Оценка параметров смеси распределений

Кирилл Павлов, 674 группа

6 декабря 2011 г.

1 Введение

В случае, когда одной модели для описания данных не хватает, используют смеси моделей. Предполагается, что исходная зависимость выражается формулой:

$$p(\boldsymbol{y} \,|\, \boldsymbol{x}) = \sum_{k=1}^{l} p(\boldsymbol{w}_k \,|\, \boldsymbol{x}) p(y \,|\, \boldsymbol{x}, \boldsymbol{w}_k) = \sum_{k=1}^{l} \pi_k p(y \,|\, \boldsymbol{x}, \boldsymbol{w}_k), \tag{1.1}$$

где $\pi_k = p(\boldsymbol{w}_k \,|\, \boldsymbol{x})$ — вероятность принадлежности модели k,

$$\sum_{k=1}^{l} \pi_k = 1. \tag{1.2}$$

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

$$p(\boldsymbol{y} \,|\, \boldsymbol{x}) = \sum_{k=1}^{l} \pi_k \prod_{i=1}^{n} p(y^i \,|\, \boldsymbol{x}^i, \boldsymbol{w}_k) = \prod_{i=1}^{n} \sum_{k=1}^{l} \pi_k p(y^i \,|\, \boldsymbol{x}^i, \boldsymbol{w}_k). \tag{1.3}$$

Введем функцию правдоподобия $Q(\boldsymbol{w_1}, \ldots, \boldsymbol{w_l}, \boldsymbol{\pi})$ как логарифм плотности вероятности данных.

$$Q(\boldsymbol{w}^1, \ldots, \boldsymbol{w}^l, \boldsymbol{\pi}) = \ln p(\boldsymbol{y} \,|\, \boldsymbol{x}) = \sum_{i=1}^{m} \ln \left[\sum_{k=1}^{l} \pi_k p(y^i \,|\, \boldsymbol{x}^i, \boldsymbol{w}_k) \right]. \tag{1.4}$$

Обозначим через $p(y, \boldsymbol{w}_k \,|\, \boldsymbol{x})$ 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

$$\sum_{k=1}^{l}\gamma_{ik}=1,\quad \forall i. \tag{1.5}$$

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

$$p(y,\boldsymbol{w}_k\,|\,\boldsymbol{x})=p(\boldsymbol{w}_k\,|\,\boldsymbol{x})p(y\,|\,\boldsymbol{x},\boldsymbol{w}_k)\equiv\pi_k p(y\,|\,\boldsymbol{x},\boldsymbol{w}_k). \tag{1.6}$$

Подставим это равенство в формулу Байеса для γ_{ik}

$$\gamma_{ik}=\frac{\pi_k p(y^i\,|\,\boldsymbol{x}^i,\boldsymbol{w}_k)}{\sum_{s=1}^{l}\pi_s p(y^i\,|\,\boldsymbol{x}^i,\boldsymbol{w}_s)}. \tag{1.7}$$

Для определения параметров смеси необходимо решить задачу максимизации правдоподобия $Q(\boldsymbol{w}^1,\ldots,\boldsymbol{w}^l,\boldsymbol{\pi})\to max$, для этого выпишем функцию Лагранжа:

$$L=\sum_{i=1}^{m}\ln\left[\sum_{k=1}^{l}\pi_k p(y^i\,|\,\boldsymbol{x}^i,\boldsymbol{w}^k)\right]-\lambda\left(\sum_{k=1}^{l}\pi_k-1\right). \tag{1.8}$$

Приравняем производные функции Лагранжа к нулю:

$$\frac{\partial L}{\partial\pi_k}=\sum_{i=1}^{m}\frac{p(y^i\,|\,\boldsymbol{x}^i,\boldsymbol{w}^k)}{\sum_{s=1}^{l}\pi_s p(y^i\,|\,\boldsymbol{x}^i,\boldsymbol{w}^s)}-\lambda=0. \tag{1.9}$$

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

$$m=\sum_{k=1}^{l}\sum_{i=1}^{m}\frac{\pi_k p(y^i\,|\,\boldsymbol{x}^i,\boldsymbol{w}^k)}{\sum_{s=1}^{l}\pi_s p(y^i\,|\,\boldsymbol{x}^i,\boldsymbol{w}^s)}=\lambda\sum_{s=1}^{l}\pi_k=\lambda. \tag{1.10}$$

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

$$\pi_k=\frac{1}{m}\sum_{i=1}^{m}\frac{\pi_k p(y^i\,|\,\boldsymbol{x}^i,\boldsymbol{w}^k)}{\sum_{s=1}^{l}\pi_s p(y^i\,|\,\boldsymbol{x}^i,\boldsymbol{w}^s)}=\frac{1}{m}\sum_{i=1}^{m}g_{ik}. \tag{1.11}$$

Производная лагранжиана по параметрам k-й модели:

$$\frac{\partial L}{\partial\boldsymbol{w}^k}=\sum_{i=1}^{m}\frac{\pi_k\frac{\partial p(y^i\,|\,\boldsymbol{x}^i,\boldsymbol{w}^k)}{\partial\boldsymbol{w}^k}}{\sum_{s=1}^{l}\pi_s p(y^i\,|\,\boldsymbol{x}^i,\boldsymbol{w}^s)}=\sum_{i=1}^{m}\frac{\pi_k p(y^i\,|\,\boldsymbol{x}^i,\boldsymbol{w}^k)}{\sum_{s=1}^{l}\pi_s p(y^i\,|\,\boldsymbol{x}^i,\boldsymbol{w}^s)}\frac{\partial\ln p(y^i\,|\,\boldsymbol{x}^i,\boldsymbol{w}^k)}{\partial\boldsymbol{w}^k}. \tag{1.12}$$

Преобразуем выражение:

$$\frac{\partial L}{\partial\boldsymbol{w}^k}=\frac{\partial}{\partial\boldsymbol{w}^k}\sum_{i=1}^{m}\gamma_{ik}\ln p(y^i\,|\,\boldsymbol{x}^i,\boldsymbol{w}^k)=0. \tag{1.13}$$

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

$$\sum_{i=1}^{m} \gamma_{ik} \ln p(y^i \,|\, \boldsymbol{x}^i, \boldsymbol{w}^k) \to \max_{\boldsymbol{w}^k}. \tag{1.14}$$

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

$$\gamma_{ik} = \frac{\pi_k p(y^i \,|\, \boldsymbol{x}^i, \boldsymbol{w}_k)}{\sum_{s=1}^{l} \pi_s p(y^i \,|\, \boldsymbol{x}^i, \boldsymbol{w}_s)}. \tag{1.15}$$

M-шаг заключается в оптимизации параметров распределений.

$$Q(\boldsymbol{w}^1, \ldots, \boldsymbol{w}^l \,|\, \boldsymbol{\pi}) \to max \tag{1.16}$$

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

$$G = (\boldsymbol{\gamma}_1, \ldots, \boldsymbol{\gamma}_l) = \begin{pmatrix} \gamma_{11} & \cdots & \gamma_{1l} \\ \vdots & \ddots & \vdots \\ \gamma_{m1} & \cdots & \gamma_{ml} \end{pmatrix}, \quad G_k = \mathrm{diag}(\boldsymbol{\gamma}_k). \tag{1.17}$$

Перейдем к рассмотрению линейный и обобщенных линейных моделей.

2 Оценка параметров смеси линейных моделей

Линейная модель имеет вид:

$$\boldsymbol{y} = X\boldsymbol{w} + \boldsymbol{\varepsilon}, \tag{2.1}$$

где $\boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{0}, B)$ — 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.

$$p(\boldsymbol{y} \,|\, X, \boldsymbol{w}) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{|\mathrm{det} B|}} \exp\left(-\frac{1}{2}(\boldsymbol{y} - X\boldsymbol{w})^\mathsf{T} B (\boldsymbol{y} - X\boldsymbol{w})\right). \tag{2.2}$$

Шаг M алгоритма примет следующий вид:

$$G_k \ln \left[\frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{|\det B|}} \right] - \frac{1}{2} \left(G_k (\boldsymbol{y} - X\boldsymbol{w})^\mathsf{T} B (\boldsymbol{y} - X\boldsymbol{w}) \right) \to \max_{\boldsymbol{w}} \tag{2.3}$$

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

$$\frac{1}{2} \boldsymbol{w}^\mathsf{T} X^\mathsf{T} G_k B X \boldsymbol{w} - \boldsymbol{w}^\mathsf{T} X^\mathsf{T} G_k B \boldsymbol{y} \to \min_{\boldsymbol{w}} \tag{2.4}$$

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

$$\boldsymbol{w}^* = \left(X^\mathsf{T} G_k B X \right)^{-1} G_k B X \boldsymbol{y}. \tag{2.5}$$

3 Оценка параметров смеси обобщенно-линейных моделей

В случае обобщенных линейный моделей функция плотности распределения имеет вид

$$p(\boldsymbol{y} \,|\, \boldsymbol{\theta}) = \exp \left(\boldsymbol{T}(\boldsymbol{y})^\mathsf{T} \boldsymbol{\eta}(\boldsymbol{\theta}) - b(\boldsymbol{\theta}) + c(\boldsymbol{y}) \right). \tag{3.1}$$

M-шаг алгоритма сводится к максимизации

$$G_k \boldsymbol{T}(\boldsymbol{y})^\mathsf{T} \boldsymbol{\eta}(\boldsymbol{\theta}) - G_k b(\boldsymbol{\theta}) + G_k c(\boldsymbol{y}) \to \max_{\boldsymbol{\theta}}. \tag{3.2}$$

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

$$G_k \boldsymbol{T}(\boldsymbol{y})^\mathsf{T} \boldsymbol{\eta}(\boldsymbol{\theta}) - G_k b(\boldsymbol{\theta}) \to \max_{\boldsymbol{\theta}}. \tag{3.3}$$

Дальнейшая минимизация зависит от конкретного семейства из обобщенного класса, вида функции $b(\theta)$.

4 Оценка параметров смеси экспертов

Понятие смеси экспертов было введено Якобсом (Jacobs) в 1991г. Предполагается, что параметры смеси π являются функциями от объекта, т.е.

$$p(\boldsymbol{y} \,|\, \boldsymbol{x}) = \sum_{k=1}^{l} \pi_k(\boldsymbol{x}) p(y \,|\, \boldsymbol{x}, \boldsymbol{w}_k). \tag{4.1}$$

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.