Московский государственный университет
имени М.В. Ломоносова
Механико-математический факультет
Кафедра математической теории
интеллектуальных систем

Оценка максимума правдоподобия параметров прямоугольнообразных распределений в гауссовском случае
Maximum likelihood estimation of the parameters of the rectangular-shaped distributions in the Gaussian case

Курсовая работа
Нерсисяна Степана Ашотовича
507 группа

Научный руководитель:
старший научный сотрудник
кафедры математической
теории интеллектуальных систем
к.ф.-м.н. Галатенко Алексей Владимирович

Москва, 2020

1 Введение

Параметрические семейства непрерывных распределений используются в огромном числе приложений для моделирования различных случайных величин. При этом семейство распределений зачастую выбирается с использованием принципа максимума энтропии, так, нормальное распределение обладает наибольшей энтропией среди всех непрерывных распределений на числовой прямой с фиксированным математическим ожиданием и дисперсией [1], в связи с чем активно используется в огромном количестве приложений. Например, в большом количестве исследований постулируется нормальность распределения экспрессии генов в гомогенной выборке образцов, хотя такой подход может подвергаться критике [2]. Более очевидным примером является равномерное распределение на отрезке $[a, b]$, содержащее в себе наибольшую "неопределенность" среди всех распределений с тем же носителем.

Однако, использование равномерного распределения в приложениях затрудняется его нерегулярностью, в частности, зависимостью носителя от параметров распределения. Для борьбы с этим явлением можно использовать распределения, плотность которых является приближением к плотности равномерного распределения с носителем, совпадающим со всей числовой прямой. Примером такого приближения может служить обобщенное нормальное распределение [3].

К сожалению использование такого подхода не решает проблему абсолютной неустойчивости равномерного распределени к выбросам в данных. Для борьбы с этим явлением в работе [4] было предложено семейство распределений, полученное с помощью "пришивания" хвостов нормального распределения к равномерной плотности. Авторы использовали свое распределение для решения задачи кластеризации, применяя технику разделения смеси распределений с помощью EM-алгоритма. Однако, в работе не сформулированы алгоритмы, выполение которых гарантированно приводит к глобальному максимуму функции правдоподобия. Более того, параметр, отвечающих за ширину нормальных хвостов, считался фиксированным.

В настоящей работе приводится обобщение данной конструкции (прямоугольнообразные распределения) и эффективный алгоритм для поиска глобального максимума функции правдоподобия в случае нормальных хвостов. Параметрами оптимизации являются не только границы отрезка равномерности, но и размер хвостов распределения. Также приведены результаты вычислительных экспериментов, демонстрирующие состоятельность алгоритма.

2 Построение распределения

Рассмотрим функцию

$$k(x|l, h) = \max(0, x - h) - \max(0, l - x)$$

при $h \geq l$, см. Рис. 1 для примера. Будем использовать ее, чтобы "расширить" моду некоторого распределения, получив плотность, форма которой близка к прямоугольнику. А именно, рассмотрим распределение, заданное четной плотностью $q(x|\alpha)$, где $x \in \mathbb{R}$ и α – набор параметров. Подстановка функции k в аргумент q позволяет создать новое распределение, имеющее плато на отрезке $[l, h]$ (прямоугольнообразное распределение):

$$RSD(x|l, h, \alpha) \propto q(k(x|l, h)|\alpha)$$

Для дальнейшей работы с этим распределением необходимо посчитать нормировочную константу:

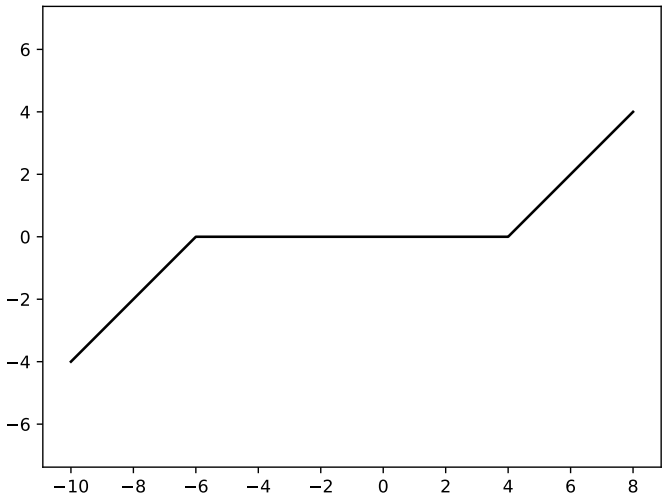

Рис. 1: График функции $k(x|l,h)$ при $l=-6,\ h=4$.

$$\int_{-\infty}^{+\infty} q(k(x|l,h)|\alpha)dx = \int_{-\infty}^{l} q(x-l|\alpha)dx + \int_{l}^{h} q(0|\alpha)dx + \int_{h}^{+\infty} q(x-h|\alpha)dx =$$
$$= 1 + q(0|\alpha)(h-l)$$

Таким образом,

$$RSD(x|l,h,\alpha) = \frac{1}{1+q(0|\alpha)(h-l)} q(k(x|l,h)|\alpha)$$

3 Оценка максимума правдоподобия

Рассмотрим задачу поиска максимума функции правдоподобия для прямоугольнообразных распределений. Дана выборка $D=(x_1,\ldots,x_N)$, где $x_1 < x_2 < \ldots < x_N$, и соответствующие веса $r_1,\ldots,r_N$ (их рассмотрение позволит перенести все полученные результаты на M-шаг EM-алгоритма). Нас интересует решение задачи оптимизации

$$NLL(l,h,\alpha) = -\sum_{i=1}^{N} r_i \log RSD(x_i|l,h,\alpha) =$$
$$= \sum_{i=1}^{N} r_i \left(\log\left(1+q(0|\alpha)(h-l)\right) - \log q(k(x_i|l,h)|\alpha)\right) \to \min$$

Также рассмотрим случай, когда q – плотность нормального распределения:

$$q(x|\sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-x^2/2\sigma^2}$$

Пример графика плотности распределения представлен на Рис. 2

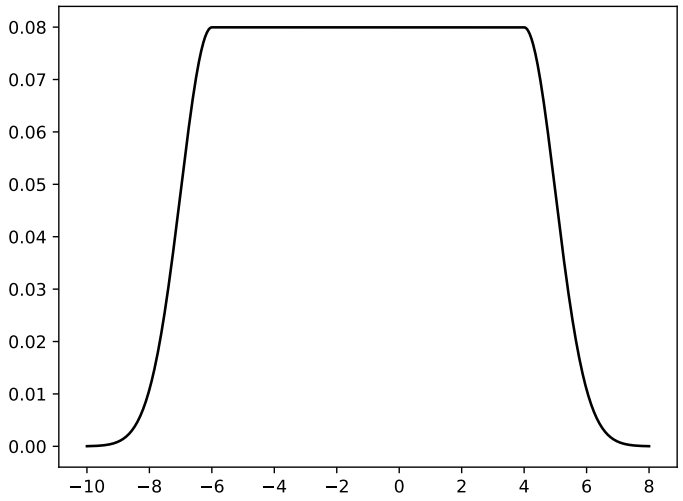

Рис. 2: График плотности гауссовской RSD при $l = -6,\ h = 4,\ \sigma = 1$.

Имеем

$$NLL(l,h,\sigma) = \sum_{i=1}^{N} r_i \left(\log\left(1 + \frac{1}{\sqrt{2\pi}\sigma}(h-l)\right) + \log\left(\sqrt{2\pi}\sigma\right) + \frac{k^2(x_i|l,h)}{2\sigma^2} \right) =$$

$$= r\log\left(\sqrt{2\pi}\sigma + h - l\right) + \frac{1}{2\sigma^2}\sum_{i=1}^{N} r_i \left(\max(0, x_i - h) - \max(0, l - x_i)\right)^2 \to \min,$$

где

$$r = \sum_{i=1}^{N} r_i$$

Функцию удобно рассматривать на областях

$$S_{u,v} = \{l, h : l \le h, l \in [x_u, x_{u+1}), h \in (x_{v-1}, x_v]\},$$

где $1 \le u < v \le N$ (см. Рис. 3); вне объединения этих областей минимум, очевидно, не достигается. Рассмотрим сужение функции $NLL(l,h,\sigma)$ на область $S_{u,v}$ (для простоты обозначим NLL за f):

$$f(l,h,\sigma) = r\log\left(\sqrt{2\pi}\sigma + h - l\right) + \frac{1}{2\sigma^2}\left(\sum_{i=1}^{u} r_i(l - x_i)^2 + \sum_{i=v}^{N} r_i(x_i - h)^2\right)$$

Приравняем частные производные функции по l, h к нулю:

$$\frac{\partial f}{\partial l} = -\frac{r}{\sqrt{2\pi}\sigma + h - l} + \frac{1}{\sigma^2}\sum_{i=1}^{u} r_i(l - x_i) = 0$$

$$\frac{\partial f}{\partial h} = \frac{r}{\sqrt{2\pi}\sigma + h - l} - \frac{1}{\sigma^2}\sum_{i=v}^{N} r_i(x_i - h) = 0$$

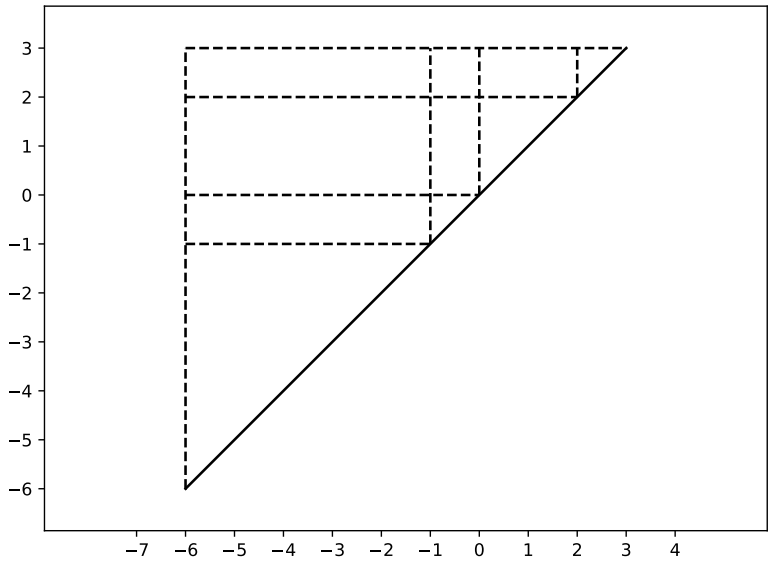

Рис. 3: Области $S_{u,v}$, построенные по выборке $D = \{-6, -1, 0, 2, 3\}$.

Складывая уравнения, получаем:

$$l \sum_{i=1}^{u} r_i + h \sum_{i=v}^{N} r_i = \sum_{i=1}^{u} r_i x_i + \sum_{i=v}^{N} r_i x_i \tag{1}$$

Для удобства введем сокращения:

$$r^l = \sum_{i=1}^{u} r_i,\ r^h = \sum_{i=v}^{N} r_i,\ x^l = \sum_{i=1}^{u} r_i x_i,\ x^h = \sum_{i=v}^{N} r_i x_i$$

Тогда уравнение (1) принимает вид

$$r^l l + r^h h = x^l + x^h \tag{2}$$

Рассмотрим $u = 1, v = N$. На нем уравнение (2) выглядит как

$$r_1 l + r_N h = r_1 x_1 + r_N x_N$$

Точка $l = x_1$, $h = x_N$, очевидно, ему удовлетворяет. Прямая, выходящая из этой точки, пересекает область $S_{1,N}$ либо справа, либо снизу. В случае пересечения справа перейдем в область $S_{2,N}$, иначе в область $S_{1,N-1}$. Лемма 1 обеспечивает тот факт, что прямая, соответствующая новой области, будет проходить через ту же точку пересечения. В то же время все прямые, ассоциированные с областями $S_{1,v}, v < N$ в случае пересечения справа, и прямые, ассоциированные с областями $S_{u,1}, u > 1$ в случае пересечения снизу, не будут пересекаться со своими областями в силу Леммы 2. Далее, рассуждение повторяется: прямая, соответствующая новой области ($S_{2,N}$ или $S_{1,N-1}$) пересекает ее границу либо справа, либо снизу. Продолжая процесс движения вправо-вниз рано или поздно дойдем до пересечения прямой $l = h$, что соответствует области $S_{u,v}$, где $v - u = 1$. Таким образом, объединение прямых, заданных уравнением (2), по всем областям $S_{u,v}$ определяет ломаную, исходящую из точки (x_1, x_N) и входящую в прямую $l = h$. Псевдокод по обходу данной ломаной представлен в Алгоритме 1.

$u, v \leftarrow 1, N$;
$l_{current}, h_{current} \leftarrow x_1, x_N$;
$r^l, r^h \leftarrow r_1, r_N$;
$x^l, x^h \leftarrow r_1 x_1, r_N x_N$;
while $v - u \geq 1$ **do**
  if $v - u > 1$ **then**
    $l_{next} \leftarrow \min(x_{u+1}, \frac{x^l + x^h - r^h x_{v-1}}{r^l})$;
    $h_{next} \leftarrow \max(x_{v-1}, \frac{x^l + x^h - r^l x_{u+1}}{r^h})$;
  else
    $l_{next} \leftarrow \frac{x^l + x^h}{r^l + r^h}$;
    $h_{next} \leftarrow \frac{x^l + x^h}{r^l + r^h}$;
  end
  if $l_{next} = x_{u+1}$ **then**
    $r^l \leftarrow r^l + r_{u+1}$;
    $x^l \leftarrow x^l + r_{u+1} x_{u+1}$;
    $u \leftarrow u + 1$;
  else
    $r^h \leftarrow r^h + r_{v-1}$;
    $x^h \leftarrow x^h + r_{v-1} x_{v-1}$;
    $v \leftarrow v - 1$;
  end
  $l_{current}, h_{current} \leftarrow l_{next}, h_{next}$;
end

Algorithm 1: Обход критической ломаной.

Таким образом, для минимизации функционала достаточно минимизировать его на ломаной. Сузим функцию f на звено ломаной: для этого выразим h из уравнения (2):

$$h = \frac{x^l + x^h}{r^h} - \frac{r^l}{r^h} l$$

Введем обозначения:

$$\begin{aligned} t &= r^l l - x^l \\ a &= \frac{1}{r^l} + \frac{1}{r^h} \\ b &= \frac{x^h}{r^h} - \frac{x^l}{r^l} \\ c &= \sum_{i=1}^{u} r_i x_i^2 + \sum_{i=v}^{N} r_i x_i^2 - \frac{(x^l)^2}{r^l} - \frac{(x^h)^2}{r^h} \end{aligned}$$

Тогда по Лемме 3 имеем

$$\begin{aligned} h - l &= -at + b \\ \sum_{i=1}^{u} r_i (l - x_i)^2 + \sum_{i=v}^{N} r_i (x_i - h)^2 &= at^2 + c \end{aligned}$$

и

$$g(t, \sigma) = f(l(t), h(t), \sigma) = r \log\left(\sqrt{2\pi}\sigma - at + b\right) + \frac{1}{2\sigma^2}\left(at^2 + c\right).$$

Приравняем частные производные по t, σ к нулю:

$$\frac{\partial g}{\partial t} = -\frac{ra}{\sqrt{2\pi}\sigma - at + b} + \frac{a}{\sigma^2}t = 0$$
$$\frac{\partial g}{\partial \sigma} = \frac{\sqrt{2\pi}r}{\sqrt{2\pi}\sigma - at + b} - \frac{1}{\sigma^3}\left(at^2 + c\right) = 0$$

Сократим в первом уравнении на a, после чего домножим его на $\sqrt{2\pi}$ и сложим со вторым:

$$\sqrt{2\pi}\sigma t = at^2 + c$$

С учетом этого преобразуем первое уравнение:

$$r\sigma^2 = \sqrt{2\pi}\sigma t - at^2 + bt = bt + c$$

Подставляя сюда

$$\sigma = \frac{at^2 + c}{\sqrt{2\pi}t}$$

получаем

$$r(at^2 + c)^2 = 2\pi t^2(bt + c)$$

или

$$ra^2t^4 - 2\pi bt^3 + 2c(ra - \pi)t^2 + rc^2 = 0$$

Остается решить данное полиномиальное уравнение 4-й степени, после чего выбрать точку локального минимума из полученных решений и границ соответствующего звена ломаной. Таким образом, весь алгоритм имеет сложность $O(N)$.

4 Вычислительные эксперименты

Алгоритм был реализован в виде класса на языке Python 3. Метод был применен к известной выборке данных Iris, содержащей численные характеристики 150 ирисов. Оценка максимального правдоподобия была получена для каждой из компонент данных независимо, см Рис. 4. На этих данных видно разнообразие форм, которые может принимать плотность распределения: от равномерной (petal width) до нормальной (sepal width) плотностей.

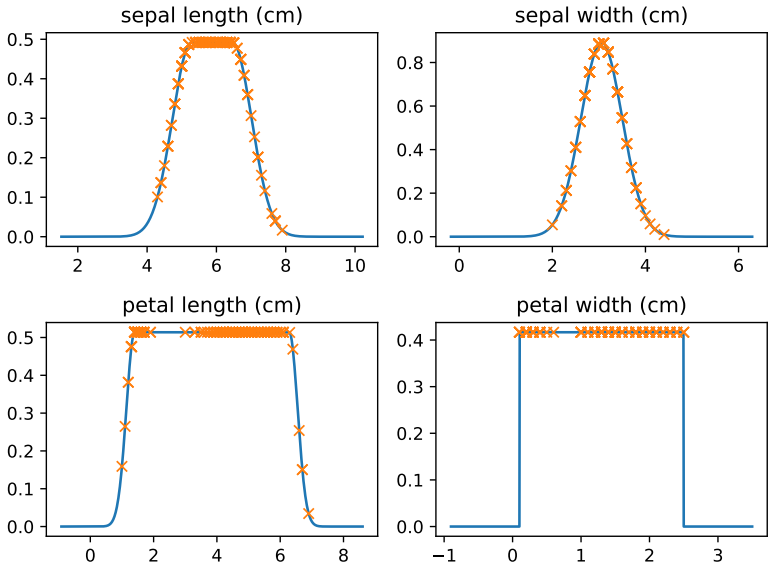

Рис. 4: Оценка максимального правдоподобия для каждой компоненты выборки Iris.

5 Вспомогательные результаты

Лемма 1. *При $v - u > 1$ верны следующие утверждения:*

1. *Прямые, заданные уравнением (1) в областях $S_{u,v}$ и $S_{u+1,v}$ пересекаются при $l = x_{u+1}$;*
2. *Прямые, заданные уравнением (1) в областях $S_{u,v}$ и $S_{u,v-1}$ пересекаются при $h = x_{v-1}$.*

Доказательство. Подставим $l = x_{u+1}$ в уравнение (1), записанное для областей $S_{u,v}$ и $S_{u+1,v}$. Легко видеть, что в обоих случаях получим

$$h = \frac{\sum_{i=1}^{u} r_i(x_i - x_{u+1}) + \sum_{i=v}^{N} r_i x_i}{\sum_{i=v}^{N} r_i}.$$

Аналогично, если подставить $h = x_v$ в уравнение (1), записанное для областей $S_{u,v}$ и $S_{u,v-1}$, получим

$$l = \frac{\sum_{i=1}^{u} r_i x_i + \sum_{i=v}^{N} r_i(x_i - x_{v-1})}{\sum_{i=1}^{u} r_i}.$$

□

Лемма 2. *Верны следующие утверждения:*

1. *Если прямая, заданная уравнением (1) в области $S_{u,v}$, проходит через правую сторону области ($l = x_{u+1}$), то все области $S_{u,\widetilde{v}}$, где $\widetilde{v} < v$, не имеют пересечений со своей прямой;*
2. *Если прямая, заданная уравнением (1) в области $S_{u,v}$, проходит через нижнюю сторону области ($h = x_{v-1}$), то все области $S_{\widetilde{u},v}$, где $\widetilde{u} > u$, не имеют пересечений со своей прямой.*

Доказательство. Докажем пункт 1. Из геометрических соображений пересечение прямой с правой стороной области означает, что пересечение той же прямой с прямой $h = x_{v-1}$ произойдет при $l > x_{u+1}$. Применяя лемму 1 получаем, что прямая, заданная уравнением в области $S_{u,v-1}$, лежит в некоторой области, расположенной правее, и не пересекается с $S_{u,v-1}$ в силу направленности прямой. В силу тех же геометрических соображений, прямая, заданная областью $S_{u,v-1}$, пересекается с прямой $h = x_{v-2}$ при $l > x_{u+2}$. Дальнейшее доказательство проводится повторением рассуждений. Пункт 2 доказывается аналогично. □

Лемма 3. *Верны следующие равенства:*

$$h - l = -\left(r^l l - x^l\right)\left(\frac{1}{r^l} + \frac{1}{r^h}\right) + \frac{x^h}{r^h} - \frac{x^l}{r^l}$$

$$\sum_{i=1}^{u} r_i(l - x_i)^2 + \sum_{i=v}^{N} r_i(x_i - h)^2 = \left(r^l l - x^l\right)^2\left(\frac{1}{r^l} + \frac{1}{r^h}\right) + \sum_{i=1}^{u} r_i x_i^2 + \sum_{i=v}^{N} r_i x_i^2 - \frac{(x^l)^2}{r^l} - \frac{(x^h)^2}{r^h}$$

Доказательство. Доказательство производится прямым вычислением:

$$h - l = \frac{x^l + x^h}{r^h} - \left(1 + \frac{r^l}{r^h}\right) l = -r^l l\left(\frac{1}{r^l} + \frac{1}{r^h}\right) + x^l\left(\frac{1}{r^l} + \frac{1}{r^h}\right) + \frac{x^h}{r^h} - \frac{x^l}{r^l} =$$
$$= -\left(r^l l - x^l\right)\left(\frac{1}{r^l} + \frac{1}{r^h}\right) + \frac{x^h}{r^h} - \frac{x^l}{r^l}$$

$$\sum_{i=1}^{u} r_i(l-x_i)^2 = \sum_{i=1}^{u} r_i\left(l^2 - 2lx_i + x_i^2\right) = r^l l^2 - 2lx^l + \sum_{i=1}^{u} r_i x_i^2 = \frac{1}{r^l}\left(r^l l - x^l\right)^2 + \sum_{i=1}^{u} r_i x_i^2 - \frac{(x^l)^2}{r^l}$$

$$\sum_{i=v}^{N} r_i(x_i-h)^2 = \sum_{i=v}^{N} r_i\left(\frac{r^l}{r^h}l - \frac{x^l+x^h}{r^h} + x_i\right)^2 =$$

$$= \sum_{i=v}^{N} r_i\left(\left(\frac{r^l}{r^h}l\right)^2 + \left(\frac{x^l+x^h}{r^h}\right)^2 + x_i^2 - 2r^l l\frac{x^l+x^h}{(r^h)^2} + 2\frac{r^l}{r^h}lx_i - 2x_i\frac{x^l+x^h}{r^h}\right) =$$

$$= \frac{(r^l l)^2}{r^h} + \frac{(x^l+x^h)^2}{r^h} + \sum_{i=v}^{N} r_i x_i^2 - 2r^l l\frac{x^l+x^h}{r^h} + 2\frac{r^l}{r^h}x^h l - 2x^h\frac{x^l+x^h}{r^h} =$$

$$= \frac{1}{r^h}\left((r^l l)^2 - 2r^l lx^l + (x^l)^2\right) + \frac{2x^l x^h + (x^h)^2}{r^h} - 2x^h\frac{x^l+x^h}{r^h} + \sum_{i=v}^{N} r_i x_i^2 =$$

$$= \frac{1}{r^h}\left(r^l l - x^l\right)^2 + \sum_{i=v}^{N} r_i x_i^2 - \frac{(x^h)^2}{r^h}$$

□

Список литературы

[1] Murphy KP. Machine learning : a probabilistic perspective. Cambridge, Mass. [u.a.]: MIT Press, 2013.

[2] Hebenstreit D, Fang M, Gu M, Charoensawan V, van Oudenaarden A, Teichmann SA. RNA sequencing reveals two major classes of gene expression levels in metazoan cells. Mol Syst Biol. 2011 Jun 7;7:497.

[3] Mahesh KV, Behnaam A.Parametric generalized Gaussian density estimation. J. Acoust. Soc. Am. 1989 Jun 86;1404.

[4] Pelleg D, Moore A. Mixtures of rectangles: Interpretable soft clustering. Proc. 18th International Conf. on Machine Learning. 2001;401-408.