2. Ймовірнісна постановка задачі розпізнавання образів

Дані про об'єкти з множини X можуть бути неточними або неповними. У цьому випадку одному опису x можуть відповідати разні відповіді. Ймовірнісна постановка полягає у такому: замість невідомої цільової залежності $y^*(x)$ припускається існування невідомого ймовірнісного розподілу на множині $X \times Y$ із щільністю $p(x, y)$, з якого випадково і незалежно вибираються спостереження $T = \{(x_i, y_i)\}_{i=1}^{m}$. Такі вибірки називаються *простими*.

Принцип максимальної правдоподібності

При ймовірнісній постановці задачі замість моделі алгоритмів $g(x, \theta)$, яка апроксимує невідому залежність $y^*(x)$, задається модель сумісної щільності розподілу об'єктів і відповідей $\varphi(x, y, \theta)$, що апроксимує невідому ймовірність $p(x, y)$. Після цього визначається значення параметру θ, при якому вибірка даних T є найбільш правдоподібною, тобто найкраще узгоджується із моделлю щільності.

Якщо спостереження у вибірці T є незалежними, то сумісна щільність всіх спостережень дорівнює добутку значень щільності $p(x, y)$ для кожного спостереження:

$$p(T) = \prod_{i=1}^{m} p(x_i, y_i).$$

Якщо апроксимувати $p(x_i, y_i)$ моделлю щільності $\varphi(x_i, y_i, \theta)$, отримуємо функцію правдоподібності

$$L(\theta, T) = \prod_{i=1}^{m} \varphi(x_i, y_i, \theta).$$

Що більше значення функції правдоподібности, то краще вибірка узгоджується з моделлю. Отже, треба шукати

$$\arg\max_{\theta} L(\theta, T).$$

Описаний вище метод **називається принципом правдоподібності**.

Принцип максимальної правдоподібності і мінімізація емпіричного ризику

Замість максимізації функції правдоподібності $L(\theta, T)$ зручніше мінімізувати функціонал $-\ln L(\theta, T)$, оскільки він є адитивним за об'єктами вибірки.

$$-\ln L(\theta, T) = -\sum_{i=1}^{m} \ln \varphi(x_i, y_i, \theta) \to \min_{\theta}.$$

Озн. 2.1. ***Ймовірнісна функція втрат*** *дорівнює*

$$\mathcal{L}(a_\theta, x) = -m \ln \varphi(x_i, y_i, \theta).$$

Що гірше пара (x_i, y_i) узгоджується з моделлю φ, то менше значення щільності $\varphi(x_i, y_i, \theta)$ і більше величина втрати $\mathcal{L}(a_\theta, x)$, і навпаки, для багатьох функцій втрат можна підібрати таку модель щільності $\varphi(x, y, \theta)$, щоб мінімізація емпіричного ризику була еквівалентною максимізації правдоподібності.

Приклад 2.1. Нехай задано модель алгоритмів $g(x, \theta)$. Припустимо, що помилки

$$\varepsilon(x, \theta) = g(x, \theta) - y^*(x)$$

мають нормальний розподіл

$$N(\varepsilon; 0, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{\varepsilon^2}{2\sigma^2}}.$$

Тоді модель щільності має вигляд

$$\varphi(x, y, \theta) = p(x)\varphi(y|x, \theta) = p(x) N(g(x, \theta) - y^*(x); 0, \sigma^2).$$

Ймовірнісна функція втрат в такому випадку є квадратичною (оскільки логарифм експоненти із квадратичним показником є квадратичною функцією).

$$-\ln \varphi(x, y, \theta) = -\ln p(x) N(g(x, \theta) - y^*(x); 0, \sigma^2) =$$

$$= C_0 + C_1 (g(x, \theta) - y^*(x))^2.$$

Перенавчання і здатність до узагальнення

Мінімізація емпіричного ризику має певні особливості, а саме: якщо мнімум функціонала якості $Q(a,T)$ досягається на алгоритмі a^*, це не гарантує, що алгоритм a^* буде добре наближати цільову залежність на довільній контрольній виборці

$$K = \{(x_i', y_i')\}_{i=1}^{n}.$$

Озн. 2.2. ***Ефекти наднавчання, або надпідгонки*** *полягає в погіршенні якості роботи алгоритму на об'єктах, які не входили до навчальної вибірки.*

Приклад 2.2. Уявімо собі метод, яки просто запам'ятовує об'єкти з навчальних вибірок і розпізнає нові об'єкти лише тоді, коли вони точно збігаються із об'єктами з навчальної вибірки. У цьому випадку емпіричний ризик дорівнює нуля, але точність розпізнавання інших вибірок теж дорівнює нулю.

Навчання — це не лише запам'ятовування, але й узагальнення.

Озн. 2.3. ***Узагальнена здатність методу*** μ *характеризується величиною* $Q(\mu(T), K)$, *якщо вибірки* T *і* K *є репрезентативними, тобто отримані шляхом простого випадкового вибору з однієї генеральної сукупності* X.

Приклад 2.3. *Метод навчання* μ *називається* ***слушним***, *якщо при заданих достатньо малих числах* ε *і* η, *має місце нерівність*

$$P_{T,K}\{Q(\mu(T), K) > \varepsilon\} < \eta. \qquad (1)$$

Озн. 2.4. ***Параметр*** ε ***називається точністю методу*** μ, *а параметр* $1-\eta$ *— його* ***надійністю****.*

Отримання оцінок типу (1) є основною задачею статистичної теоріхї навчання.

Емпіричні оцінки здатності до узагальнення

Емпіричні оцінки застосовуються, коли неможливо отримати теоретичні. Нехай дано вибірку $S = \{(x_i, y_i)\}_{i=1}^{M}$. Розіб'ємо її N способами на диз'юнктні підвибірки: навчальну $T_j = \{(x_i, y_i)\}_{i=1}^{m}$ в контрольну $K_j = \{(x_i, y_i)\}_{i=1}^{n}$, де $n + m = M$.

Озн. 2.5. *Для кожного розбиття* $j = 1, 2, ..., N$ *побудуємо алгоритм* $a_j = \mu(T_j)$ *і обчислимо функціонал* $Q_j = Q(a_j, K_j)$. *Середнє арифметичне значень* Q_j *по всіх розбиттях називається оцінкою* ***ковзного контролю, або кросс-валідацією.***

$$CV(\mu, S) = \frac{1}{N}\sum_{j=1}^{N} Q(\mu(T_j), K_j).$$

Стандартом є методика *t*×*q*-кроссвалідації, коли вибірка випадково розбивається на *q* блоків однакової довжини *t* разів. Переваги: кожний об'єкт враховується *t* разів. Недоліки: обчислювальна складність, велика диспрсія, неповне використання навчальних даних.

Література

1. Воронцов К.В. Машинное обучение. (Курс лекций). ВмиК МГУ: Москва, 2009. http://www.machinelearning.ru/wiki/index.php?title= Машинное_обучение_(курс_лекций%2C_К.В.Воронцов)
2. Дуда Р., Харт П. Распознавание образов и анализ сцен. — М.: Мир, 1976.
3. Фукунага К. Введение в статистическую теорию распознавания образов. — М.: Наука, 1979.
4. Главач В., Шлезингер М.И. Десять лекций по статистическому и структурному распознаванию образов. К.: Наукова думка, 2004. www.irtc.org.ua/image/Files/Schles/esh10_full.pdf.
5. Вапник В.Н., Червоненкис А.Я. Теория распознавания образов М.: Наука, 1974. — 416 с.