1 Структурное обучение

Заданы множества $\mathcal{X}, \mathcal{Y}$ и выборка наблюдений $(x_t, y_t)_{t=1}^n$. Задачей структурного обучения (SL) [1] является построение отображения $f : \mathcal{X} \to \mathcal{Y}$.

Рассматривается следующий класс алгоритмов решения задачи SL. Задана *дискриминантная функция* $Q(x, y, \mathbf{w})$, где $\mathbf{w}$ — вектор параметров. Значением отображения f является решение задачи максимизации дискриминантной функции Q по y:

$$f(x, \mathbf{w}) = \arg\max_{y \in \mathcal{Y}} Q(x, y, \mathbf{w}). \tag{1}$$

Отметим, что отображение f, определяемое уравнением (1), зависит от переменной x, а также от вектора параметров $\mathbf{w}$. Оптимальные параметры $\mathbf{w}$ определяются из предположения о том, что значения отображения $f(x_t, \mathbf{w})$ на объектах обучающей выборки x_t должны быть структурно близкими к значениям переменной y_t.

Другими словами, на множестве $\mathcal{Y}$ определена функция потерь $\Delta : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, где $\Delta(y, \hat{y})$ определяет значение потери предсказания $\hat{y}$ при условии, что истинным значением является y. Для нахождения оптимальных параметров $\mathbf{w}$ рассматривается задача минимизации эмпирического риска

$$L = \sum_{i=1}^{n} \Delta(y_t, f(x_t, \mathbf{w})), \tag{2}$$

Стандартные подходы решения задачи структурного обучения. Для решения задачи (2) в [1] предлагается следующая методология. Предполагается, что дискриминантная функция линейно зависит от параметров, $Q(x, y, \mathbf{w}) = \langle \mathbf{w}, \Psi(x, y) \rangle$, где вектор $\Psi(x, y)$ является совместным признаковым описанием объекта x и целевой переменной y.

Поскольку целевая функция в (2) не является выпуклой, предлагается оптимизировать ее выпуклую верхнюю грань:

$$\Delta(y_t, f(x_t)) \leq \max_{y \in \mathcal{Y}} \left(\Delta(y_t, y) + Q(x_t, y, \mathbf{w}) - Q(x_t, y_t, \mathbf{w}) \right).$$

Минимизация этой функции совместно с рассмотрением линейной дискриминантной функции приводит к структурному методу опорных векторов [2].

2 Предсказание структуры суперпозиций

Структура на множестве суперпозиций. Рассматривается задача предсказания структуры суперпозиций (SSF) — задача построения отображения $f : x \mapsto y$, где целевая переменная y является композицией базисных функций $y = h_0 \circ h_{k_1} \circ ... \circ h_{k_y}$, выбранных с возвращением из множества базисных функций $\mathcal{H} = \{h_0, h_1, ..., h_r\}$. Каждой базисной функции h_k соответствует арность $a_k \geq 0$. Крайним левым элементом суперпозиции y

является функция $h_0 : h_0(x) = x$, имеющая арность $a_0 = 1$. Кроме того, предполагается, что существует непустое подмножество $\mathcal{H}_0 \subset \mathcal{H}$ функций нулевой арности, называемых свободными переменными.

Поставим в соответствие целевой переменной y крашеное дерево $\Gamma = (V, E)$. На множестве вершин задана функция раскраски $h : V \to \mathcal{H}$. Цвет $h(v)$ вершины $h(v)$ является базисной функцией и определяет количество дочерних вершин у v: оно совпадает с арностью функции $h(v)$. Таким образом, допустимыми деревьями являются деревья, имеющие свои корнем вершину $v_0 : h(v_0) = h_0$, в листьях содержащие функции нулевой арности, а количество дочерних вершин для всех остальных элементов совпадает с соответствующими арностями.

Распределение на цветах. Рассматривается следующее предположение о распределении данных. Предполагается, что на множестве цветов-базисных функций задано распределение $\mathbf{w}$, $w_{ij} \equiv w(h_i, h_j)$ — вероятность наблюдать цвет h_j в виде цвета дочерней вершины для вершины с цветом h_i.

Принимается гипотеза «независимости от посторонних вершин»: вероятность того, что вершина v в дереве Γ имеет цвет h_j, зависит только от цвета родителя этой вершины $h(v_p)$: $p(h(v) = h_j \mid \Gamma \backslash v) = p(h(v_p), h(v) = h_j)$.

Для определенного таким образом распределения выпишем функцию правдоподобия дерева y, которая в силу независимости раскладывается в произведение вероятностей ребер:

$$P(y) = \prod_{(i,j)\in E} w(h(v_i), h(v_j)).$$

В качестве дискриминантной функции $Q(x, y, \mathbf{w})$ принимается правдоподобие дерева $P(y)$:

$$Q(x, y, \mathbf{w}) = \prod_{(i,j)\in E} w(h(v_i), h(v_j)).$$

Оптимальное значение целевой переменной y задается деревом, максимизирующим дискриминантную функцию:

$$f(x, \mathbf{w}) = \arg\max_{y\in\mathcal{Y}} \prod_{(i,j)\in E} w(h(v_i), h(v_j)). \tag{3}$$

Для определения оптимальных параметров w_{ij} определим функцию потери $\Delta(y, \hat{y})$ в виде количества несовпадающих элементов бинарных векторов $\mathbf{y}$ и $\hat{\mathbf{y}}$:

$$\Delta(y, \hat{y}) = \sum_{i,j} |y_{ij} - \hat{y}_{ij}|, \tag{4}$$

где элемент вектора $\mathbf{y}$, индексируемый y_{ij}, равен 1 в случае, если последовательность цветов (h_i, h_j) принадлежит множеству ребер E_y дерева y, и равен 0 в противном случае. Отметим, что размерность вектора $\mathbf{y}$ равна $|\mathcal{H}|^2$ и совпадает с размерностью вектора $\mathbf{w}$.

Оптимальные параметры $\hat{\mathbf{w}}$ должны минимизировать функцию эмпирического риска (4). Отметим, что в случае, когда вектор $\mathbf{w}$ совпадает с вектором $\mathbf{y}$, задача максимизации дискриминантной функции (3) в качестве решения содержит элемент y_i. Исходя из этих соображений, будем искать параметры w_{ij} в виде функций от элемента x: $w_{ij} = \theta_{ij}(x)$ с помощью решения задачи минимизации

$$\hat{\theta}_{ij} = \arg\min_{\theta_{ij}} \|y_{ij} - \theta_{ij}(x)\|.$$

Оценка параметров. Исходя из вышесказанного, задача прогнозирования структуры суперпозиций состоит из двух последовательных этапов.

1. Оценка параметров распределения $\hat{\mathbf{w}}$ путем минимизации отклонения параметра $w_{ij} = \theta_{ij}(x)$ от элемента вектора y_{ij}:

$$\hat{\theta}_{ij} = \arg\min_{\theta_{ij}} \|y_{ij} - \theta_{ij}(x)\|. \tag{5}$$

2. Поиск оптимального элемента $f(x)$, максимизирующего дискриминантную функцию (3):

$$f(x, \hat{\mathbf{w}}) = \arg\max_{y \in \mathcal{Y}} \prod_{(i,j) \in E} \hat{w}(h(v_i), h(v_j)).$$

Для решения задачи (5) предлагается стандартный метод многоклассовой классификации...

Для решения задачи (3) предлагается алгоритм на основе динамического программирования. Утверждается, что алгоритм находит оптимальное решение за $O(|\mathcal{H}|^3)$ вычислений.

Алгоритм максимизации дискриминантной функции. Алгоритм основывается на принципе динамического программирования. На шаге k алгоритм хранит массив из $|\mathcal{H}|$ элементов, элемент i которого содержит стоимость оптимального дерева Γ_i^k с корнем в вершине, раскрашенной h_i, с количеством вершин не более k.

На шаге $k+1$ для каждой базисной функции $h_{i'}$ выполняется процедура присоединения корня $(h_{i'}, \Gamma_i^k)$ для всех деревьев $\Gamma_i^k, i = 1, ..., |\mathcal{H}|$. Если максимальная стоимость построенных деревьев превосходит стоимость дерева $\Gamma_{i'}^k$, то происходит замена оптимального дерева $\Gamma_{i'}^k$ на дерево $\Gamma_{i'}^{k+1} \equiv (h_{i'}, \Gamma_i^k)$ с максимальной стоимостью. Иначе дерево остается прежним, $\Gamma_{i'}^{k+1} \equiv \Gamma_i^k$.

Если после итерации $k+1$ не произошло ни одной замены дерева, то алгоритм останавливается, а его решением является наилучшее из деревьев с присоединенным корнем (h_0, Γ_i^k). Утверждается, что количество шагов k на превосходит количество базисных функций $k \leq O(|\mathcal{H}|)$. На каждом шаге алгоритм выполняет $O(|\mathcal{H}|^2)$ проверок.

Список литературы

[1] Ioannis Tsochantaridis, Thorsten Joachims, Thomas Hofmann, and Yasemin Altun. Large margin methods for structured and interdependent output variables. In *Journal of Machine Learning Research*, pages 1453–1484, 2005.

[2] Chun-Nam John Yu and Thorsten Joachims. Learning structural svms with latent variables. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 1169–1176. ACM, 2009.