Оценка параметров смеси распределений

К. В. Павлов
kirill.pavlov@phystech.edu
Московский физико-технический институт

В работе рассматриваются способы построения смеси моделей и экспертов. Предлагается EM-алгоритм для совместного нахождения параметров моделей и их весов в смеси, а так же для нахождения параметров смеси обобщенных линейных моделей.

Ключевые слова: *смеси моделей, обобщенно-линейные модели, смеси экспертов.*

Введение

При решении задачи анализа данных строится модель — отображение известных характеристик объекта в неизвестные. Часто оказывается, что качество алгоритма можно улучшить с помощью комбинирования нескольких моделей [3, p. 653–676]. Например, можно обучить l моделей и в качестве ответа выводить усредненный ответ по всем моделям. Подобные комбинации моделей называются комитетами. Один из наиболее важных случаев комитета является бустинг. Алгоритмы в комитет добавляются последовательно и их параметры зависят от уже созданного на момент добавления комитета. Другим важным частным случаем комитета является смесь экспертов. В этом случае ответы алгоритмов взвешиваются в зависимости от области пространства, в которой находится объект. Рассмотрим способы построения композиций.

Общий подход к оценке параметров моделей

В случае, когда одной модели для описания данных не хватает, используют смеси моделей. Предполагается, что исходная зависимость $p(\boldsymbol{y}\,|\,\boldsymbol{x})$ выражается как композиция моделей $p(y\,|\,\boldsymbol{x}, \boldsymbol{w}_k)$ формулой:

$$p(\boldsymbol{y}\,|\,\boldsymbol{x}) = \sum_{k=1}^{l} p(\boldsymbol{w}_k\,|\,\boldsymbol{x}) p(y\,|\,\boldsymbol{x}, \boldsymbol{w}_k) = \sum_{k=1}^{l} \pi_k p(y\,|\,\boldsymbol{x}, \boldsymbol{w}_k), \tag{1}$$

где $\pi_k = p(\boldsymbol{w}_k\,|\,\boldsymbol{x})$ — вероятность принадлежности к модели k. На π_k накладываются условия нормировки: вероятность каждой модели неотрицательна и сумма вероятностей равна единице.

$$\sum_{k=1}^{l} \pi_k = 1, \quad \pi_k \geqslant 0 \quad \forall k. \tag{2}$$

Далее предполагается, что объекты в выборке независимы и плотность совместного распределения преобразуется в произведение плотностей распределения каждого объекта.

$$p(\boldsymbol{y}\,|\,\boldsymbol{x}) = \sum_{k=1}^{l} \pi_k \prod_{i=1}^{n} p(y^i\,|\,\boldsymbol{x}^i, \boldsymbol{w}_k) = \prod_{i=1}^{n} \sum_{k=1}^{l} \pi_k p(y^i\,|\,\boldsymbol{x}^i, \boldsymbol{w}_k). \tag{3}$$

В формуле 3 произведена смена порядка суммирования перемножения. Используя принцип максимума правдоподобия, будет максимизировать $p(\boldsymbol{y}\,|\,\boldsymbol{x})$. Проще это делать, введя функцию правдоподобия $Q(\boldsymbol{w_1}, \ldots, \boldsymbol{w_l}, \boldsymbol{\pi})$ как логарифм плотности вероятности данных.

$$Q(\boldsymbol{w}^1, \dots, \boldsymbol{w}^l, \boldsymbol{\pi}) = \ln p(\boldsymbol{y} \mid \boldsymbol{x}) = \sum_{i=1}^{m} \ln \left[\sum_{k=1}^{l} \pi_k p(y^i \mid \boldsymbol{x}^i, \boldsymbol{w}_k) \right]. \tag{4}$$

Обозначим через $p(y, \boldsymbol{w}_k \mid \boldsymbol{x})$ вероятность того, что объект $(\boldsymbol{x}, y)$ был порожден компонентой $\boldsymbol{w}_k$, $\gamma_{ik} = p(\boldsymbol{w}_k | y^i, \boldsymbol{x}^i)$ — вероятность того, что i-объект порожден j - компонентой. Каждый объект был порожден какой-либо моделью, по формуле полной вероятности

$$\sum_{k=1}^{l} \gamma_{ik} = 1, \quad \forall i. \tag{5}$$

Для произвольного объекта $(\boldsymbol{x}, y)$ вероятность его получения моделью w_k по формуле условной вероятности равна:

$$p(y, \boldsymbol{w}_k \mid \boldsymbol{x}) = p(\boldsymbol{w}_k \mid \boldsymbol{x}) p(y \mid \boldsymbol{x}, \boldsymbol{w}_k) \equiv \pi_k p(y \mid \boldsymbol{x}, \boldsymbol{w}_k). \tag{6}$$

Подставим это равенство в формулу Байеса для γ_{ik}

$$\gamma_{ik} = \frac{\pi_k p(y^i \mid \boldsymbol{x}^i, \boldsymbol{w}_k)}{\sum_{s=1}^{l} \pi_s p(y^i \mid \boldsymbol{x}^i, \boldsymbol{w}_s)}. \tag{7}$$

Для определения параметров смеси необходимо решить задачу максимизации правдоподобия $Q(\boldsymbol{w}^1, \dots, \boldsymbol{w}^l, \boldsymbol{\pi}) \to \max$, это можно сделать с использованием функции Лагранжа [1], которая имеет вид:

$$L = \sum_{i=1}^{m} \ln \left[\sum_{k=1}^{l} \pi_k p(y^i \mid \boldsymbol{x}^i, \boldsymbol{w}^k) \right] - \lambda \left(\sum_{k=1}^{l} \pi_k - 1 \right). \tag{8}$$

Необходимым условием экстремума функции является равенство нулю первых производных. Приравняем производную функции Лагранжа по π_k к нулю:

$$\frac{\partial L}{\partial \pi_k} = \sum_{i=1}^{m} \frac{p(y^i \mid \boldsymbol{x}^i, \boldsymbol{w}^k)}{\sum_{s=1}^{l} \pi_s p(y^i \mid \boldsymbol{x}^i, \boldsymbol{w}^s)} - \lambda = 0. \tag{9}$$

Умножим обе части равенства на π_k и просуммируем по $k = 1..l$

$$m = \sum_{k=1}^{l} \sum_{i=1}^{m} \frac{\pi_k p(y^i \mid \boldsymbol{x}^i, \boldsymbol{w}^k)}{\sum_{s=1}^{l} \pi_s p(y^i \mid \boldsymbol{x}^i, \boldsymbol{w}^s)} = \lambda \sum_{s=1}^{l} \pi_k = \lambda. \tag{10}$$

Получилось, что $\lambda = m$ необходимое условие минимума. В выражении для производной $\frac{\partial L}{\partial \pi_k}$ заменим λ на m и домножим обе части равенства на π_k:

$$\pi_k = \frac{1}{m} \sum_{i=1}^{m} \frac{\pi_k p(y^i \mid \boldsymbol{x}^i, \boldsymbol{w}^k)}{\sum_{s=1}^{l} \pi_s p(y^i \mid \boldsymbol{x}^i, \boldsymbol{w}^s)} = \frac{1}{m} \sum_{i=1}^{m} \gamma_{ik}. \tag{11}$$

Равенство 11 позволяет находить коэффициенты π_k смеси модели при известных γ_{ik}. Вычислим производную функции Лагранжа по параметрам k-й модели:

$$\frac{\partial L}{\partial \boldsymbol{w}^k} = \sum_{i=1}^{m} \frac{\pi_k \frac{\partial p(y^i \mid \boldsymbol{x}^i, \boldsymbol{w}^k)}{\partial \boldsymbol{w}^k}}{\sum_{s=1}^{l} \pi_s p(y^i \mid \boldsymbol{x}^i, \boldsymbol{w}^s)} = \sum_{i=1}^{m} \frac{\pi_k p(y^i \mid \boldsymbol{x}^i, \boldsymbol{w}^k)}{\sum_{s=1}^{l} \pi_s p(y^i \mid \boldsymbol{x}^i, \boldsymbol{w}^s)} \frac{\partial \ln p(y^i \mid \boldsymbol{x}^i, \boldsymbol{w}^k)}{\partial \boldsymbol{w}^k}. \tag{12}$$

Преобразуем выражение:

$$\frac{\partial L}{\partial \boldsymbol{w}^k} = \frac{\partial}{\partial \boldsymbol{w}^k} \sum_{i=1}^{m} \gamma_{ik} \ln p(y^i \,|\, \boldsymbol{x}^i, \boldsymbol{w}^k) = 0. \tag{13}$$

Полученное равенство совпадает с необходимым условием максимума в задаче максимизации взвешенного правдоподобия:

$$\sum_{i=1}^{m} \gamma_{ik} \ln p(y^i \,|\, \boldsymbol{x}^i, \boldsymbol{w}^k) \to \max_{\boldsymbol{w}^k}. \tag{14}$$

В общем случае задача оптимизации $Q(\boldsymbol{w}^1, \ldots, \boldsymbol{w}^l, \boldsymbol{\pi}) \to \max$ трудна, для её решения используют EM-алгоритм, заключающийся в итеративном повторении двух шагов. На E-шаге вычисляются ожидаемые значения вектора скрытых переменных γ_{ik} по текущему приближения параметров моделей $(\boldsymbol{w}_1, \ldots, \boldsymbol{w}_l)$. На M-шаге решается задача максимизации правдоподобия Q при начальном приближении параметров моделей и значений γ_{ik}.

E-шагу соответствует выражение

$$\gamma_{ik} = \frac{\pi_k p(y^i \,|\, \boldsymbol{x}^i, \boldsymbol{w}_k)}{\sum_{s=1}^{l} \pi_s p(y^i \,|\, \boldsymbol{x}^i, \boldsymbol{w}_s)}. \tag{15}$$

M-шаг заключается в оптимизации параметров распределений.

$$Q(\boldsymbol{w}^1, \ldots, \boldsymbol{w}^l \,|\, \boldsymbol{\pi}) \to max \tag{16}$$

Формула на M-шаге может упроститься для случая конкретного распределения. Для упрощения дальнейших рассуждений введем обозначения

$$G = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_l) = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1l} \\ \vdots & \ddots & \vdots \\ \gamma_{m1} & \cdots & \gamma_{ml} \end{pmatrix}, \quad G_k = \begin{pmatrix} \gamma_{1k} & & 0 \\ & \ddots & \\ 0 & & \gamma_{mk} \end{pmatrix}. \tag{17}$$

Перейдем к рассмотрению линейный и обобщенных линейных моделей.

Оценка параметров смеси линейных моделей

Линейная модель имеет вид:

$$\boldsymbol{y} = X\boldsymbol{w} + \boldsymbol{\varepsilon}, \tag{18}$$

где $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0}, B)$ — вектор нормально распределенных ошибок. В данной постановке вектор $\boldsymbol{y}$ является нормальным с математическим ожиданием $\mathsf{E}(y \,|\, \boldsymbol{x}) = \mu = \boldsymbol{x}^{\mathsf{T}}\boldsymbol{w}$, и корреляционной матрицей B. Плотность распределения $\boldsymbol{y}$ задается формулой:

$$p(\boldsymbol{y} \,|\, X, \boldsymbol{w}) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{|\det B|}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - X\boldsymbol{w})^{\mathsf{T}} B (\boldsymbol{y} - X\boldsymbol{w})\right). \tag{19}$$

Применим для задачи описанный EM-алгоритм.Шаг E сводится к применению формулы 15, а шаг M алгоритма принимает следующий вид:

$$G_k \ln\left[\frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{|\det B|}}\right] - \frac{1}{2}\left(G_k (\boldsymbol{y} - X\boldsymbol{w})^{\mathsf{T}} B (\boldsymbol{y} - X\boldsymbol{w})\right) \to \max_{\boldsymbol{w}} \tag{20}$$

Первое слагаемое не зависит от $\boldsymbol{w}_k$, его можно не учитывать. Преобразование второго слагаемого дает

$$\frac{1}{2}\boldsymbol{w}^{\mathsf{T}}X^{\mathsf{T}}G_kBX\boldsymbol{w} - \boldsymbol{w}^{\mathsf{T}}X^{\mathsf{T}}G_kB\boldsymbol{y} \to \min_{\boldsymbol{w}} \tag{21}$$

Задача квадратична по $\boldsymbol{w}$, решение находится аналитически

$$\boldsymbol{w}^* = (X^{\mathsf{T}}G_kBX)^{-1}G_kBX\boldsymbol{y}. \tag{22}$$

Оценка параметров смеси обобщенно-линейных моделей

В случае обобщенных линейный моделей функция плотности распределения имеет вид

$$p(\boldsymbol{y}\,|\,\boldsymbol{\theta}) = \exp\left(\boldsymbol{T}(\boldsymbol{y})^{\mathsf{T}}\boldsymbol{\eta}(\boldsymbol{\theta}) - b(\boldsymbol{\theta}) + c(\boldsymbol{y})\right). \tag{23}$$

M-шаг алгоритма сводится к максимизации

$$\boldsymbol{T}(\boldsymbol{y})^{\mathsf{T}}G_k\boldsymbol{\eta}(\boldsymbol{\theta}) - b(G_k\boldsymbol{\theta}) + c(G_k\boldsymbol{y}) \to \max_{\boldsymbol{\theta}}. \tag{24}$$

Последнее слагаемое не зависит от параметров модели θ, что позволяет упростить функционал

$$\boldsymbol{T}(\boldsymbol{y})^{\mathsf{T}}G_k\boldsymbol{\eta}(\boldsymbol{\theta}) - b(G_k\boldsymbol{\theta}) \to \max_{\boldsymbol{\theta}}. \tag{25}$$

Дальнейшая минимизация зависит от конкретного семейства из обобщенного класса распределений.

Оценка параметров смеси экспертов

Понятие смеси экспертов было введено Джорданом и Якобсом в 1991г [2]. Предполагается, что параметры смеси π являются функциями от объекта, т.е.

$$p(\boldsymbol{y}\,|\,\boldsymbol{x}) = \sum_{k=1}^{l}\pi_k(\boldsymbol{x})p(y\,|\,\boldsymbol{x},\boldsymbol{w}_k). \tag{26}$$

Компоненты $\pi_k(\boldsymbol{x})$ называются функциями селективности, а $p(y\,|\,\boldsymbol{x},\boldsymbol{w}_k)$ экспертами. Функция селективности отвечает за компетентность эксперта в определенной области.

Оказывается [4], что наличие функции компетенции допускает решение задачи с помощью EM-алгоритма, причем, E-шаг остается прежним:

$$\gamma_{ik} = \frac{\pi_k(\boldsymbol{x}^i)p(y^i\,|\,\boldsymbol{x}^i,\boldsymbol{w}_k)}{\sum_{s=1}^{l}\pi_s(\boldsymbol{x}^i)p(y^i\,|\,\boldsymbol{x}^i,\boldsymbol{w}_s)}. \tag{27}$$

M-шаг принимает вид:

$$\pi_k = \frac{1}{m}\sum_{i=1}^{m}\gamma_{ik}. \tag{28}$$

$$\sum_{i=1}^{m}\gamma_{ik}(\boldsymbol{x}^i)\ln p(y^i\,|\,\boldsymbol{x}^i,\boldsymbol{w}^k) \to \max_{\boldsymbol{w}^k}. \tag{29}$$

Уравнение 29 можно решить с помощью метода итеративно перевзвешенных наименьших квадратов (IRLS).

Литература

[1] *Воронцов К. В.* Курс лекций: Линейные методы классификации. — 2009. — 01. `http://www.machinelearning.ru/wiki/images/6/68/Voron-ML-Lin.pdf`.

[2] Adaptive mixtures of local experts / R. A. Jacobs, M. I. Jordan, S. J. Nowlan, G. E. Hinton // *Neural Computation.* — 1991. — no. 3. — Pp. 79–87.

[3] *Bishop C. M.* Pattern Recognition and Machine Learning. — Springer, Series: Information Science and Statistics, 2006. — 740 pp.

[4] *Jordan M. I., Jacobs R. A.* Hierarchical mixtures of experts and the EM algorithm // *Neural Computation.* — 1994. — no. 6. — Pp. 181–214. `citeseer.ist.psu.edu/article/jordan94hierarchical.html`.