \leq Q(z,\alpha) \leq B$ dla $\alpha \in \Lambda$

$(z_1, \dots, z_l)$ – próbka

$q(\alpha) = (Q(z_1,\alpha), \dots, Q(z_l,\alpha))$ dla $\alpha \in \Lambda$

$\{q(\alpha) : \alpha \in \Lambda\}$ jest podzbiorem kostki l-wymiarowej, której krawędzie mają długość $|B-A|$



Dla tego zbioru istnieje minimalna ε-sieć w metryce C. Niech $N = N^\Lambda(\varepsilon; z_1, \dots, z_l)$ będzie liczbą elementów minimalnej ε-sieci dla $\{q(\alpha) : \alpha \in \Lambda\}$.

$$H^\Lambda(\varepsilon; z_1, \dots, z_l) = \ln N^\Lambda(\varepsilon; z_1, \dots, z_l)$$

losowa VC-entropia zbioru $\{Q(z,\alpha) : \alpha \in \Lambda\}$ dla $(z_1, \dots, z_l)$

$$H^\Lambda(\varepsilon; l) = \int H^\Lambda(\varepsilon, z_1, \dots, z_l)\, d\overbrace{P(z_1, \dots, z_l)}^{= P(z_1)\cdot \dots \cdot P(z_l)}$$

VC-entropia zbioru $\{Q(z,\alpha) : \alpha \in \Lambda\}$ na próbkach o długości l.

Zbiór $\{q(\alpha) : \alpha \in \Lambda\}$ ma minimalną ε-sieć $\cancel{q(\alpha_1), \dots, q(\alpha_N)}$ jeśli:

1. Istnieje $N = N^\Lambda(\varepsilon, z_1, \dots, z_l)$ wektorów $q(\alpha_1), \dots, q(\alpha_N)$ takich, że dla każdego $q(\alpha^*)$ dla $\alpha^* \in \Lambda$ istnieje r $(1 \leq r \leq N)$ $q(\alpha_r)$ który jest ε-bliski do tego wektora $q(\alpha^*)$ tzn. $C(q(\alpha^*), q(\alpha_r)) = \max\limits_{1 \leq i \leq l} |Q(z_i,\alpha^*) - Q(z_i,\alpha_r)| \leq \varepsilon$
2. N jest ~~każdą~~ minimalną liczbą wektorów o powyższej własności.

$q(\alpha_1), \dots, q(\alpha_N)$ – nazywamy wtedy minimalną ε-siecią dla $\{q(\alpha) : \alpha \in \Lambda\}$.

Tw. Jednostajne dwustronne zbieżność ryzyka empirycznego do ryzyka faktycznego

tzn.

$$\lim_{l\to\infty} P\left\{(z_1,\dots,z_l):\ \sup_{\alpha\in\Lambda} |R(\alpha)-R_{emp}(\alpha)|>\varepsilon\right\}=0 \quad \text{dla każdego } \varepsilon$$

~~jest nowelnetowaniu~~ ma miejsce wtedy i tylko wtedy

$$(*)\qquad \lim_{l\to\infty} \frac{H^{\Lambda}(\varepsilon;l)}{l}=0 \quad \text{dla każdego } \varepsilon.$$

Uwaga. Niewielka modyfikacja warunku (*) pozwala otrzymać warunek konieczny i dostateczny na to by miała miejsce jednostajna jednostronna zbieżność ryzyka empirycznego do ryzyka faktycznego tzn.

$$\lim_{l\to\infty} P\left\{(z_1,\dots,z_l):\ \sup_{\alpha\in\Lambda} \left(R(\alpha)-R_{emp}(\alpha)\right)>\varepsilon\right\}=0 \quad \text{dla każdego } \varepsilon.$$

$$\text{VC-dim}(H)\ ;\ H = \{Q(z,\alpha) : \alpha \in \Lambda\}$$

1. $N^{\Lambda}(z_1,\dots,z_l) = \operatorname{card}\{(Q(z_1,\alpha),\dots,Q(z_l,\alpha)) : \alpha \in \Lambda\}$

2. $G(l) = \sup\limits_{(z_1,\dots,z_l)} \ln N^{\Lambda}(z_1,\dots,z_l)$

3. $\text{VC-dim}(H) = \begin{cases} \max\{l : G(l) = l \ln 2\} & \text{jeśli } \exists l \ G(l) \neq l \ln 2 \\ +\infty & \text{w. pp.} \end{cases}$

<u>Uwaga</u> Jeśli $N^{\Lambda}(z_1,\dots,z_l) = 2^l$ to mówimy, że $\{Q(z,\alpha) : \alpha \in \Lambda\}$ <u>szatkuje</u> (rozbija) $(z_1,\dots,z_l)$.

Tw. Funkcja wzrostu $G(l)$ dla $\{Q(z,\alpha): \alpha \in \Lambda\}$; $Q(z,\alpha) \in \{0,1\}$

spełnia warunek

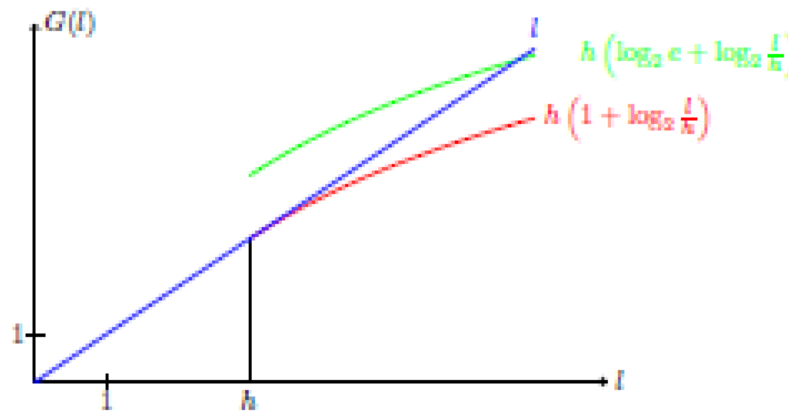
$$G^{\Lambda}(l) = l \ln 2$$

albo jest ograniczona

$$G^{\Lambda}(l) \begin{cases} = l \ln 2 & \text{dla } l \leq h \\ \leq \ln\left(\sum_{i=0}^{h} \binom{l}{i}\right) \leq h\left(1 + \ln \frac{l}{h}\right) & \text{dla } l > h \end{cases}$$

h jest największą taką liczbą, że $G^{\Lambda}(l) = l \ln 2$

h nazywamy wymiarem Vapnika-Chervonenkisa dla Λ $(\{Q(z,\alpha): \alpha \in \Lambda\})$



Dziękuję za rys. przygotowany przez Pana Grzegorza Bokotę korygujący poprzednią wersję

Tw. Funkcja wzrostu $G(l)$ dla $\{Q(z,\alpha) : \alpha \in \Lambda\}$; $Q(z,\alpha) \in \{0,1\}$

spełnia warunek

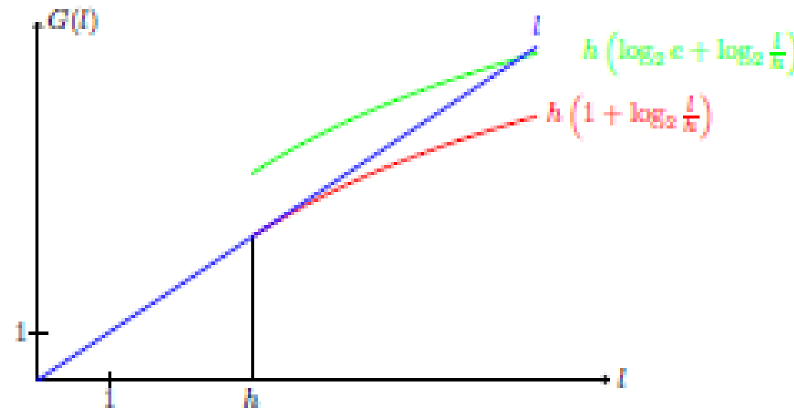
$$G^{\Lambda}(l) = l \ln 2$$

albo jest ograniczona

$$G^{\Lambda}(l) \begin{cases} = l \ln 2 & \text{dla } l \leq h \\ \leq \ln\left(\sum_{i=0}^{h} \binom{l}{i}\right) \leq h\left(1 + \ln \frac{l}{h}\right) & \text{dla } l > h \end{cases}$$

h jest największą taką liczbą, że $G^{\Lambda}(l) = l \ln 2$

h nazywamy wymiarem Vapnika-Chervonenkisa dla Λ $(\{Q(z,\alpha): \alpha \in \Lambda\})$



Dziękuję za rys. przygotowany przez Pana Grzegorz Bokotę korygujący poprzednia wersję

<u>Příklad</u>

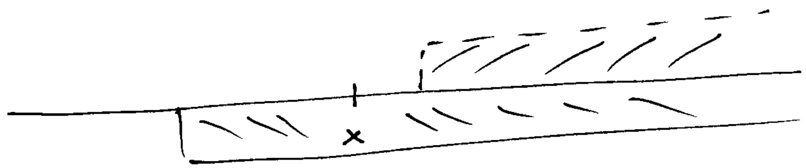
$\Omega = R$

$$Q(x, \alpha) = \begin{cases} 1 & x \geq \alpha \\ 0 & x < 0 \end{cases}$$

$H = \{ Q(x, \alpha) : \alpha \in R \}$

$\dim_{VC}(H) = 1$

$l = 1$:



$l = 2$: $\quad \mathrm{card}\{ (Q(x_1, \alpha), Q(x_2, \alpha)) : \alpha \in R \} = \mathrm{card}\{00, 11, 01\} < 2^2$

$$H = H_\Lambda = \{ Q(z,\alpha) : \alpha \in \Lambda \} \;; \quad Q(z,\alpha) \in \{0,1\} \;; \quad z \in Z$$

$$\mathcal{H} = \{ Z_\alpha \subseteq Z : \alpha \in \Lambda \} \quad \text{gdzie} \quad Z_\alpha = \{ z \in Z : Q(z,\alpha) = 1 \}$$

$X \subseteq Z$ jest "rozbijany" przez H $(\mathcal{H})$ wtt

$$\forall Y \subseteq X \;\; \exists \alpha \in \Lambda \;\; (Y = Z_\alpha \cap X)$$

"każdy podzbiór X może być wybrany przez H $(\mathcal{H})$"

$z = (z_1, \dots, z_\ell)$ (próbka o wyjściach powiedzmy różnych)

$q(\alpha, z) = (Q(z_1,\alpha), \dots, Q(z_\ell,\alpha))$

$N^\Lambda(z) = \operatorname{card}\{ q(\alpha, z) : \alpha \in \Lambda \} = 2^\ell$ wtt zbiór $\{z_1, \dots, z_\ell\}$ jest rozbijany przez H $(\mathcal{H})$

$$G^\circ(\ell) \stackrel{\downarrow}{=} G^\circ_H(\ell) = \sup_{(z_1,\dots,z_\ell)} N^\Lambda(z_1,\dots,z_\ell)$$

$= G^\circ(\ell, H)$

$$G(\ell) = G_H(\ell) = \ln \sup_{(z_1,\dots,z_\ell)} N(z_1,\dots,z_\ell)$$

$$G^{\bullet}(\ell) = G^{\bullet}_{H}(\ell) = \sup_{(z_1 \dots z_\ell)} N^{1}(z_1 \dots z_\ell)$$

Zamiast $G^{\circ}_{H}(\ell)$ piszemy też $G^{\bullet}(\ell, H)$

Przykład

$$H = \{[a, +\infty) : a \in R\}$$

$\underline{\dim_{VC}(H) = 1}$

$$G^{\bullet}(n) = n + 1 \qquad\qquad G^{\bullet}(n) = \binom{n}{0} + \binom{n}{1}.$$

$$G^{\bullet}(1) = 2^1$$

$$G^{\bullet}(2) = 3 < 2^2$$

Przykład

$$H = \{(a, b) : a < b,\ a, b \in R\}$$

$\underline{\dim_{VC}(H) = 2}$

$$G^{\bullet}(n) = \binom{n+1}{2} + 1 \qquad\qquad G^{\bullet}(n) = \binom{n}{0} + \binom{n}{1} + \binom{n}{2}$$

$$G^{\bullet}(1) = 2^1$$

$$G^{\bullet}(2) = 2^2$$

$$G^{\bullet}(3) = 7 < 2^3$$

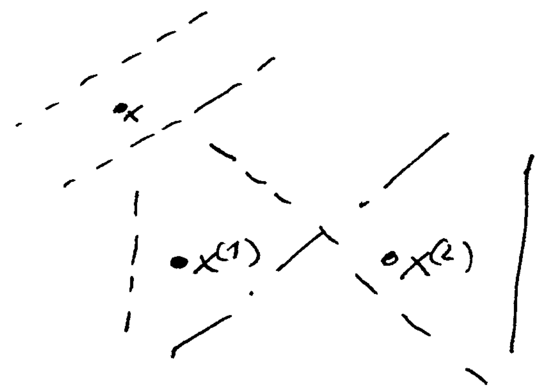
Przykład 2

$\Lambda = R^3$, $\alpha = (a, b, c) \in R^3$; $x = (x_1, x_2) \in R^2$

$$Q(x, \alpha) = \begin{cases} 1 & \text{jeśli } ax_1 + bx_2 + c \geq 0 \\ 0 & \text{w p.p.} \end{cases}$$

$$H = \{ Q(x,\alpha) : \alpha \in R^3 \}$$

$$\text{VC-dim}(H) = 3$$

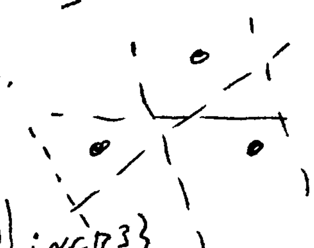


$l = 1$ $\quad \text{card}\{(Q(x,\alpha)) : \alpha \in R^3\} = \text{card}\{0, 1\} = 2^1$

$l = 2$ $\quad \text{card}\{(Q(x^{(1)},\alpha), Q(x^{(2)},\alpha)) : \alpha \in R^3\} = 2^2$

$l = 3$ $\quad \text{card}\{(Q(x^{(1)},\alpha), Q(x^{(2)},\alpha), Q(x^{(3)},\alpha)) : \alpha \in R^3\} = 2^3$

$x^{(1)}, x^{(2)}, x^{(3)}$ nie są współliniowe np.



$l = 4$ $\quad \text{card}\{(Q(x^{(1)},\alpha), Q(x^{(2)},\alpha), Q(x^{(3)},\alpha), Q(x^{(4)},\alpha)) : \alpha \in R^3\}$
$< 2^4$

$$\text{VC-dim}(H) = 3$$

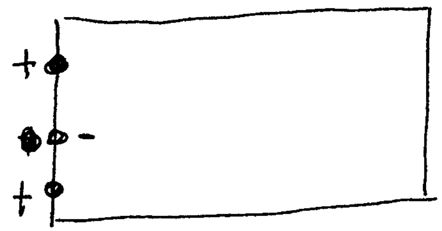
Przykład 3

Prostokąty || do osi.

Dla $l = 1, 2, 3, 4$ $\quad G(l) = l \ln 2$.

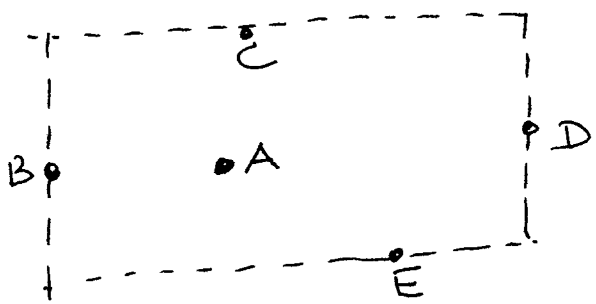
$l = 5$

- jeśli 3 punkty są współliniowe to $\quad N^{-1}(z_1, \dots, z_5) < 2^l$



Nie umiemy prostokąt taki, by punkt był na zewnątrz a + + wewnątrz

- jeśli żadne z 3 punktów nie są współliniowe to z pięciu punktów wybieramy 4, których współrzędne osiągają max, min na pierwszej i drugiej współrzędnej; powstaje prostokąt. Nie umiemy prostokąt taki by A był na zewnątrz a B, C, D, E wewnątrz.

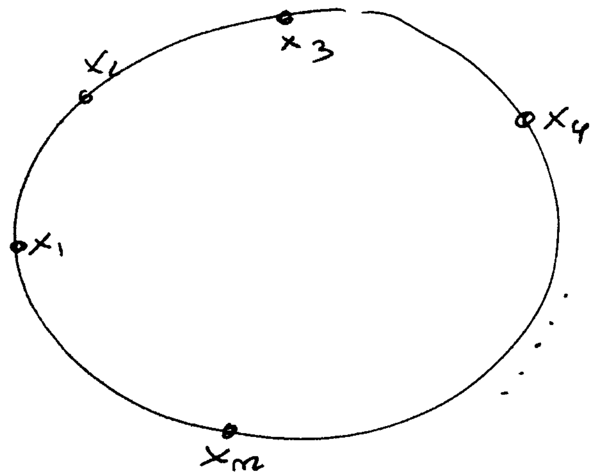


Przykład

H – zbiór ~~obszarów~~ hipotez wyznaczonych przez wielokąty na płaszczyźnie

$\dim_{VC}(H) = +\infty$

Rozważmy n punktów na okręgu : $x_1 \dots x_n$



Każdy podzbiór tego zbioru punktów należy do obszaru wyznaczonego przez wielokąt o wierzchołkach w tych punktach. (pozostałe punkty znajdą się na zewnątrz tego obszaru).

$$\Phi(x,\alpha) = \Theta\left(\sum_{k=1}^{n} \alpha^k x^k\right)$$
$\alpha = (\alpha^1, \dots, \alpha^n)$
$x = (x^1, \dots, x^k)$

$$\Theta(t) = \begin{cases} 1 & t > 0 \\ \cancel{0}\,-1 & t \le 0 \end{cases}$$

zbiór decyzji $1, -1$

Przykład 4

H : hiperpłaszczyzny przechodzące przez O : $\sum_{k=1}^{n} \alpha^k u_k = 0$

$Q(z,\alpha) = \cancel{\Theta\left(\sum_{k=1}^{n} \alpha^k u_k\right)}$ ~~gdzie Θ - funkcja progowa~~

~~$\alpha = (\alpha^1, \dots, \alpha^n)$~~

$$Q(z,\alpha) = L(\underset{\{-1,1\}}{\omega}, \Phi(x,\alpha)) \in \{0,1\}$$

VC-dim(H) = n

Łatwo zauważyć, że zbiór n punktów ~~odpowiadających~~ ~~wektorom jednostkowym~~ $(0 \dots \underset{i}{1} \dots 0)$ może być "poszatkowany" przez H.

Pokażemy, że nie istnieje $n+1$ punktów takich, że H rozbije zbioru tych punktów.

Przypuśćmy, że istnieją takie punkty $u_1, \dots, u_{n+1}$, że H rozbije zbiór $\{u_1, \dots, u_{n+1}\}$.

Wtedy istniałyby 2^{n+1} wektorów $\alpha_i \in R^n$ $(i = 1, \dots, 2^{n+1})$ takich, że macierz (z_{ij}) gdzie $z_{ij} = u_i^T \alpha_j$ ma kolumny z dowolnym rozkładem znaków. Wobec tego macierz

$$A = \begin{pmatrix} z_{1,1} & \dots & z_{1,2^{n+1}} \\ \vdots & & \vdots \\ z_{n+1,1} & \dots & z_{n+1,2^{n+1}} \end{pmatrix}$$

której kolumny mają 2^{n+1} możliwych rozkładów znaków

$$\operatorname{sign}(A) = \begin{pmatrix} - & - & \dots & + \\ - & \vdots & & \\ \vdots & & & \\ - & + & \dots & + \end{pmatrix}$$

Wtedy wiersze ~~kolumny~~ A byłyby liniowo niezależne, bo dla każdych stałych nie równych jednocześnie 0 $c_1, \dots, c_{n+1}$ takie, że mamy $c_1 \begin{pmatrix} z_{1,1} \\ \vdots \\ z_{1,2^{n+1}} \end{pmatrix} + \dots + c_{n+1} \begin{pmatrix} z_{n+1,1} \\ \vdots \\ z_{n+1,2^{n+1}} \end{pmatrix} \neq 0$. Przypuśćmy, że nie.

~~Dlaczego~~ ~~Bo~~ Istnieje i takie, że $(z_{1,i}, \dots, z_{n+1,i})$ ma ten sam rozkład znaków co $(c_1, \dots, c_{n+1})$ a więc $\sum_{j=1}^{n+1} c_j z_{ji} > 0$ $(\neq 0)$. Sprzeczność: w R^n istniałoby n+1 wektorów liniowo niezależnych.

$$G^{\bullet}_{H}(\ell) = \sup_{(z_1,\dots,z_\ell)} N^{\Lambda}(z_1,\dots,z_\ell) \qquad H=\{\Phi(x,\alpha): \alpha\in\Lambda\}$$

Lemat Sauera

Niech $\dim_{VC}(H)=d<+\infty$ oraz $\ell\geq 1$. Wtedy $G^{\bullet}_{H}(\ell)\leq 1+\binom{\ell}{1}+\dots+\binom{\ell}{d}$.

Uwaga. $\binom{\ell}{d}=0$ dla $d>\ell\geq 1$.

Dla $d\geq \ell$ lemat oznacza, że $G^{\bullet}_{H}(\ell)\leq 1+\binom{\ell}{1}+\dots+\binom{\ell}{\ell}+0+\dots+0=2^{\ell}$

co wynika z definicji.

Niech $\Phi(d,\ell)=1+\binom{\ell}{1}+\dots+\binom{\ell}{d}$. $\Phi(d,\ell)<2^{\ell}$ dla $\ell>d$.

Dla $\ell>d$ $\Phi(d,\ell)$ jest wielomianem stopnia d względem ℓ.

$\Phi(0,\ell)=1$ dla $\ell\geq 1$

$\Phi(d,1)=2$ dla $d\geq 1$

-1 $\quad \Phi(d,\ell-1)+\Phi(d-1,\ell-1) = 1+\binom{\ell-1}{1}+\dots+\binom{\ell-1}{d}+ \qquad \binom{a}{b}=\binom{a-1}{b}+\binom{a-1}{b-1}$

$$1+\binom{\ell-1}{1}+\dots+\binom{\ell-1}{d-1}$$

$$\overline{\quad 1+\binom{\ell}{1}+\dots+\binom{\ell}{d}=\Phi(d,\ell) \quad}$$

1. $d=0$ wtedy dla każdego x i hipotezy $h\in H$ $h(x)$ ma tę samą wartość [~~0 albo 1~~] jedną z wartości jakie przyjmuje $\Phi(x,\alpha)$.
   Wtedy $G^{\bullet}_{H}(\ell)=1$ dla każdej próbki o dł. $\ell\geq 1$ $G^{\bullet}_{H}(\ell)=1\leq\Phi(0,\ell)=1$.

2. $\ell=1$ i $d\geq 1$ $\Phi(d,1)=2$
   $G^{\bullet}_{H}(\ell)=G^{\bullet}_{H}(1)\leq 2=\Phi(d,1)$

Dowód przez indukcję ze względu na $d+l$.

Jeśli $d+l=2$: $d=0,\ l=2$ – przypadek 1 } teza zachodzi.
$d=1\ \ l=1$ " 2 }

Przypuśćmy, że teza zachodzi dle wszystkich przypadków gdy $d+l\leq k$ gdzie k jest jakąś liczbą ≥ 2.

Niech ~~...~~ $\dim_{VC}(H)=d$ i niech $\mathbf{z}$ będzie próbką o dł. l takiej, że $d+l=k+1$. Przypadki $(0,k+1)$, $(\cancel{k},1)$ są pokryte przez warunki brzegowe. Możemy więc założyć, że $d\geq 1$ i $l\geq 2$.

Można założyć, że jeśli $\mathbf{z}=(x_1,\omega_1),\ldots,(x_l,\omega_l)$ to ciąg $x_1,\ldots,x_l$ (z $z_1 \ldots z_l$) nie zawiera identycznych wyrazów [inaczej musimy ~~odrzucać powtarzające się wyrazy~~ pozostawić tylko jedno wystąpienie każdego wyrazu i otrzymamy krótszy ciąg i teza wynika z zał. indukcyjnego].

Niech $E=\{z_1,\ldots,z_l\}$ i niech $H_E=H|E=\{h|E : h\in H\}$. Zbiór $H|E$ jest skończony i $N^{-1}(z_1,\ldots,z_l)=|H_E|$. Pokażemy, że $|H_E|\leq\Phi(d,l)$.

$F=E-\{z_l\}$, $\quad H_F=H|F$

$$|H_E|=|H_F|+|H_*|$$

, gdzie H_* to zbiór hipotez h z H_F które powstają z $h_1,h_2\in H_E$ takich, że $h_1|F=h_2|F=h$ oraz $h_1(z_l)\neq h_2(z_l)$.

$\mathbf{z}'=(z_1,\ldots,z_{l-1})$

$(d+(l-1)\leq k)$
z zał. ind.

$$|H_F|=N^{-1}(z_1^*,\ldots,z_{l-1}^*)\leq G_H^{\bullet}(l-1)\leq\Phi(d,l-1)$$

Mamy $\dim_{VC}(H_*)=d-1$. W przeciwnym przypadku istniałaby próbka $\mathbf{z}''=(z_1'',\ldots,z_d'')$ "rozbłokowana" przez H_* ale wtedy H rozbłokowałby $(z_1'',\ldots,z_d'',\cancel{z_l})$ ~~...~~ co jest sprzeczne z zał. $\dim_{VC}(H)=d$.

$$(d-1)+(m-1)\le k$$

Wobec tego

$$|H_*| = N^1_{H^*}(z^*_1,\dots,z^*_{l-1}) \leqslant G_{H^*}(l-1) \overset{\text{z zał. ind.}}{\leqslant} \Phi(d-1,l-1)$$

Ostatecznie ~~Vlex~~

$$N^1_H(z_1,\dots,z_l) = |H_E| = |H_F| + |H_*| \leqslant$$

$$\leqslant \Phi(d,l-1) + \Phi(d-1,l-1) = \Phi(d,l).$$

<u>Wniosek</u>

1. $G^\circ_H(l) \leqslant (l+1)^d$ jeśli $\dim_{VC}(H) = d < +\infty$ oraz l – dowolne

<u>Dowód</u>

$$\sum_{i=0}^{d}\binom{l}{i} \leqslant \sum_{i=0}^{d}\frac{l^i}{i!} \leqslant \sum_{i=0}^{d}\frac{l^i\, d!}{i!\,(d-i)!} = (l+1)^d$$

2. $G^\circ_H(l) \leqslant \left(\frac{ne}{d}\right)^d$ jeśli $\dim_{VC}(H) = d < +\infty$ oraz $l \geqslant d$.

$$\left(\frac{d}{l}\right)^d \sum_{i=0}^{d}\binom{l}{i} \leqslant \sum_{i=0}^{d}\left(\frac{d}{l}\right)^i\binom{l}{i} \leqslant \sum_{i=0}^{l}\left(\frac{d}{l}\right)^i\binom{l}{i} = \left(1+\frac{d}{l}\right)^l \leqslant e^d.$$

<u>Uwaga</u> Jeśli $l \geqslant d = \dim_{VC}(H)$ to $G_H(l) = \ln G^*_H(l) \leqslant d \ln \frac{le}{d} =$

$$= d\left(1 + \ln\frac{l}{d}\right)$$

Structural risk minimization (SRM)

Tw. $\{Q(z,\alpha) : \alpha \in \Lambda\}$; $0 \leq Q(z,\alpha) \leq B$ dla $\alpha \in \Lambda$ $(*)$
wymiar VC równy h.

Nierówność

$$R(\alpha) \leq R_{emp}(\alpha) + \frac{B\varepsilon}{2}\left(1 + \sqrt{1 + \frac{4R_{emp}(\alpha)}{B\varepsilon}}\right)$$

zachodzi z prawdopodobieństwem $\geq 1-\eta$ gdzie $\varepsilon = 4\,\dfrac{h\left(\ln\frac{2l}{h}+1\right)-\ln\eta}{l}$
dla wszystkich funkcji z $(*)$

Dla $\frac{l}{h}$ "dużego" drugi składnik jest zaniedbywalny i faktyczne ryzyko $R(\alpha)$ jest bliskie do ryzyka empirycznego $R_{emp}(\alpha)$. Jeśli jednak $\frac{l}{h}$ jest "małe" to ~~drugiego~~ drugiego składnika nie można pominąć i optymalizacja $R_{emp}(\alpha)$ nie gwarantuje małej wartości ryzyka. Należy minimalizować oba składniki. SMR ma na celu minimalizację funkcjonału ryzyka ze względu na ryzyko empiryczne jak i ze względu na wymiar VC zbioru funkcji.

Załóżmy, że dla $S = \{Q(z,\alpha) : \alpha \in \Lambda\}$ określona jest struktura tzn. rodzina $\{S_k\}$, $S_k = \{Q(z,\alpha) : \alpha \in \Lambda_k\}$ podzbiorów S takie, że $S_1 \subset S_2 \subset \ldots \subset S_k \subset \ldots$ oraz $S^* = \bigcup S_i$. Struktura ta jest dopuszczalna jeśli

1. S^* jest wszędzie gęsty w S
2. wymiar VC dla każdego zbioru S_k jest skończony (oznaczamy go przez h_k)
3. ~~elementy~~ S_k zawierają funkcje wspólnie ograniczone tzn. dla pewnego B_k : $0 \leq Q(z,\alpha) \leq B_k$ dla $\alpha \in \Lambda_k$

Zasada SRM

- dla zadanej próbki $(z_1, .., z_l)$ [a więc l ustalone!] należy wybrać S_n gdzie $n = n(l)$ oraz funkcję z S_n która minimalizuje prawą stronę nierówności w twierdzeniu.

Jeśli $n \nearrow$ to $R_{emp}(\alpha) \searrow$

$n \nearrow$ to drugi składnik w oszacowaniu nierówności $\nearrow$

Wobec tego SRM proponuje pewien trade-off między jakością aproksymacji i złożonością funkcji.

Tw. Dla każdego rozkładu prawdopodobieństwa metoda SRM zapewnia zbieżność do rozwiązania najlepszego z prawdopodobieństwem 1.

Tw. SRM dla struktur dopuszczalnych pozwala wyznaczyć $Q(z, \alpha_l^{n(l)})$ dla których ciąg wartości ryzyka $R(\alpha_l^{n(l)})$ zbiega do $R(\alpha_{opt})$ <u>asymptotycznie</u> <u>z szybkością</u>

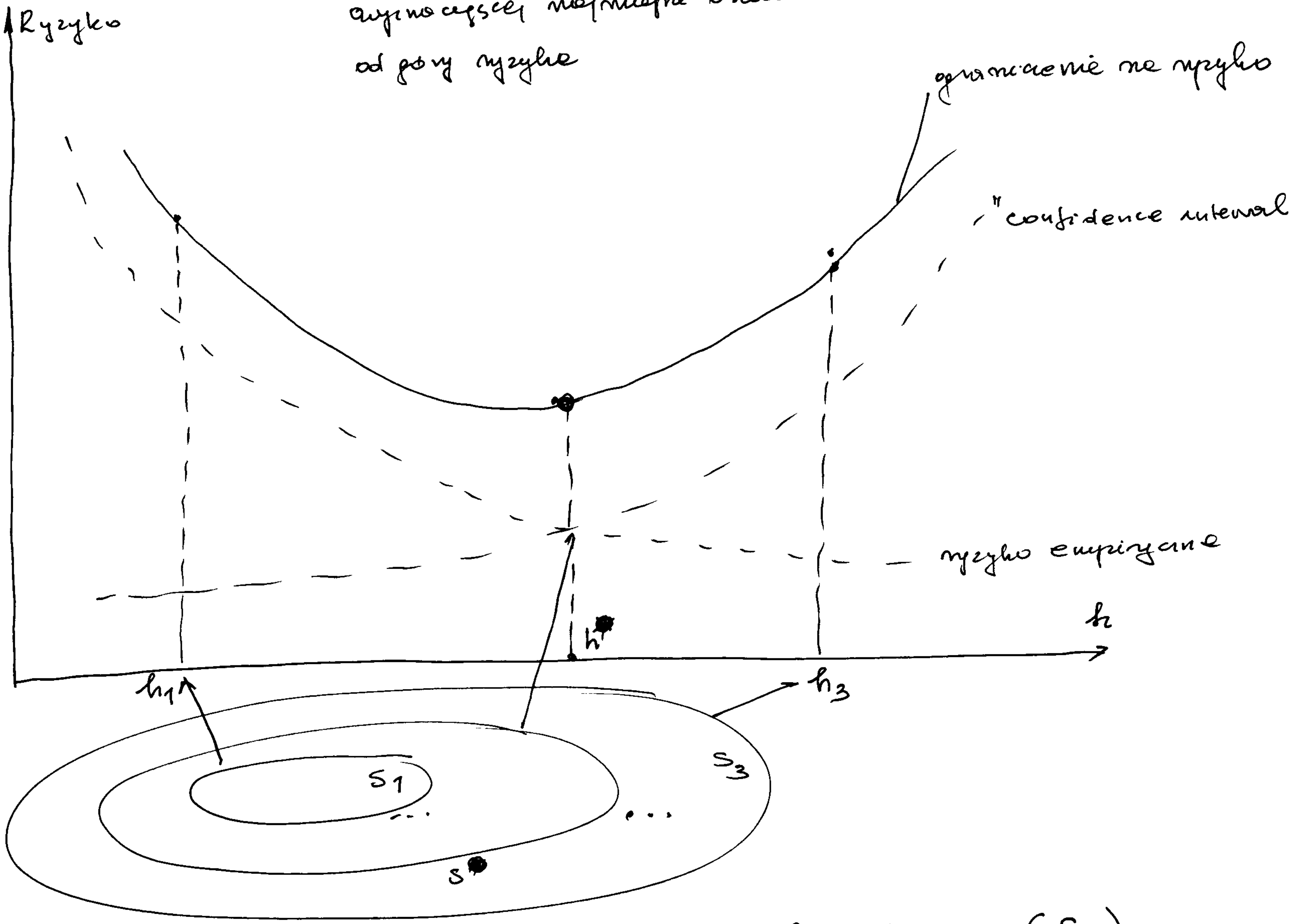
$$V(l) = r_{n(l)} + B_{n(l)}\sqrt{\frac{h_{n(l)} \ln l}{l}} \qquad (**)$$

jeśli $n = n(l)$ jest takie, że $\lim\limits_{l\to\infty} \frac{B^2_{n(l)}\, h_{n(l)} \ln l}{l} = 0$.

W (**) B_n jest ograniczeniem dla funkcji z S_n a $r_n(l)$ jest szybkością aproksymacji: $r_n = \inf\limits_{\alpha\in\Lambda_n} Q(z,\alpha)\,dP(z) - \inf\limits_{\alpha\in\Lambda} \int Q(z,\alpha)\,dP(z)$.

zmienne losowe ξ_l, $l=1,\dots$ zbiegają do ξ_0 asymptotycznie z szybkością $V(l)$ jeśli istnieje stała C taka, że $V^{-1}(l)\,|\xi_l - \xi_0| \xrightarrow[l\to\infty]{D} C$.

poszukiwanie "warstwy" S^* wyznaczającej najmniejsze oszacowanie od góry ryzyka



$h_1 = \dim_{VC}(S_1)$

$h^* = \dim_{VC}(S^*)$

$h_3 = \dim_{VC}(S_3)$