Лекція 6. Розмірність Вапніка-Червоненкіса

Модель навчання з учителем складається з трьох компонентів: 1) векторів $\vec{x}$ з фіксованою, але невідомою функцією розподілу ймовірностей $F_X(\vec{x})$, 2) учителя, який генерує відклик $d = f(\vec{x}, v)$ відповідно до фіксованої, але теж невідомої умовної функції розподілу $F_X(\vec{x}|d)$ із урахуванням шуму v, 3) машина, що навчається, яка реалізує відображення $y = F(\vec{x}, \vec{w})$, де y — відклик, обчислений машиною у відповідь вхідний на сигнал $\vec{x}$, $\vec{w}$ — набір вільних параметрів (ваг), вибраних з простору W.

Мета навчання з учителем — вибрати конкретну функцію $F(\vec{x}, \vec{w})$, яка оптимально в статистичному розумінні апроксимує очікуваний відклик d. Цей вибір залежить від незалежних однаково розподілених прикладів навчання $T = \{(\vec{x}_i, d_i)\}, i = 1,..., N.$ Кожна пара вибирається машиною з множини T відповідно до деякої узагальненої функції розподілу ймовірностей $F_{X,D}(\vec{x}, d)$, яка теж фіксована, але невідома.

Позначимо як $L(d, F(\vec{x}, \vec{w}))$ *міру втрат* або *незбіжності* між бажаним відкликом d, що відповідає вхідному вектору $\vec{x}$, та відкликом $F(\vec{x}, \vec{w})$, обчисленим машиною. Як правило, цю функцію задають як квадрат різниці між d і $F(\vec{x}, \vec{w})$:

$$L(d, F(\vec{x}, \vec{w})) = (d - F(\vec{x}, \vec{w}))^2.$$

Очікувана величина втрат визначається *функціоналом ризику*:

$$R(\vec{w}) = \int L(d, F(\vec{x}, \vec{w}))\, dF_{X,D}(\vec{x}, d).$$

Отже, навчання з учителем зводиться до мінімізації функціонала ризику $R(\vec{w})$ в класі апроксимацій $\{F(\vec{x}, \vec{w}), \vec{w} \in W\}$. Зважаючи на те, що функція $F_{X,D}(\vec{x}, d)$ є невідомою, будемо використовувати індуктивний *принцип мінімізації емпіричного ризику*.

Функціонал емпіричного ризику має вигляд

$$R_{emp}(\vec{w}) = \frac{1}{N} \sum_{i=1}^{N} L(d_i, F(\vec{x}_i, \vec{w})).$$

Основні переваги функціонала емпіричного ризику полягають у такому:

1. Він *не залежить* явно від невідомої функції розподілу $F_{X,D}(\vec{x}, d)$
2. Його можна мінімізувати за вектором ваг $\vec{w}$.

Принцип мінімізації емпіричного ризику формулюється так:

1. Замість функціонала $R(\vec{w})$ будується функціонал емпіричного ризику $R_{emp}(\vec{w})$ на базі множини $T = \{(\vec{x}_i, d_i)\}$, $i = 1,..., N.$
2. Нехай w_{emp} — вектор вагових коефіцієнтів, що мінімізує функціонал емпіричного ризику $R_{emp}(\vec{w})$ в просторі ваг W. Тоді будемо вважати, що виконуються такі умови:

$$\forall \varepsilon > 0 \ \lim_{N \to \infty} P\left(\left| \inf_{\vec{w} \in W} R(\vec{w}_{emp}) - \inf_{\vec{w} \in W} R(\vec{w}) \right| > \varepsilon \right).$$

$$\forall \varepsilon > 0 \ \lim_{N \to \infty} P\left(\sup_{\vec{w} \in W} \left| R(\vec{w}) - R_{emp}(\vec{w}) \right| > \varepsilon \right) = 0.$$

Умови 2) означають збіжність за ймовірністю значень $R\left(\vec{w}_{emp}\right)$ до мінімуму фактичного ризику $R\left(\vec{w}\right)$ і рівномірну збіжність функціонала емпіричного ризику $R_{emp}\left(\vec{w}\right)$ до функціонала фактичного ризику $R\left(\vec{w}\right)$.

Зауваження. Практичне тлумачення принципу мінімізації емпіричного ризику полягає у тому, що при збільшенні навчальної множини правдоподібність тих функцій апроксимації, що не суперечать навчальній множині, збільшується, а точка мінімуму емпіричного функціонала прямує за ймовірністю до точки мінімуму функціонала фактичного ризику.

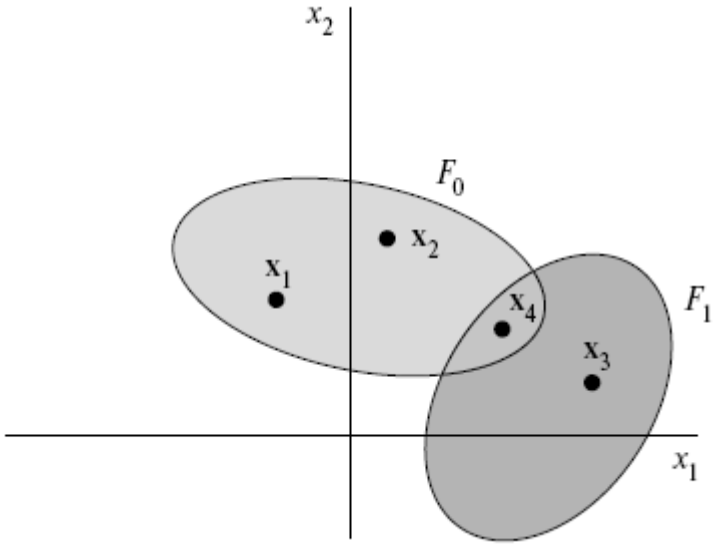
Вимір Вапніка-Червоненкіса

Розглянемо задачу бінарної класифікації: $d \in \{0,1\}$. Нехай F — множина вирішальних правил, що реалізуються машиною: $F = \left\{F\left(\vec{x},\vec{w}\right)\right\}$, а $L = \left\{x_i, i = 1,...,N\right\}$. Вирішальне правило розбиває множину L на дві частини L_0 і L_1, що не перетинаються:

$$F\left(\vec{x},\vec{w}\right) = \begin{cases} 0, x \in L_0, \\ 1, x \in L_1. \end{cases}$$

Позначимо як $\Delta_F\left(L\right)$ кількість вирішальних правил, що реалізуються машиною, $\Delta_F\left(l\right)$— максимум $\Delta_F\left(L\right)$ на множині L, для яких $L = \left|l\right|$, де $\left|L\right|$ — кількість елементів в L. Сукупність вирішальних правил F називається *розбиттям* множини L, якщо $\Delta_F\left(L\right) = 2^{|L|}$, тобто функції з F можуть реалізувати усі можливі вирішальні правила. Функція $\Delta_F\left(l\right)$ називається *функцією зростання*.

Приклад 1. Розглянемо двовимірний простір X і навчальну вибірку, що складається з точок x_1, x_2, x_3, x_4.



Приклад 1 (см. Хайкин, Нейронные сети, стр. 147).

Вирішальні функції F_0 і F_1 відповідають гіпотезам 0 і 1 і породжують множини $D_0 = \{G_0 = \{x_1, x_2, x_4\}, G_1 = \{x_3\}\}$ і $D_1 = \{G_0 = \{x_1, x_2\}, G_1 = \{x_3, x_4\}\}$ відповідно. Оскільки $|G| = 4$ і $\Delta_F(G) = 2^4 = 16.$

Означення. Виміром Вапніка-Червоненкіса (VC-dimension) множини вирішальних правил F називається потужність найбільшої множини L, розбиттям якої є множина F. Якщо така потужність може бути скільки завгодно великою, вимір VC вважається нескінченим.

Інакше кажучи, $VC(F)$ — це максимальне число образів, на яких машина може бути навчена без помилок для всіх можливих бінарних маркувань.

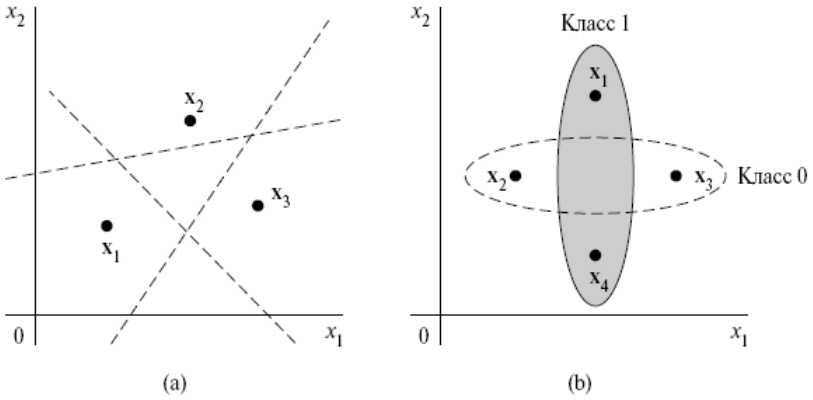
Приклад 2. Розглянемо *m*-вимірний простір X і вирішальне правило

$$F : y = \varphi(\vec{w}^T \vec{x} + b),$$

де w — *m*-вимірний вектор ваг, а b — поріг. Функція φ визначається як порогова:

$$\varphi(v) = \begin{cases} 1, v \geq 0, \\ 0, v < 0. \end{cases}$$

Розмірність VC сукупності порогових вирішальних правил дорівнює $VC(F) = m + 1$.



Приклад 2 (см. Хайкин, Нейронные сети, стр. 148).

Як бачимо на рис. (a), очки можна відокремити лініями трьома способами, а на рис. (b) — чотири точки лініями уже не розділяються, тому $VC(F) = 2 + 1 = 3.$

Зв'язок VC і статистичної теорії навчання

Практичне значення VC-виміру полягає в тому, що кількість прикладів, необхідних для навчання розпізнавальної системи, пропорційна VC-виміру. Оскільки точне значення VC отримати складно, цей показник часто замінюють оцінкою. Зокрема,

1) для довільної нейронної мережі прямого розповсюдження із пороговою функцію активації Хевісайда показник VC складає $O(WlogW)$, де W — загальна кількість вільних параметрів мережі;
2) для довільної нейронної мережі прямого розповсюдження із сігмоїдальною функцію активації показник VC складає $O(W^2)$, де W — загальна кількість вільних параметрів мережі.

Взагалі, багатошарові мережі прямого розповсюдження мають *скінченний* показник VC.

Розглянемо бінарну класифікацію. В такому випадку функція втрат має вигляд:

$$L\left(d, F\left(\vec{x}, \vec{w}\right)\right) = \begin{cases} 0, F\left(\vec{x}, \vec{w}\right) = d, \\ 1, F\left(\vec{x}, \vec{w}\right) \neq d. \end{cases}$$

За цих умов функціонал ризику $R\left(\vec{w}\right)$ тлумачиться як *помилка класифікації* $P\left(\vec{w}\right)$, а функціонал емпіричного ризику $R_{emp}\left(\vec{w}\right)$ — як *помилка навчання* (частотою помилок в процесі навчання) $v\left(\vec{w}\right)$.

За принципом мінімізації емпіричного ризику

$$\forall \varepsilon > 0 \quad \lim_{N\to\infty} P\left(\left|P\left(\vec{w}\right) - v\left(\vec{w}\right)\right| > \varepsilon\right) = 0,$$

$$\forall \varepsilon > 0 \quad \lim_{N\to\infty} P\left(\sup_{\vec{w}\in W}\left|P\left(\vec{w}\right) - v\left(\vec{w}\right)\right| > \varepsilon\right) = 0.$$

Показник VC накладає обмеження на швидкість рівномірної збіжності. Зокрема, якщо покласти VC = *h*,

$$P\left(\sup_{\vec{w}\in W}\left|P\left(\vec{w}\right) - v\left(\vec{w}\right)\right| > \varepsilon\right) < \left(\frac{2eN}{h}\right)^h e^{-\varepsilon^2 N}.$$

Нехай

$$\alpha = P\left(\sup_{\vec{w}}\left|P\left(\vec{w}\right) - v\left(\vec{w}\right)\right| \geq \varepsilon\right).$$

Тоді

$$P\left(P\left(\vec{w}\right) < v\left(\vec{w}\right) + \varepsilon\right) = 1 - \alpha,$$

звідки

$$\alpha = \left(\frac{2eN}{h}\right)^h e^{-\varepsilon^2 N}.$$

Якщо $\varepsilon_0\left(N, h, \alpha\right)$ — значення ε, що задовольняє попередню нерівність, то $\varepsilon_0\left(N, h, \alpha\right) = \sqrt{\frac{h}{N}\left[\log\frac{2N}{h} + 1\right]} - \frac{1}{N}\log\alpha$ — верхня межа довірчого інтервалу для помилки класифікації. Вапнік і Червоненкіс ввели в розгляд величину

$$\varepsilon_1\left(N, h, \alpha, v\right) = 2\varepsilon_0^2\left(N, h, \alpha\right)\left(1 + \sqrt{1 + \frac{v\left(\vec{w}\right)}{\varepsilon_0^2\left(N, h, \alpha\right)}}\right).$$

Звідси випливає, що

1) в загальному випадку швидкість рівномірної збіжності задовольняє нерівність

$$P\left(\vec{w}\right) \leq v\left(\vec{w}\right) + \varepsilon_1\left(N, h, \alpha, v\right).$$

2) при малих ймовірностях помилки навчання $v\left(\vec{w}\right)$ має місце нерівність

$$P\left(\vec{w}\right) \leq v\left(\vec{w}\right) + 4\varepsilon_0\left(N, h, \alpha\right).$$

3) при великих ймовірностях помилки навчання $v\left(\vec{w}\right)$ має місце нерівність

$$P\left(\vec{w}\right) \leq v\left(\vec{w}\right) + \varepsilon_0\left(N, h, \alpha\right).$$