

Оценка параметров смеси распределений

Кирилл Павлов, 674 группа

7 декабря 2011 г.

1 Введение

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

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

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

$$\sum_{k=1}^l \pi_k = 1. \quad (1.2)$$

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

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

Введем функцию правдоподобия $Q(\mathbf{w}_1, \dots, \mathbf{w}_l, \boldsymbol{\pi})$ как логарифм плотности вероятности данных.

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

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

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

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

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

Подставим это равенство в формулу Байеса для γ_{ik}

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

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

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

Приравняем производные функции Лагранжа к нулю:

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

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

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

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 :

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

Производная лагранжиана по параметрам k -й модели:

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

Преобразуем выражение:

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

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

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

В общем случае задача оптимизации $Q(\mathbf{w}^1, \dots, \mathbf{w}^l, \boldsymbol{\pi}) \rightarrow \max$ 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 приближению параметров моделей $(\mathbf{w}_1, \dots, \mathbf{w}_l)$. На M -шаге решается задача максимизации правдоподобия Q при начальном приближении параметров моделей и значений γ_{ik} .

E -шагу соответствует выражение

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

M -шаг заключается в оптимизации параметров распределений.

$$Q(\mathbf{w}^1, \dots, \mathbf{w}^l | \boldsymbol{\pi}) \rightarrow \max \quad (1.16)$$

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

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

Перейдем к рассмотрению линейных и обобщенных линейных моделей.

2 Оценка параметров смеси линейных моделей

Линейная модель имеет вид:

$$\mathbf{y} = X\mathbf{w} + \boldsymbol{\varepsilon}, \quad (2.1)$$

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

$$p(\mathbf{y} | X, \mathbf{w}) = \frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{|\det B|}} \exp\left(-\frac{1}{2}(\mathbf{y} - X\mathbf{w})^\top B(\mathbf{y} - X\mathbf{w})\right). \quad (2.2)$$

Шаг M алгоритма примет следующий вид:

$$G_k \ln \left[\frac{1}{(2\pi)^{\frac{n}{2}} \sqrt{|\det B|}} \right] - \frac{1}{2} (G_k(\mathbf{y} - X\mathbf{w})^\top B(\mathbf{y} - X\mathbf{w})) \rightarrow \max_{\mathbf{w}} \quad (2.3)$$

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

$$\frac{1}{2} \mathbf{w}^\top X^\top G_k B X \mathbf{w} - \mathbf{w}^\top X^\top G_k B \mathbf{y} \rightarrow \min_{\mathbf{w}} \quad (2.4)$$

Задача квадратична по \mathbf{w} , решение находится аналитически

$$\mathbf{w}^* = (X^\top G_k B X)^{-1} G_k B X \mathbf{y}. \quad (2.5)$$

3 Оценка параметров смеси обобщенно-линейных моделей

В случае обобщенных линейных моделей функция плотности распределения имеет вид

$$p(\mathbf{y} | \boldsymbol{\theta}) = \exp(\mathbf{T}(\mathbf{y})^\top \boldsymbol{\eta}(\boldsymbol{\theta}) - b(\boldsymbol{\theta}) + c(\mathbf{y})). \quad (3.1)$$

M -шаг алгоритма сводится к максимизации

$$G_k \mathbf{T}(\mathbf{y})^\top \boldsymbol{\eta}(\boldsymbol{\theta}) - G_k b(\boldsymbol{\theta}) + G_k c(\mathbf{y}) \rightarrow \max_{\boldsymbol{\theta}}. \quad (3.2)$$

Последнее слагаемое не зависит от параметров модели $\boldsymbol{\theta}$, что позволяет упростить функционал

$$G_k \mathbf{T}(\mathbf{y})^\top \boldsymbol{\eta}(\boldsymbol{\theta}) - G_k b(\boldsymbol{\theta}) \rightarrow \max_{\boldsymbol{\theta}}. \quad (3.3)$$

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

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.

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

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

Оказывается (Jordan and Jacobs, 1994), что наличие функции компетенции допускает решение задачи с помощью *EM*-алгоритма, причем, *E*-шаг остается прежним:

$$\gamma_{ik} = \frac{\pi_k(\mathbf{x}^i) p(y^i | \mathbf{x}^i, \mathbf{w}_k)}{\sum_{s=1}^l \pi_s(\mathbf{x}^i) p(y^i | \mathbf{x}^i, \mathbf{w}_s)}. \quad (4.2)$$

M-шаг принимает вид:

$$\pi_k = \frac{1}{m} \sum_{i=1}^m g_{ik}. \quad (4.3)$$

$$\sum_{i=1}^m \gamma_{ik}(\mathbf{x}^i) \ln p(y^i | \mathbf{x}^i, \mathbf{w}^k) \rightarrow \max_{\mathbf{w}^k}. \quad (4.4)$$

Уравнение 4.4 можно решить с помощью метода итеративно перевзвешенных наименьших квадратов (IRLS).