Машинное обучение

MIPT DIHT

22 января 2015 г.

1 Билет №1

1.1 Байесовские методы классификации. Наивный байесовский классификатор

Постановка задачи Пусть X — множество объектов, Y — конечное множество имён классов, множество $X \times Y$ является вероятностным пространством с плотностью распределения $p(x, y) = P(y)p(x|y)$. Вероятности появления объектов каждого из классов $P_y = P(y)$ называются априорными вероятностями классов. Плотности распределения $p_y(x) = p(x|y)$ называются функциями правдоподобия классов. Вероятностная постановка задачи классификации разделяется на две независимые подзадачи.

Problem 1. Имеется простая выборка $X^l = (x_i, y_i)_{i=1}^l$ из неизвестного распределения $p(x, y) = P_y p_y(x)$. Требуется построить эмпирические оценки априорных вероятностей $\hat{P}_y$ и функций правдоподобия $\hat{p}_y(x)$ для каждого из классов $y \in Y$.

Problem 2. По известным плотностям распределения $p_y(x)$ и априорным вероятностям P_y всех классов $y \in Y$ построить алгоритм a(x), минимизирующий вероятность ошибочной классификации.

Первая задача имеет множество решений, поскольку многие распределения $p(x, y)$ могли бы породить одну и ту же выборку X^l

Функционал среднего риска Событие вида «$x \in \Omega$ при условии, что x принадлежит классу y»:

$$P(\Omega|y) = \int_\Omega p_y(x)dx$$

Definition 1. Функционалом среднего риска называется ожидаемая величина потери при классификации объектов алгоритмом a:

$$R(a) = \sum_{y \in Y} \sum_{s \in Y} \lambda_{ys} P_y P(a_s|y)$$

Оптимальное байесовское решающее правило

Theorem 1. *Если известны априорные вероятности P_y и функции правдоподобия $p_y(x)$, то минимум среднего риска $R(a)$ достигается алгоритмом*

$$a(x) = arg \min_{s \in Y} \sum_{y \in Y} \lambda_{ys} P_y p_y(x)$$

Proof Для произвольного t из Y запишем функционал среднего риска:

$$R(a) = \sum_{y \in Y} \sum_{s \in Y} \lambda_{ys} P_y P(A_s|y) =$$

$$= \sum_{y \in Y} \lambda_{ys} P_y P(A_t|y) + \sum_{\substack{s \in Y \\ t}} \sum_{y \in Y} \lambda_{ys} P_y P(A_s|y)$$

Из формулы полной вероятности $P(A_t|y) = 1 - \sum_{\substack{s\in Y\\t}} P(A_s|y)$

$$R(a) = \sum_{y\in Y} \lambda_{ys} P_y + \sum_{\substack{s\in Y\\t}} \sum_{y\in Y} (\lambda_{ys} - \lambda_{yt}) P_y P(a_s|y) =$$

$$= const(a) + \sum_{\substack{s\in Y\\t}} \int_{A_s} \sum_{y\in Y} (\lambda_{ys} - \lambda_{yt}) P_y p_y(x)$$

Обозначим $g_s(x) = \sum_{y\in Y} \lambda_{ys} P_y p_y(x)$

$$R(a) = const(a) + \sum_{\substack{s\in Y\\t}} \int_{A_s} (g_s(x) - g_t(x))dx$$

В выражении неизвестны только области A_s. Функционал R(a) есть сумма $|Y| - 1$ слагаемых $I(A_s) = \int_{A_s} (g_s(x) - g_t(x)dx$, каждое из которых зависит только от одной области A_s. Минимум $I(A_s)$ достигается, когда A_s совпадает с областью неположительности подынтегрального выражения. В силу произвольности t:

$$A_s = \{x \in X | g_s(x) \leqslant_t (x), \forall t \in Y, t \neq s\}$$

С другой стороны $a_s = \{x \in X | a(x) = s\}$. Значит $a(x) = s \iff s = arg \min_{t\in Y} g_t(x)$ •

Theorem 2. *Если P_y и $p_y(x)$ известны, $\lambda_{ss} = 0, \lambda_{ys} \equiv \lambda_y \forall y, s \in Y$, то минимум среднего риска достигается алгоритмом*

$$a(x) = arg \max_{y\in Y} \lambda_y P_y p_y(x)$$

Наивный байесовский классификатор

Problem 3. Задано множество объектов $X^m = x_1, \ldots, x_m$, выбранных случайно и независимо согласно неизвестному распределению p(x). Требуется построить эмпирическую оценку плотности — функцию $\hat{p}(x)$, приближающую p(x) на всём X. Допустим, что объекты $x \in X$ описываются n числовыми признаками $f_j : X \to R, j = 1, \ldots, n$. Обозначим через $x == (\xi_1, ..., \xi_n)$ произвольный элемент пространства объектов $X = \mathbb{R}^n$, где $\xi_j = f_j(x)$.

Proposition 1. *Признаки $f_1(x), \ldots, f_n(x)$ являются независимыми случайными величинами. Следовательно, функции правдоподобия классов представимы в виде $p_y(x) = p_{y1}(\xi_1)\Delta\Delta\Delta p_{yn}(\xi_n), y \in Y$, где $p_{yj}(\xi_j)$ — плотность распределения значений j-го признака для класса y.*

Подставим эмпирические оценки одномерных плотностей $\hat{p}_{yj}(\xi)$ в предположение и затем в теорему 2. Получим алгоритм

$$a(x) = arg \max_{y\in Y} \left(ln\lambda_y \hat{P}_y + \sum)j{=}1^n ln\hat{p}_{yj}(\xi_j) \right)$$

2 Билет №2

2.1 Логистическая регрессия

Базовые предположения. Пусть классов два, $Y = -1, +1$, объекты описываются n числовыми признаками $f_j : X \to R, j = 1, \ldots, n$. Будем полагать $X = \mathbb{R}^n$ отождествляя объекты с их признаковыми описаниями: $x \equiv (f_1(x), \ldots, f_n(x))$.

2.2 L1 и L2 регуляризации

2.3 Теорема об оптимальности

3 Билет №3

3.1 Метод опорных векторов. Решение двойственной задачи

3.2 Спрямляющее пространство

3.3 Ядра

4 Билет №4

4.1 Многомерная линейная регрессия

Метод наименьших квадратов

$$Q(\alpha, X^l) = \sum_{i=1}^{l} w_i \Big(f(x_i, \alpha) - y_i)^2 \Big) \to \min_{\alpha}$$

где w_i - вес i-го объекта $Q(\alpha^*, X^l$ - остаточная сумма квадратов **Многомерная логрегрессия** $f_1(x) \dots f_n(x)$ - числовые признаки. Модель многомерной регрессии:

$$f(x, \alpha) = \sum_{j=1}^{n} \alpha_j f_j(x), \alpha \in \mathbb{R}^n$$

Функионал квадрата ошибки:

$$Q(\alpha, X^l) = \sum_{i=1}^{l} (f(x_i, \alpha) - y_i)^2 = ||F\alpha - y||^2 \to \min_{\alpha}$$

Необхожимое условие минимума в матричном виде:

$$\frac{\partial Q}{\partial \alpha}(\alpha) = 2F^T(F\alpha - y) = 0$$

откуда следует нормальная система задачи метода наименьших квадратов: $F^T F\alpha = F^T y$ Решение системы: $\alpha^* = (F^T F)^{-1} F^T y = F^+ y$. Значение функционала: $Q(\alpha^*) = ||P_F y - y||^2$, где $P_F = FF^+ = (F^T F)^{-1} F^T$ - проекционная матрица

4.2 Лассо Тибширани

$$\begin{cases} Q(\alpha) = ||F\alpha - y||^2 \to \min_{\alpha}, \\ \sum_{j=1}^{n} |\alpha_j| \leqslant \varkappa \end{cases}$$

Лассо приводит к отбору признаков. После замены переменных

$$\begin{cases} \alpha_j = \alpha_j^+ - \alpha_j^- \\ |\alpha_j| = \alpha_j^+ + \alpha_j^- \end{cases} , \alpha_j^+, \alpha_j^- \geqslant 0$$

ограничения принимают канонический вид

$$\sum_{j=1}^{n} \alpha_j^+ + \alpha_j^- \leqslant \varkappa$$

Чем меньше $\varkappa$, тем больше таких j что $\alpha_j^+ = \alpha_j^- = 0$

4.3 Гребневая регрессия

Штраф за увеличение норма вектора весов $||\alpha||$:

$$Q_\tau(\alpha) = ||F\alpha - y||^2 + \frac{1}{\sigma}||\alpha||^2$$

где $\tau = \frac{1}{\sigma}$ - неотрицательный параметр регуляризации. **Вероятностная интерпретация** Априорное распределение вектора α - гауссовское с ковариационной матрицей σI_n Модифицированное МНК-решение:

$$\alpha_\tau^* = (F^T F + \tau I_n)^{-1} F^T y$$

5 Билет №5

5.1 Бустинг

5.2 Алгоритм AdaBoost