Непараметрические байесовские методы. Процессы Дирихле.

Антон Осокин

15 сентября 2014 г.

Аннотация

Этот конспект посвящён описанию случайных процессов Дирихле и методов работы с ними. Основная задача конспекта состоит в определении понятия процесса Дирихле "на пальцах" и в изложении алгоритмов, которые можно использовать на практике. При этом целый ряд важных теоретических моментов обходится стороной или даётся без доказательств.

Конспект построен следующим образом. Раздел 1 содержит основные факты о распределении Дирихле, особое внимание уделяется методам генерации выборки из распределения Дирихле. Раздел 2 вводит понятие процесса Дирихле и описывает его различные представления. Раздел 3 содержит описание модели смеси распределений с априорным распределением, заданным в виде процесса Дирихле. Приводятся алгоритмы приближённого байесовского вывода в данной модели: алгоритмы MCMC (Monte-Carlo Markov chain) и алгоритм вариационного вывода. Разделы 4 и 5 содержат краткие описания расширений процесса Дирихле и его приложений.

1 Распределение Дирихле и его свойства

Основной источник материала по распределению Дирихле и его свойствам – технический отчёт [7].

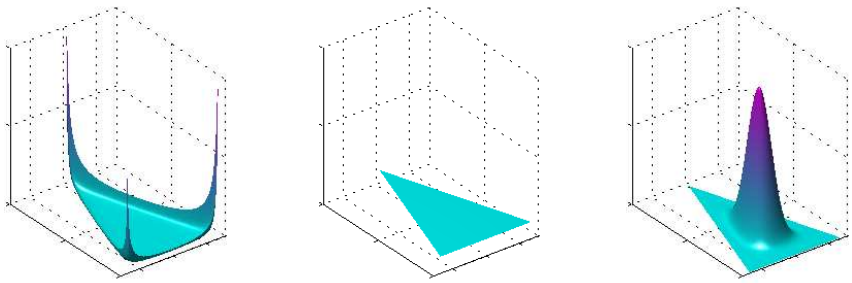
1.1 Определение и основные свойства

Определение. Распределение Дирихле – непрерывное распределение вероятностей, носителем которого является k-мерный симплекс $\Delta_k = \{\boldsymbol{q} \in \mathbb{R}^k \mid \sum_{i=1}^k q_i = 1,\ q_j \geq 0, j = 1, \dots, k\}$.

Плотность вероятности распределения Дирихле задаётся следующей формулой:

$$\mathrm{Dir}(\boldsymbol{q} \mid \boldsymbol{\alpha}) = \frac{\Gamma(\sum_{i=1}^k \alpha_i)}{\prod_{i=1}^k \Gamma(\alpha_i)} \prod_{i=1}^k q_i^{\alpha_i - 1}, \quad \boldsymbol{q} \in \Delta_k.$$

где $\boldsymbol{\alpha} = (\alpha_i)_{i=1}^k$, $\alpha_i > 0$ – вектор параметров распределения, $\Gamma(\cdot)$ – гамма-функция. Графики плотности распределения Дирихле при различных значениях параметров приведены на рис. 1.



$\alpha_1 = \alpha_2 = \alpha_3 = 0.1$ $\alpha_1 = \alpha_2 = \alpha_3 = 1$ $\alpha_1 = \alpha_2 = \alpha_3 = 10$

Рис. 1: Различные виды плотности распределения Дирихле для $k = 3$.

Связь с бета-распределением. В случае $k = 2$ распределение Дирихле тесно связано с бета-распределением:

$$\text{Beta}(x \mid \alpha, \beta) = \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \quad 0 \le x \le 1,$$

где α и β – параметры, $\alpha > 0$, $\beta > 0$. Если $x \sim \text{Beta}(x \mid \alpha, \beta)$, то $(x, 1-x) \sim \text{Dir}((x, 1-x) \mid (\alpha, \beta))$.

Связь с мультиномиальным распределением. Распределение Дирихле является сопряжённым к мультиномиальному распределению:

$$\text{Mult}(\boldsymbol{x} \mid k, n, \boldsymbol{q}) = \frac{n!}{x_1! \dots x_k!} \prod_{i=1}^{k} q_i^{x_i}, \quad x_i = 0, \dots, n,\ i = 1, \dots, k,\ \sum_{i=1}^{k} x_i = n,$$

где $n \in \mathbb{N}$, $k \in \mathbb{N}$, $\boldsymbol{q} \in \Delta_k$ – параметры распределения. n – общее количество испытаний, k – количество возможных исходов каждого из испытаний, $\boldsymbol{q}$ – вектор вероятностей выпадения каждого из исходов в каждом испытании.

Если $\boldsymbol{x} \sim \text{Mult}(\boldsymbol{x} \mid k, n, \boldsymbol{q})$ и на $\boldsymbol{q}$ задано априорное распределение Дирихле $\boldsymbol{q} \sim \text{Dir}(\boldsymbol{q} \mid \boldsymbol{\alpha})$, то апостериорное распределение на $\boldsymbol{q}$ также является распределением Дирихле: $p(\boldsymbol{q} \mid \boldsymbol{x}) = \text{Dir}(\boldsymbol{q} \mid \boldsymbol{\alpha} + \boldsymbol{x})$. Доказательство приведено в [7, п. 1.2].

Свойство накопления (aggregation property). Суммы компонент вектора, распределённого по Дирихле, также распределены по Дирихле.

Пусть $\boldsymbol{q} \sim \text{Dir}(\boldsymbol{q} \mid \boldsymbol{\alpha})$ и пусть множество $\{1, \dots, k\}$ разбито на ℓ непересекающихся множеств $A_1, \dots, A_\ell$. Тогда

$$\left(\sum_{i \in A_1} q_i, \dots, \sum_{i \in A_\ell} q_i\right) \sim \text{Dir}\left(\hat{\boldsymbol{q}} \;\middle|\; \sum_{i \in A_1} \alpha_i, \cdots \sum_{i \in A_\ell} \alpha_i\right), \quad \hat{\boldsymbol{q}} \in \Delta_\ell.$$

Доказательство свойства накопления приведено в [7, п. 2.3.1].

Нейтральность (neutrality). Каждая компонента вектора $\boldsymbol{q} \sim \text{Dir}(\boldsymbol{\alpha})$ влияет на распределение остальных только через нормировку. Т.е. случайная величина q_i и случайный вектор $\left(\frac{1}{1-q_i}\boldsymbol{q}_{\setminus i}\right)$ являются независимыми.[1] Доказательство свойства нейтральности распределения Дирихле приведено в [7, п. 2.2.2].

[1] Символами $\boldsymbol{q}_{\setminus i}$ обозначается вектор $\boldsymbol{q}$, из которого изъята компонента номер i.

Маргинальные распределения. Следствием свойства накопления является то, что маргинальное распределение одной компоненты q_i вектора $\boldsymbol{q}$, распределённого по Дирихле, является бета-распределением. Если $\boldsymbol{q} \sim \mathrm{Dir}(\boldsymbol{q} \mid \boldsymbol{\alpha})$, то $q_i \sim \mathrm{Beta}(q_i \mid \alpha_i, \alpha_0 - \alpha_i)$, где $\alpha_0 = \sum_{i=1}^{k} \alpha_i$.

Также можно показать, что $\left(\frac{1}{1-q_i}\boldsymbol{q}_{\setminus i}\right) \sim \mathrm{Dir}(\boldsymbol{\alpha}_{\setminus i})$.

1.2 Генерация выборки из распределения Дирихле

1.2.1 Набор гамма-распределений

Один из наиболее практичных способов генерации выборки из распределения Дирихле состоит в генерации набора k величин из гамма-распределения[2] с параметрами, соответствующими параметрам распределения Дирихле, и последующей нормировке:

$$c_i \sim \mathrm{Gam}(\alpha_i, 1), \quad i = 1, \ldots, k,$$
$$q_i = \frac{c_i}{\sum_{j=1}^{k} c_j}, \quad i = 1, \ldots, k.$$

Доказательство корректности схемы приведено в [7, п. 2.3].

Недостатком этой схемы является то, что она не последовательная, т.е. для предъявления i-й компоненты генерируемого элемента, необходимо знать все k чисел c_i, $i = 1, \ldots, k$. Этот недостаток может быть критичен при больших k и, в частности, препятствует обобщению этой схемы на случай процесса Дирихле, в котором k бесконечно.

1.2.2 "Ломка палки" (Stick-breaking)

Подход "Ломка палки" представляет собой последовательную схему генерации компонент из маргинальных и условных распределений. Первая компонента генерируется из маргинального распределения $p(q_1)$, вторая – из условного распределения $p(q_2 \mid q_1)$, третья – из $p(q_3 \mid q_1, q_2)$ и т.д. Своё название данный подход получил из-за аналогии с итеративным отламыванием кусочков длины q_i от палки длины 1.

Из свойства накопления распределения Дирихле следует, что маргинальное распределение $p(q_1)$ – бета-распределение с параметрами α_1, $\sum_{i=2}^{k} \alpha_i$. Сгенерируем компоненту $q_1 = v_1 \sim \mathrm{Beta}(\alpha_1, \sum_{i=2}^{k} \alpha_i)$ при помощи метода, описанного в разделе 1.2.1 (бета-распределение – частный случай распределения Дирихле).

Дальнейшая схема основана на свойстве нейтральности распределения Дирихле. Выполним сэплирование первой компоненты v_2 из распределения $\mathrm{Dir}(\boldsymbol{\alpha}_{\setminus i})$ (аналогично шагу, описанному выше). Вторую компоненту генерируемого вектора $\boldsymbol{q}$ можно вычислить как $q_2 = v_2(1 - v_1)$. Таким образом выполняется $k-1$ шаг. В качестве последней k-й компоненты берется величина, обеспечивающая выполнение равенства $\sum_i q_i = 1$ (длина оставшейся палки).

[2] Гамма-распределение $\mathrm{Gam}(\lambda \mid a, b)$ – вероятностное распределение действительной положительной переменной λ. Плотность гамма-распределения имеет вид:

$$\mathrm{Gam}(\lambda \mid a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda), \quad \lambda > 0, a > 0, b > 0.$$

Итоговая схема:

$$v_i \sim \text{Beta}\left(\alpha_i,\ \sum_{j=i+1}^{k} \alpha_j\right), \quad i = 1, \ldots, k-1,$$
$$v_k = 1,$$
$$q_i = v_i \prod_{j=1}^{i-1}(1 - v_j), \quad i = 1, \ldots, k.$$

Легко убедиться, что равенство $\sum_{i=1}^{k} q_i = 1$ выполнено.

1.2.3 Урновая схема

Урновая схема (урны Пойя) для распределения Дирихле носит скорее теоретический характер, но её обобщение будет использоваться для генерации выборки из процесса Дирихле.

Пусть есть урна (ящик) и шары k различных цветов. Положим в урну по α_i шару каждого цвета (α_i может быть нецелым числом). На каждом шаге случайно вытягиваем 1 шар из урны (шар определённого цвета вытягивается из урны с вероятностью, пропорциональной количеству (нецелому) шаров этого цвета в урне), после чего возвращаем его обратно, добавив ещё один шар такого же цвета. При бесконечном количестве шагов пропорции шаров разных цветов будут составлять вектор, сгенерированный из распределения Дирихле $\text{Dir}(\boldsymbol{\alpha})$.

2 Процессы Дирихле

Основными источниками материалов по процессам Дирихле являются технический отчёт [7], обзорная статья [15], материалы лекций [13].

2.1 Определение

Случайный процесс – это функция двух аргументов $\xi(\omega, x) : \Omega \times \mathcal{X} \to \mathbb{R}$, где ω – элемент множества элементарных исходов Ω, отвечающий за случайность, а x – элемент множества индексов $\mathcal{X}$ (индексирующий элемент). Соответственно, множество $\xi(\cdot, x) = \{\xi(\omega, x)\}_{\omega\in\Omega}$ является случайной величиной, а множество $\xi(\omega, \cdot) = \{\xi(\omega, x)\}_{x\in\mathcal{X}}$ является функцией, отображающей $\mathcal{X}$ на $\mathbb{R}$.

Процесс Дирихле является случайной вероятностной мерой, т.е распределением вероятностей над вероятностными мерами (распределением над распределениями случайной величины). Множеством значений рассматриваемых случайных величин является вообще говоря бесконечное множество. В рамках данного конспекта будет считать таковым множеством $\mathbb{R}^d$. Таким образом, реализация случайного процесса $\xi(\omega, \cdot)$ – распределение над множеством $\mathbb{R}^d$. Индексирующий элемент – измеримое подмножество $\mathbb{R}^d$. Индексируемые случайные величины – $\xi(\cdot, A)$, где A – измеримое подмножество $\mathbb{R}^d$ (элемент σ-алгебры над $\mathbb{R}^d$).

Процесс Дирихле задаётся двумя параметрами: базовым распределение H – распределением (вероятностной мерой) над $\mathbb{R}^d$, и коэффициентом концентрации $\alpha > 0$. Будем говорить, что случайное распределение вероятностей (случайная вероятностная мера) G распределена согласно процессу Дирихле с параметрами H и α, если для любого конечного измеримого (“хорошего”) разбиения $(A_1, \ldots, A_n)$, $A_i \cap A_j = \emptyset$, $\bigcup_{i=1}^{n} A_i = \mathbb{R}^d$ случайный вектор $(G(A_1), \ldots, G(A_n))$ распределён согласно распределению Дирихле $\text{Dir}(\alpha H(A_1), \ldots, \alpha H(A_2))$. Обозначение: $G \sim \text{DP}(H, \alpha)$.

Таблица 1: Сравнение гауссовского процесса и процесса Дирихле

	Гауссовский процесс	Процесс Дирихле
Индексирующий элемент	вектор $x \in \mathbb{R}^d$	измеримое подмножество $A \subseteq \mathbb{R}^d$
Реализация	функция $f : \mathbb{R}^d \to \mathbb{R}$	вероятностная мера $G : 2^{\mathbb{R}^d} \to [0, 1]$
Параметры	$\mu(x)$ – мат. ожидание , $C(x', x'')$ – ковар. функция	H – вер. мера на $\mathbb{R}^d$, $\alpha > 0$ – коэфф. концентрации
Одномерная проекция	$\xi(\cdot, x) \sim \mathcal{N}\Big(\mu(x), C(x, x)\Big)$	$\xi(\cdot, A) \sim \text{Beta}\Big(\alpha H(A), \alpha(1-H(A))\Big)$
Многомерная проекция	$x_1, \ldots, x_n \in \mathbb{R}^d$ $(\xi(\cdot, x_1), \ldots, \xi(\cdot, x_n))$ $\sim \mathcal{N}\Big((\mu(x_i))_i, (C(x_i, x_j))_{i,j}\Big)$	$\{A_1, \ldots, A_n\}$ – разбиение $\mathbb{R}^d$ $(\xi(\cdot, A_1), \ldots, \xi(\cdot, A_n))$ $\sim \text{Dir}\Big(\alpha H(A_1), \ldots, \alpha H(A_n)\Big)$

Таблица 1 сопоставляет понятия, связанные с процессом Дирихле, с аналогами для гауссовского процесса.

Если бы множество значений случайных величин было бы конечным (а не $\mathbb{R}^d$): $\{\xi_1, \ldots, \xi_n\}$, то процесс Дирихле являлся бы распределением Дирихле над векторами $(G(\xi_1), \ldots G(\xi_n)) \sim \text{Dir}(\alpha H(\xi_1), \ldots, \alpha H(\xi_n))$.

Заметим, что данное выше определение неконструктивно, а значит возникает вопрос корректности, т.е. существования и единственности. Существует несколько подходов к доказательству корректности, но их рассмотрение выходит за рамки данного конспекта. Здесь же ограничимся лишь констатацией факта, что для любой вероятностной меры H над $\mathbb{R}^d$ и любого $\alpha > 0$ процесс Дирихле $\text{DP}(H, \alpha)$ существует и единственен.

Важным теоретическим свойством процесса Дирихле является то, что с вероятностью 1 реализация процесса Дирихле является дискретной вероятностной мерой, т.е. соответствующую плотность можно записать в виде обобщённой функции

$$p(x) = \sum_{i=1}^{\infty} p_i \delta_{\theta_i}(x), \tag{1}$$

где $\{p_i\}_{i=1}^{\infty}$, $p_i > 0$, $\sum_{i=1}^{\infty} p_i = 1$ и $\{\theta_i\}_{i=1}^{\infty}$, $\theta_i \in \mathbb{R}^d$ – последовательности случайных величин, а $\delta_{\theta_i}(x)$ – дельта-функция Дирака с параметром θ_i.[3] Величины p_i называются вероятностями атомов дискретной меры, величины θ_i – позициями атомов.

2.2 Условные распределения

В рамках данного текста мы имеем дело с распределениями над сложными структурами, а именно другими распределениями. Следующие результаты необходимы для того, чтобы осуществлять байесовский вывод в моделях с процессами Дирихле, а именно вычислять апостериорное распределение, если априорное распределение является процессом Дирихле.

[3]Неформально, дельта-функцией с параметром a является плотность распределения, в котором вся вероятностная масса сосредоточена в точке $a \in \mathbb{R}^d$. Это означает, что $\delta_a(x) \geq 0$, $\delta_a(x) = 0$ при $x \neq a$. В точке a дельта-функция принимает специальное значение, обеспечивающее выполнение равенства $\int_{\mathbb{R}^d} \delta_a(x)dx = 1$. Заметим, что для "обычной" функции $f : \mathbb{R}^d \to \mathbb{R}$ выполнено равенство $\int_{\mathbb{R}^d} f(x)\delta_a(x)dx = f(a)$.

Рассмотрим следующую простую вероятностную модель:

$$\begin{aligned} G &\sim \mathrm{DP}(H, \alpha), \\ \theta_1, \ldots, \theta_n \mid G &\sim G. \end{aligned} \tag{2}$$

Здесь сначала согласно априорному распределению $\mathrm{DP}(H, \alpha)$ выбирается вероятностная мера G. Затем из вероятностной меры G генерируется выборка точек $\theta_i \in \mathbb{R}^d$.

Можно показать [13, п. 2.4], что апостериорное распределение на меру G является процессом Дирихле:

$$G \mid \theta_1, \ldots, \theta_n \sim \mathrm{DP}\left(\frac{\alpha}{\alpha+n}H + \frac{1}{\alpha+n}\sum_{i=1}^{n}\delta_{\theta_i},\ \alpha+n\right). \tag{3}$$

Аналогично можно найти условное распределение на значение новой переменной θ_{n+1}, если значения переменных $\theta_1, \ldots, \theta_n$ известны:

$$\theta_{n+1} \mid \theta_1, \ldots, \theta_n \sim \frac{\alpha}{\alpha+n}H + \frac{1}{\alpha+n}\sum_{i=1}^{n}\delta_{\theta_i}. \tag{4}$$

Заметим, что при записи этих распределений допущена некоторая вольность, и мера H складывается с плотностями распределений δ_{θ_i}. Такую запись следует интерпретировать как сложение двух мер, одна из которых представлена плотностью распределения.

2.3 Представления процесса Дирихле

В данном разделе описывается два способа задавать процесс Дирихле, каждый из которых в дальнейшем будет использоваться в алгоритмах вывода.

2.3.1 Процесс “Китайский ресторан” (Chinese restaurant process, CRP)

Условное распределение (4) позволяет генерировать выборку $\theta_1, \ldots, \theta_n$ из модели (2) без генерации и хранения меры G в каком-либо виде. Заметим, что распределения, из которых генерируется выборка представляют собой смеси базового распределения H и распределений, задаваемыми дельта-функциями с центрами в уже сгенерированных точках. Процесс генерации можно записать так:

$$\theta_{n+1} \mid \theta_1, \ldots, \theta_n \sim \begin{cases} H, & \text{с вероятностью } \frac{\alpha}{\alpha+n}, \\ \delta_c, & \text{с вероятностью } \frac{\sum_{j=1}^{n}[\theta_j = c]}{\alpha+n}. \end{cases} \tag{5}$$

В первом случае значение θ_{n+1} с вероятностью 1 (если H – непрерывное распределение) не совпадает со значениями, присутствующими среди $\theta_1, \ldots, \theta_n$. Во втором же случае с вероятностью 1 выбирается значение, которое уже встречалось, причём вероятность увидеть конкретное значение тем больше, чем чаще оно уже наблюдалось (“богатый становится богаче”). Схема (5) часто называется урновой схемой Блэквела-МакКвина.

При работе схемы (5) на каждом шаге существует разбиение точек $\theta_1, \ldots, \theta_n$ на группы. Точки i и j относятся к одной группе, если значения θ_i и θ_j совпадают. Такие разбиения часто называют кластеризациями точек. Процесс, генерирующий кластеризации точек (игнорирующий конкретные значения θ_i), и распределение над кластеризациями (с неизвестным заранее числом кластеров) обычно называют процессом “Китайский ресторан” (обозначение: $\mathrm{CRP}(\alpha, n)$). Название произошло от следующей метафоры: посетитель номер $n+1$ заходит в ресторан и либо (с

вероятностью $\frac{\alpha}{\alpha+n}$) садится за свободный стол, либо (с вероятностью $\frac{\sum_{j=1}^{n}[\theta_j=c]}{\alpha+n}$) подсаживается за стол c.

Можно вычислить математическое ожидание количества кластеров m в кластеризации, генерируемой процессом "Китайский ресторан" $\mathrm{CRP}(\alpha, n)$. Заметим, что при генерации очередной метки i новый кластер образуется с вероятностью $\frac{\alpha}{\alpha+i-1}$, причём все эти события независимы. Отсюда следует, что мат. ожидание можно вычислить так:

$$\mathbb{E}_{\mathrm{CRP}(\alpha,n)}\, m = \sum_{i=1}^{n} \frac{\alpha}{\alpha+i-1} = \alpha(\psi(\alpha+n)+\psi(\alpha)) \simeq \alpha \log\left(1+\frac{n}{\alpha}\right), \quad \text{при } n \to +\infty.$$

Здесь $\psi(\cdot)$ – дигамма функция.

Обратим внимание на два момента. Во-первых, параметр α фактически линейно влияет на среднее количество кластеров. Подбирая параметр α, можно влиять на ожидаемое среднее количество кластеров. Во-вторых, количество кластеров растёт логарифмически с ростом числа точек n. В некоторых случаях такое поведение количества кластеров является нежелательным (рост числа кластеров слишком медленный). Существуют расширения процесса Дирихле (процесс Питмана-Йора), направленые на изменение этого свойства.

2.3.2 Процесс "Ломка палки"

Заметим, что процесс Дирихле с вероятностью 1 генерирует дискретные распределения, т.е. распределения вида (1). Здесь $\{p_i\}_{i=1}^{\infty}$ и $\{\theta_i\}_{i=1}^{\infty}$ – последовательности случайных величин. Величины θ_i определяют точки в $\mathbb{R}^d$, в которых концентрируется масса, величины p_i определяют вероятностную массу, расположенную в точках θ_i. Задавая эти две последовательности напрямую, можно получить конструктивное определение процесса Дирихле [13, п. 2.3]:

$$(\{p_i\}_{i=1}^{\infty}, \{\theta_i\}_{i=1}^{\infty}) \sim DP(\alpha, H).$$

Можно показать, что все величины θ_i независимы и распределены одинаково согласно закону H. Величины p_i не могут быть независимыми, поскольку сумма ряда $\sum_{i=1}^{\infty} p_i$ должна равняться 1. Пусть p_1 сгенерировано из некоторого одномерного распределения на отрезке [0,1], тогда величина p_2 должна генерироваться из распределения на отрезке $[0, 1-p_1]$, аналогично величина p_{n+1} – из отрезка $[0, 1-p_1-\cdots-p_n]$. Одним из способов генерации последовательности таких случайных величин является схема, аналогичная процедуре "Ломка палки" для генерации выборки из распределения Дирихле. Сначала из некоторого распределения на отрезке $[0,1]$ генерируется последовательность независимых одинаково-распределённых случайных величин $\{v_i\}_{i=1}^{\infty}$, затем величины p_i вычисляются при помощи перенормировки. Можно показать, что если все величины v_i генерируются из распределения $\mathrm{Beta}(1,\alpha)$, то (1) – реализации процесса Дирихле. Итоговая схема выглядит так:

$$\begin{aligned} &v_1, \ldots, v_i, \ldots \sim \mathrm{Beta}(1,\alpha), \quad p_i = v_i \prod_{j=1}^{i-1}(1-v_j), \\ &\theta_1, \ldots, \theta_i, \ldots \sim H. \end{aligned} \tag{6}$$

Обычно данная конструкция используется в алгоритмах, реализующих схему вариационного вывода. Распределение на $\{p_i\}_{i=1}^{\infty}$ соответствующее схеме (6) иногда обозначают $\mathrm{GEM}(\alpha)$, где аббревиатура соответствует первым буквам фамилий авторов.

Для корректного использования этого представления необходимо либо доказать его эквивалентность определениям данным выше, либо доказывать выполнения свойств (3), (4) исходя из этого представления. Рассмотрение этих вопросов выходит за рамки данного конспекта.

3 Смесь распределений с априорным распределением, заданным процессом Дирихле

3.1 Определение модели

С помощью процесса Дирихле может быть задана смесь распределений, состоящая из бесконечного числа компонент. Пусть каждая компонента смеси представляет собой распределение с параметрами $\theta \in \mathbb{R}^d$: $p(x \mid \theta)$. Пусть на параметры θ задано априорное распределение в виде реализации процесса Дирихле с параметрами H и $\alpha > 0$. Тогда генерация n наблюдаемых объектов выборки проводится по следующей схеме:

$$\begin{aligned} &G \sim \mathrm{DP}(H, \alpha), \\ &\theta_i \sim G, \quad i = 1, \ldots, n, \\ &x_i \sim p(x_i \mid \theta_i), \quad i = 1, \ldots, n \end{aligned} \tag{7}$$

Заметим, что если проинтегрировать совместное распределение на x_i и θ_i при условии меры G по θ_i, то получится смесь распределений с бесконечным количеством компонент:

$$\int_{\mathbb{R}^d} p(\theta_i \mid G) p(x_i \mid \theta_i) d\theta_i = \int_{\mathbb{R}^d} p(x_i \mid \theta_i) G(d\theta_i) = \int_{\mathbb{R}^d} p(x_i \mid \theta_i) \sum_{j=1}^{\infty} p_j \delta_{\hat{\theta}_j}(d\theta_i) = \sum_{j=1}^{\infty} p_j p(x_i \mid \hat{\theta}_j),$$

где мера G задаётся последовательностями $\{p_j\}_{j=1}^{\infty}$ и $\{\hat{\theta}_j\}_{j=1}^{\infty}$.

Задача байесовского вывода состоит в построении апостериорного распределения на меру G по конечной выборке $x_1, \ldots, x_n$ и фиксированному априорному распределению $\mathrm{DP}(H, \alpha)$. Обычно построить такое апостериорное распределение аналитически не удаётся. Исключением является ситуация (3), когда наблюдаются напрямую переменные θ_i, или, что эквивалентно, когда $p(x_i \mid \hat{\theta}_j)$ – дельта-функции.

В данном тексте рассматриваются две группы методов приближённого байесовского вывода: MCMC и вариационный вывод. В ситуации, когда распределения $p(x \mid \theta)$ и H принадлежат классу экспоненциальных распределений и образуют сопряжённую пару, все алгоритмы упрощаются. В рамках данного текста будет рассматриваться только такой случай. В модельном примере, рассматриваемом ниже, условие сопряжённости выполнено.

3.1.1 Модельный пример

Для демонстрации всех алгоритмов будем использовать следующий сквозной пример: $p(x \mid \theta)$ – нормальное распределение $\mathcal{N}(x \mid \mu, \sigma_x I)$, вектор параметров θ состоит из мат. ожиданий компонент μ, распределение H также является нормальным: $\mathcal{N}(\mu \mid 0, \sigma_\mu I)$. Величины σ_x и σ_μ – заранее заданные параметры, значения которых фиксированы. Примеры реализаций выборок из данной модели приведены на рис. 2.

Легко вычислить апостериорное распределение на центр кластера μ при известной выборке $x_1, \ldots, x_n$ (если известно, что все точки принадлежат к этому кластеру):

$$p(\mu \mid x_1, \ldots, x_n) = \frac{p(x_1, \ldots, x_n \mid \mu) p_H(\mu)}{\int_{\mathbb{R}^d} p(x_1, \ldots, x_n \mid \mu) p_H(\mu) d\mu} = \mathcal{N}\left(\mu \;\middle|\; \frac{\sigma_n}{\sigma_x} \sum_{i=1}^{n} x_i,\ \sigma_n I\right), \text{ где } \frac{1}{\sigma_n} = \frac{1}{\sigma_\mu} + \frac{n}{\sigma_x}.$$

Здесь $p_H(\mu)$ – плотность распределения, задаваемого мерой H.

$$
\begin{aligned}
&G \sim \mathrm{DP}(\mathcal{N}(\mu \mid 0, \sigma_\mu I),\ \alpha),\\
&\mu_i \sim G, \quad i = 1, \ldots, n,\\
&x_i \sim \mathcal{N}(x_i \mid \mu_i, \sigma_x I),\ i = 1, \ldots, n
\end{aligned}
$$

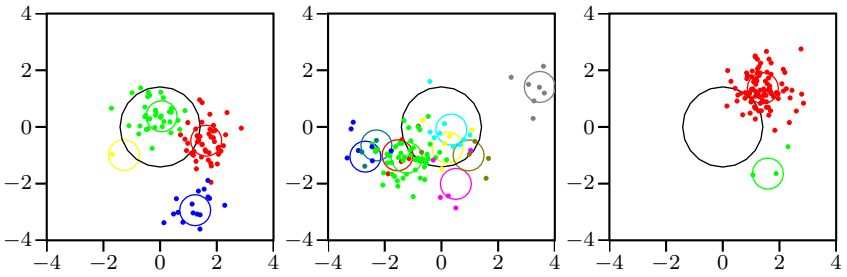


Рис. 2: Модельный пример смеси распределений с априорным распределением, заданным процессом Дирихле. Слева представлена вероятностная модель. Используются следующие значения параметров: $n = 100$, $\alpha = 1.5$, $\sigma_\mu = 2$, $\sigma_x = 0.3$. Три правых диаграммы – реализации этой модели. Точками отмечены значения переменных x_i. Цветами показаны сгенерированные кластеризации. Каждому кластеру соответствует переменная μ_k, показанная на диаграммах центрам цветных окружностей. Чёрные окружности соответствуют априорному распределению на генерацию центров кластеров H.

Маргинальное распределение на x тоже можно найти:

$$p(x) = \int_{\mathbb{R}^d} p(x \mid \mu) p_H(\mu) d\mu = \mathcal{N}(x \mid 0, (\sigma_\mu + \sigma_x) I).$$

Далее везде алгоритмы будут приводится в общем виде и для этого конкретного примера.

3.2 MCMC

Методы MCMC состоят в генерации выборки $\{\boldsymbol{\theta}_\ell\}$, $\boldsymbol{\theta}_\ell = \{\theta_i^\ell\}_{i=1}^n$ из апостериорного распределения $p(\boldsymbol{\theta} \mid \boldsymbol{x})$. Классическим текстом по применению методов MCMC к смесям распределений с процессами Дирихле является работа [12]. Ещё одно описание методов можно найти в [13, п. 2.5].

3.2.1 Простейший метод

Простейший метод генерации выборки $\{\boldsymbol{\theta}_\ell\}$ основан на схеме Гиббса, т.е. итеративной генерации элемента θ_i из условного распределения $p(\theta_i \mid \boldsymbol{\theta}_{\setminus i}, \boldsymbol{x})$, где $\boldsymbol{\theta}_{\setminus i} = (\theta_j)_{j \neq i}$.

Используя свойства условной независимости, задаваемые схемой (7), получаем, что $p(\theta_i, x_i \mid \boldsymbol{\theta}_{\setminus i}, \boldsymbol{x}_{\setminus i}) = p(\theta_i \mid \boldsymbol{\theta}_{\setminus i}) p(x_i \mid \theta_i)$. По формуле Байеса можно найти искомое маргинальное распределение:

$$p(\theta_i \mid \boldsymbol{\theta}_{\setminus i}, \boldsymbol{x}) = \frac{p(\theta_i \mid \boldsymbol{\theta}_{\setminus i}) p(x_i \mid \theta_i)}{\int p(\theta_i \mid \boldsymbol{\theta}_{\setminus i}) p(x_i \mid \theta_i) d\theta_i}.$$

Распределение $p(\theta_i \mid \boldsymbol{\theta}_{\setminus i})$ можно найти из (4), используя свойство взаимозаменяемости (exchangeability): с точки зрения процесса Дирихле порядок генерируемых точек не имеет значения.

$$p(\theta_i \mid \boldsymbol{\theta}_{\setminus i}) = \frac{\alpha}{\alpha + n - 1} p_H(\theta_i) + \frac{1}{\alpha + n - 1} \sum_{j \neq i} \delta_{\theta_j}(\theta_i). \tag{8}$$

Заметим, что использование этой формулы скрывает от нас маргинализацию по всем распределениям G (по априорному процессу Дирихле $\mathrm{DP}(H, \alpha)$).

Найдём апостериорное распределение $p(\theta_i \mid \boldsymbol{\theta}_{\setminus i}, \boldsymbol{x})$ при априорном $p(\theta_i \mid \boldsymbol{\theta}_{\setminus i})$ и правдоподобии $p(x_i \mid \theta_i)$. Априорное распределение (8) имеет вид смеси распределений, а значит и апостериорное будет иметь вид смеси, причём смеси апостериорных распределений к компонентам априорного:

$$p(\theta_i \mid \boldsymbol{\theta}_{\setminus i}, \boldsymbol{x}) = \frac{1}{Z}\left(\alpha\left(\int p(x_i \mid \theta) p_H(\theta) d\theta\right) q(\theta_i \mid x_i) + \sum_{j \neq i} \delta_{\theta_j}(\theta_i) p(x_i \mid \theta_j)\right), \tag{9}$$

где $q(\theta_i \mid x_i)$ – апостериорное распределение на θ_i, если $p_H(\theta)$ – априорное. Нормировочную константу Z можно вычислить аналитически:

$$Z = \alpha\left(\int p(x_i \mid \theta) p_H(\theta) d\theta\right) + \sum_{j \neq i} p(x_i \mid \theta_j).$$

Заметим, что если $p(x \mid \theta)$ и $p_H(\theta)$ образуют сопряжённую пару, то выражения для $p(\theta_i \mid \boldsymbol{\theta}_{\setminus i}, \boldsymbol{x})$ и Z можно получить аналитически.

В модельном примере (сек. 3.1.1) выражение (9) принимает следующий вид:

$$p(\mu_i \mid \boldsymbol{\mu}_{\setminus i}, \boldsymbol{x}) = \frac{1}{Z}\alpha\, \mathcal{N}(x_i \mid 0, (\sigma_\mu + \sigma_x)I)\, \mathcal{N}(\mu_i \mid (\sigma_1/\sigma_x)x_i,\ \sigma_1 I) + \frac{1}{Z}\sum_{j \neq i} \delta_{\mu_j}(\mu_i)\, \mathcal{N}(x_i \mid \mu_j, \sigma_x I),$$

где

$$Z = \alpha\, \mathcal{N}(x_i \mid 0, (\sigma_\mu + \sigma_x)I) + \sum_{j \neq i} \mathcal{N}(x_i \mid \mu_j, \sigma_x I).$$

Распределение $p(\mu_i \mid \boldsymbol{\mu}_{\setminus i}, \boldsymbol{x})$ представляет собой смесь нормального распределения и дельта-функций, а значит из него можно легко генерировать выборку.

Данный метод сходится к истинному апостериорному распределению очень медленно, что делает его неприменимым на практике. В частности, в рамках этого метода не предусмотрена возможность изменения значения компоненты θ_i. Возможно только добавление точек с новыми значениями (при удалении старых).

3.2.2 Схема МакИчерна

Для построения более эффективной схемы сэмплирования будем использовать представление, содержащее кластеризацию точек в явном виде [12, 11]. Пусть переменная $z_i \in \mathbb{N}$ содержит идентификатор кластера, к которому принадлежит объект x_i, а переменная $\theta_k \in \mathbb{R}^d$ содержит параметры k-й компоненты смеси:

$$\begin{aligned} (z_1, \ldots, z_n) &\sim \mathrm{CRP}(\alpha, n), \\ \theta_k &\sim H, \quad k = 1, \ldots, K, \\ x_i &\sim p(x_i \mid \theta_{z_i}), \quad i = 1, \ldots, n. \end{aligned} \tag{10}$$

Для применения схемы Гиббса сэмплирования из распределения $p(\boldsymbol{z}, \boldsymbol{\theta} \mid \boldsymbol{x})$ необходимо найти условные распределения $p(z_i \mid \boldsymbol{z}_{\setminus i}, \boldsymbol{\theta}, \boldsymbol{x})$ и $p(\theta_k \mid \boldsymbol{z}, \boldsymbol{\theta}_{\setminus k}, \boldsymbol{x})$.

По определению CRP (5):

$$p(z_i \mid \boldsymbol{z}_{\setminus i}) = \begin{cases} \frac{\sum_{j \neq i}[z_j = c]}{\alpha + n - 1}, & \text{если } z_i \text{ присоединяется к кластеру } c, \\ \frac{\alpha}{\alpha + n - 1}, & \text{если } z_i \text{ образует новый кластер.} \end{cases} \tag{11}$$

Схема (10) позволяет в явном виде записать плотность распределения на $\boldsymbol{x}$ и $\boldsymbol{\theta}$ при условии кластеризации $\boldsymbol{z}$:

$$p(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{z}) = \prod_{k=1}^{K} p_H(\theta_k) \prod_{i=1}^{n} p(x_i \mid \theta_{z_i}).$$

Отсюда можно найти условное распределение на θ_k:

$$p(\theta_k \mid \boldsymbol{\theta}_{\setminus k}, \boldsymbol{z}, \boldsymbol{x}) \propto p(\boldsymbol{x}, \boldsymbol{\theta} \mid \boldsymbol{z}) \propto p_H(\theta_k) \prod_{i:\, z_i = k} p(x_i \mid \theta_k). \tag{12}$$

Перемножение (10) и (11) позволяет записать $p(\boldsymbol{\theta}, \boldsymbol{x}, z_i \mid \boldsymbol{z}_{\setminus i})$, откуда, в свою очередь, можно найти условно распределение на z_i. Вероятность присоединения z_i к существующему кластеру можно получить сразу:

$$p(z_i = k \mid \boldsymbol{z}_{\setminus i}, \boldsymbol{\theta}, \boldsymbol{x}) = \frac{n_{\setminus i,k}}{Z_i(\alpha + n - 1)} p(x_i \mid \theta_k), \tag{13}$$

где $n_{\setminus i,k} = \sum_{j \neq i} [z_j = k]$. В случае же если z_i создаёт новый кластер, то в модели необходимо учесть ещё и новое значение θ_{z_i}, т.е.

$$p(z_i = \text{new}, \theta_{\text{new}} \mid \boldsymbol{z}_{\setminus i}, \boldsymbol{\theta}, \boldsymbol{x}) = \frac{\alpha}{Z_i(\alpha + n - 1)} p(x_i \mid \theta_{\text{new}}) p_H(\theta_{new}).$$

Интегрируя по θ_{new}, получаем

$$p(z_i = \text{new} \mid \boldsymbol{z}_{\setminus i}, \boldsymbol{\theta}, \boldsymbol{x}) = \frac{\alpha}{Z_i(\alpha + n - 1)} \int p(x_i \mid \theta_{\text{new}}) p_H(\theta_{\text{new}}) d\theta_{\text{new}}. \tag{14}$$

Нормировочную константу Z_i можно найти, просуммировав (13) и (14).

В рассматриваемом модельном примере (сек. 3.1.1) распределение (12) выглядит так:

$$p(\mu_k \mid \boldsymbol{\mu}_{\setminus k}, \boldsymbol{z}, \boldsymbol{x}) = \mathcal{N}\left(\mu_k \;\middle|\; \frac{\sigma_{n_k}}{\sigma_x} \sum_{i:\, z_i = k} x_i,\ \sigma_{n_k} I\right), \quad \text{где } n_k = \sum_{i=1}^{n} [z_i = k]. \tag{15}$$

Распределение переменной z_i при условии остальных:

$$p(z_i = k \mid \boldsymbol{z}_{\setminus i}, \boldsymbol{\mu}, \boldsymbol{x}) = \frac{n_{\setminus i,k}}{Z_i(\alpha + n - 1)} \mathcal{N}(x_i \mid \mu_k, \sigma_x I), \tag{16}$$

$$p(z_i = \text{new} \mid \boldsymbol{z}_{\setminus i}, \boldsymbol{\mu}, \boldsymbol{x}) = \frac{\alpha}{Z_i(\alpha + n - 1)} \mathcal{N}(x_i \mid 0, (\sigma_\mu + \sigma_x) I), \tag{17}$$

где Z_i – нормировочная константа.

$$Z_i = \sum_{k:\, n_{\setminus i,k} > 0} \frac{n_{\setminus i,k}}{\alpha + n - 1} \mathcal{N}(x_i \mid \mu_k, \sigma_x I) + \frac{\alpha}{\alpha + n - 1} \mathcal{N}(x_i \mid 0, (\sigma_\mu + \sigma_x) I). \tag{18}$$

Пример работы алгоритма МакИчерна на модельном примере показан на рис. 3.

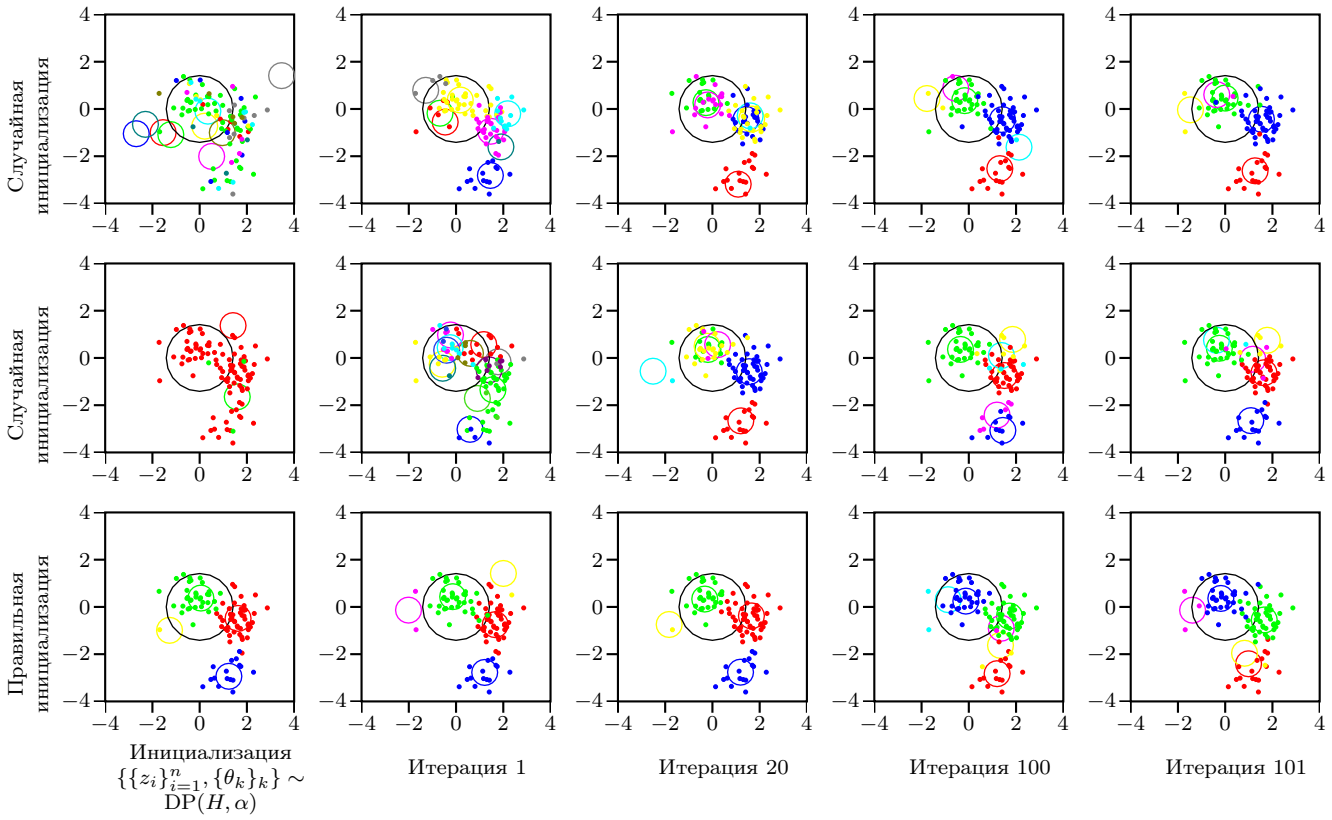


Рис. 3: Пример работы алгоритмов МСМС на модельном примере. Входом алгоритма является набор точек $\{x_i\}_{i=1}^n$ (левая картинка, рис. 2), параметры $\alpha = 1.5$, $\sigma_\mu = 2$, $\sigma_x = 0.3$. Перед запуском алгоритма проводится инициализация кластеризации $\{z_i\}_{i=1}^n$ и параметров кластеров $\{\theta_k\}_k$ (первая колонка). Первые две строки соответствуют инициализациям, сгенерированным из априорного распределения $\{\{z_i\}_{i=1}^n, \{\theta_k\}_k\} \sim \mathrm{DP}(H, \alpha)$ (центральная и правая картинки, рис. 2), нижняя строка соответствует правильной инициализации (элемент выборки, для которого генерировались данные – левая картинка, рис. 2). Далее в каждой строке представлены текущие значения переменных $\boldsymbol{z}$ и $\boldsymbol{\theta}$ на некоторых итерациях схемы МакИчерна. Можно ожидать, что в третьей строке марковская цепь изначально находится в стационарном состоянии, а в первых двух – сошлась к нему. Цветовое кодирование диаграмм аналогично рис. (2).

3.2.3 Коллапсированная схема МакИчерна

Описанную выше схему (12), (13), (14) можно сколлапсировать и избавиться от необходимости сэмплировать значения переменных θ_k. Для этого надо в (13) применить трюк аналогичный (14) и проинтегрировать по $\boldsymbol{\theta}_k$.

$$p(z_i = k, \theta_k \mid \boldsymbol{z}_{\setminus i}, \boldsymbol{x}) = p(z_i = k \mid \theta_k, \boldsymbol{z}_{\setminus i}, \boldsymbol{x}) p(\theta_k \mid \boldsymbol{z}_{\setminus i}, \boldsymbol{x}) = \frac{n_{\setminus i,k}}{Z_i(\alpha + n - 1)} p(x_i \mid \theta_k) p(\theta_k \mid \boldsymbol{x}_{\setminus i,k}),$$

где $\boldsymbol{x}_{\setminus i,k} = (x_j \mid j \neq i, z_j = k)$ и $p(\theta_k \mid \boldsymbol{x}_{\setminus i,k}) \propto p_H(\theta_k) \prod_{j:\, z_j = k,\, j \neq i} p(x_j \mid \theta_k)$. Интегрируя по θ_k, получаем условную вероятность:

$$p(z_i = k \mid \boldsymbol{z}_{\setminus i}, \boldsymbol{x}) = \frac{n_{\setminus i,k}}{Z_i(\alpha + n - 1)} \int p(x_i \mid \theta_k) p(\theta_k \mid \boldsymbol{x}_{\setminus i,k}) d\theta_k. \tag{19}$$

Нормировочную константу Z_i можно вычислить, суммируя (14) и (19).

В случае модельного примера (сек. 3.1.1) коллапсированная схема МакИчерна принимает следующий вид:

$$p(z_i = k \mid \boldsymbol{z}_{\setminus i}, \boldsymbol{x}) = \frac{n_{\setminus i,k}}{Z_i(\alpha + n - 1)} \mathcal{N}(x_i \mid \mu_{\setminus i,k}, (\sigma_x + \sigma_{n_{\setminus i,k}})I), \tag{20}$$

$$p(z_i = \text{new} \mid \boldsymbol{z}_{\setminus i}, \boldsymbol{x}) = \frac{\alpha}{Z_i(\alpha + n - 1)} \mathcal{N}(x_i \mid 0, (\sigma_\mu + \sigma_x)I), \tag{21}$$

где

$$p(\mu_k \mid \boldsymbol{x}_{\setminus i,k}) = \mathcal{N}(\mu_k \mid \mu_{\setminus i,k},\ \sigma_{n_{\setminus i,k}} I), \text{ где } \mu_{\setminus i,k} = \frac{\sigma_{n_{\setminus i,k}}}{\sigma_x} \sum_{j:\, z_j = k,\, j \neq i} x_j,\ \ \frac{1}{\sigma_{n_{\setminus i,k}}} = \frac{1}{\sigma_\mu} + \frac{n_{\setminus i,k}}{\sigma_x}, \tag{22}$$

$$Z_i = \sum_{k:\, n_{\setminus i,k} > 0} \frac{n_{\setminus i,k}}{\alpha + n - 1} \mathcal{N}(x_i \mid \mu_{\setminus i,k}, (\sigma_x + \sigma_{n_{\setminus i,k}})I) + \frac{\alpha}{\alpha + n - 1} \mathcal{N}(x_i \mid 0, (\sigma_\mu + \sigma_x)I). \tag{23}$$

Обратим внимание, что на каждой итерации схемы Гиббса требуется сэмплировать только значения переменных $\boldsymbol{z}$ согласно (20) и (21). Значения переменных $\boldsymbol{\mu}$ можно генерировать из (22) только тогда, когда они необходимы для каких-либо дальнейших вычислений.

Сравним схемы (20)-(23) и (15)-(17). С одной стороны, схема (20)-(23) требует больше вычислений, чем схема (15)-(17), но, с другой стороны, в ней реализуется схема Гиббса для распределения меньшего количества переменных, что может приводит к более быстрой сходимости.

3.3 Вариационный вывод

Вариационного подход к задаче байесовского вывода был применён к модели смеси распределений с априорным распределением в виде процесса Дирихле в работе [3].

Вариационный подход подразумевает поиск приближения к апостериорному распределению при помощи решения оптимизационной задачи по выбранному семейству приближений. При этом семейство приближений выбирается так, чтобы с ним было удобно работать (например, часто используются полностью факторизованные распределения). Пусть $\boldsymbol{x}$ – наблюдаемые переменные, $\boldsymbol{w}$ – скрытые переменные, $\boldsymbol{\varphi}$ – параметры, $p(\boldsymbol{x}, \boldsymbol{w} \mid \boldsymbol{\varphi})$ – полное правдоподобие, $q(\boldsymbol{w})$ – распределение на скрытые переменные, принадлежащее семейству приближений. Тогда

$$\log p(\boldsymbol{x} \mid \boldsymbol{\varphi}) \geq \mathbb{E}_q \log p(\boldsymbol{x}, \boldsymbol{w} \mid \boldsymbol{\varphi}) - \mathbb{E}_q \log q(\boldsymbol{w}) = \mathcal{L}(q), \tag{24}$$

где $\mathcal{L}(q)$ – вариационная нижняя оценка на неполное правдоподобие. Вариационный подход состоит в максимизации $\mathcal{L}(q)$ по q из выбранного семейства приближений.

Для применения вариационного вывода к модели (7) будем использовать определение процесса Дирихле через "Ломку палки" (6). Модели смеси можно записать в следующем виде:

$$\begin{aligned}
&v_i \sim \text{Beta}(1, \alpha), && i = 1, \dots, \infty \\
&p_i = v_i \prod_{j=1}^{i-1}(1 - v_j), && i = 1, \dots, \infty \\
&\theta_i \sim H, && i = 1, \dots, \infty \\
&z_j \sim \text{Discrete}(p_1, \dots, p_i, \dots), && j = 1, \dots, n, \\
&x_j \sim p(x_j \mid \theta_{z_j}), && j = 1, \dots, n.
\end{aligned} \tag{25}$$

В этой модели наблюдаемые переменные – $\boldsymbol{x}$, скрытые – $\boldsymbol{w} = (\boldsymbol{v}, \boldsymbol{\theta}, \boldsymbol{z})$, параметры – $\boldsymbol{\varphi} = (\alpha, H)$. Логарифм полного правдоподобия выглядит так:

$$\log p(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{\theta}, \boldsymbol{z} \mid \alpha, H) = \sum_{i=1}^{\infty} \log p(v_i \mid \alpha) + \sum_{i=1}^{\infty} \log p_H(\theta_i) + \sum_{j=1}^{n} \left(\log p(z_j \mid \boldsymbol{v}) + \log p(x_j \mid \theta_{z_j}) \right). \tag{26}$$

В качестве семейства приближений q будем использовать полностью факторизованное семейство распределений, усечённое по некоторому параметру T:

$$q(\boldsymbol{v}, \boldsymbol{\theta}, \boldsymbol{z}) = \prod_{t=1}^{T-1} q_t^v(v_t) \prod_{t=1}^{T} q_t^\theta(\theta_t) \prod_{i=1}^{n} q_i^z(z_i) \quad \cdot \quad \prod_{t=T}^{\infty} \delta_1(v_t) \prod_{t=T+1}^{\infty} \delta_0(\theta_t). \tag{27}$$

Усечение по параметру T означает, что все переменные v_t и θ_t с большими индексами реально не участвуют в модели. Заметим, что истинное апостериорное распределение модели (25) зависит от бесконечного числа переменных, а приближение q – существенно зависит лишь от конечного.

Дальнейшее применение вариационного подхода состоит в максимизации $\mathcal{L}(q)$ (24) по семейству (27). Для решения задачи оптимизации методом покоординатного подъёма необходимо получить формулы пересчёта для $q_t^v(v_t)$, $q_t^\theta(\theta_t)$, $q_i^z(z_i)$. Для этого воспользуемся основным результатом вариационного вывода [2, ур. 10.9]:

$$q_i(w_i) \propto \exp(\mathbb{E}_{\prod_{j \neq i} q_j(w_j)} \log p(\boldsymbol{x}, \boldsymbol{w}))$$

Пересчёт $q_t^\theta(\theta_t)$. Запишем мат. ожидание (26) по всем переменным, кроме θ_t:

$$\mathbb{E}_{q_{\setminus \theta_t}} \log p(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{\theta}, \boldsymbol{z} \mid \alpha, H) = \log p_H(\theta_t) + \sum_{i=1}^{n} \log p(x_i \mid \theta_t)\, \mathbb{E}_{q_i^z}[z_i = t] + \text{const},$$

что означает, что если $\gamma_{it} = \mathbb{E}_{q_i^z}[z_i = t]$ известно, то

$$q_t^\theta(\theta_t) \propto p_H(\theta_t) \prod_{i=1}^{n} (p(x_i \mid \theta_t))^{\gamma_{it}}, \tag{28}$$

что можно вычислить, если $p(x_i \mid \theta_t)$ и $p_H(\theta_t)$ образуют сопряжённую пару. При этом $q_t^\theta(\theta_t)$ будет принадлежать тому же семейству, что и $p_H(\theta_t)$.

Пересчёт $q_t^v(v_t)$. Запишем мат. ожидание (26) по всем переменным, кроме v_t:

$$\mathbb{E}_{q_{\setminus v_t}} \log p(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{\theta}, \boldsymbol{z} \mid \alpha, H) = \\ (\alpha - 1)\log(1 - v_t) + \sum_{j=1}^{n} \Big(\log v_t \, \mathbb{E}_{q_j^z}[z_j = t] + \log(1 - v_t)\, \mathbb{E}_{q_j^z}[z_j - 1 \geq t] \Big) + \text{const}\,. \quad (29)$$

При известных $\gamma_{jt} = \mathbb{E}_{q_j^z}[z_j = t]$ можно найти $q_t^v(v_t)$:

$$q_t^v(v_t) = \text{Beta}\left(v_t \;\middle|\; 1 + \sum_{j=1}^{n} \gamma_{jt},\; \alpha + \sum_{j=1}^{n} \sum_{s=t+1}^{T} \gamma_{js} \right). \quad (30)$$

Заметим, что из усечённости предположения факторизации (27), следует, что $q_j^z(z_j = t) = 0$ при $t > T$. Отсюда следует, что верхний предел суммирования по s в формуле (30) равен T.

Пересчёт $q_j^z(z_j)$. Запишем мат. ожидание (26) по всем переменным, кроме z_j:

$$\mathbb{E}_{q_{\setminus z_j}} \log p(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{\theta}, \boldsymbol{z} \mid \alpha, H) = \mathbb{E}_{q_{z_j}^v} \log v_{z_j} + \sum_{i=1}^{z_j - 1} \mathbb{E}_{q_i^v} \log(1 - v_i) + \mathbb{E}_{q_{z_j}^\theta} \log p(x_j \mid \theta_{z_j}) + \text{const}\,.$$

Если известны $\mathbb{E}_{q_k^v} \log v_k$, $\mathbb{E}_{q_k^v} \log(1 - v_k)$, $\mathbb{E}_{q_k^\theta} \log p(x_j \mid \theta_k)$, то дискретное распределение $q_j^z(z_j)$ можно найти путём нормировки выражения выше:

$$q_j^z(z_j) \propto \exp\Big(\mathbb{E}_{q_{\setminus z_j}} \log p(\boldsymbol{x}, \boldsymbol{v}, \boldsymbol{\theta}, \boldsymbol{z} \mid \alpha, H) \Big)\,. \quad (31)$$

Поскольку $q_k^v \sim \text{Beta}(a, b)$, то $\mathbb{E}_{q_k^v} \log v_k = \psi(a) - \psi(a + b)$ и $\mathbb{E}_{q_k^v} \log(1 - v_k) = \psi(b) - \psi(a + b)$, где $\psi(\cdot)$ – дигамма функция. Мат. ожидание $\mathbb{E}_{q_k^\theta} \log p(x_j \mid \theta_k) = \int \log p(x_j \mid \theta) q_k^\theta(\theta) d\theta$ также обычно можно вычислить, если q_k^θ принадлежит тому же семейству, что и H, а $p(x_j \mid \theta)$ и H образуют сопряжённую пару.

Заметим, что $\gamma_{jt} = \mathbb{E}_{q_j^z}[z_j = t] = q_j^z(t)$, поскольку q_j^z – дискретное распределение.

Модельный пример. Для применения схемы вариационного вывода на модельном примере (сек. 3.1.1) необходимо получить формулу пересчёта $q_t^\theta(\mu_t)$ (28) и вычислить интеграл $\int p(x_j \mid \mu) q_k^\theta(\mu) d\mu$ для пересчёта (31).

По (28) нетрудно видеть, что $q_t^\theta(\mu_t)$ будет нормальным распределением:

$$q_t^\theta(\mu_t) = \mathcal{N}(\mu_t \mid \hat{\mu},\; \sigma I), \text{ где } \frac{1}{\sigma} = \frac{1}{\sigma_\mu} + \frac{1}{\sigma_x} \sum_{i=1}^{n} \gamma_{it},\;\; \hat{\mu} = \frac{\sigma}{\sigma_x} \sum_{i=1}^{n} \gamma_{it} x_i.$$

В этом случаем можно вычислить и интеграл в (31):

$$\int \log p(x_j \mid \mu) q_k^\theta(\mu) d\mu = -\frac{n}{2} \log(2\pi\sigma_x) - \frac{1}{2\sigma_x}(x_j^\intercal x_j - 2\hat{\mu}^\intercal x_j + n\sigma + \hat{\mu}^\intercal \hat{\mu}).$$

Пример работы полученного алгоритма представлен на рис. 4.

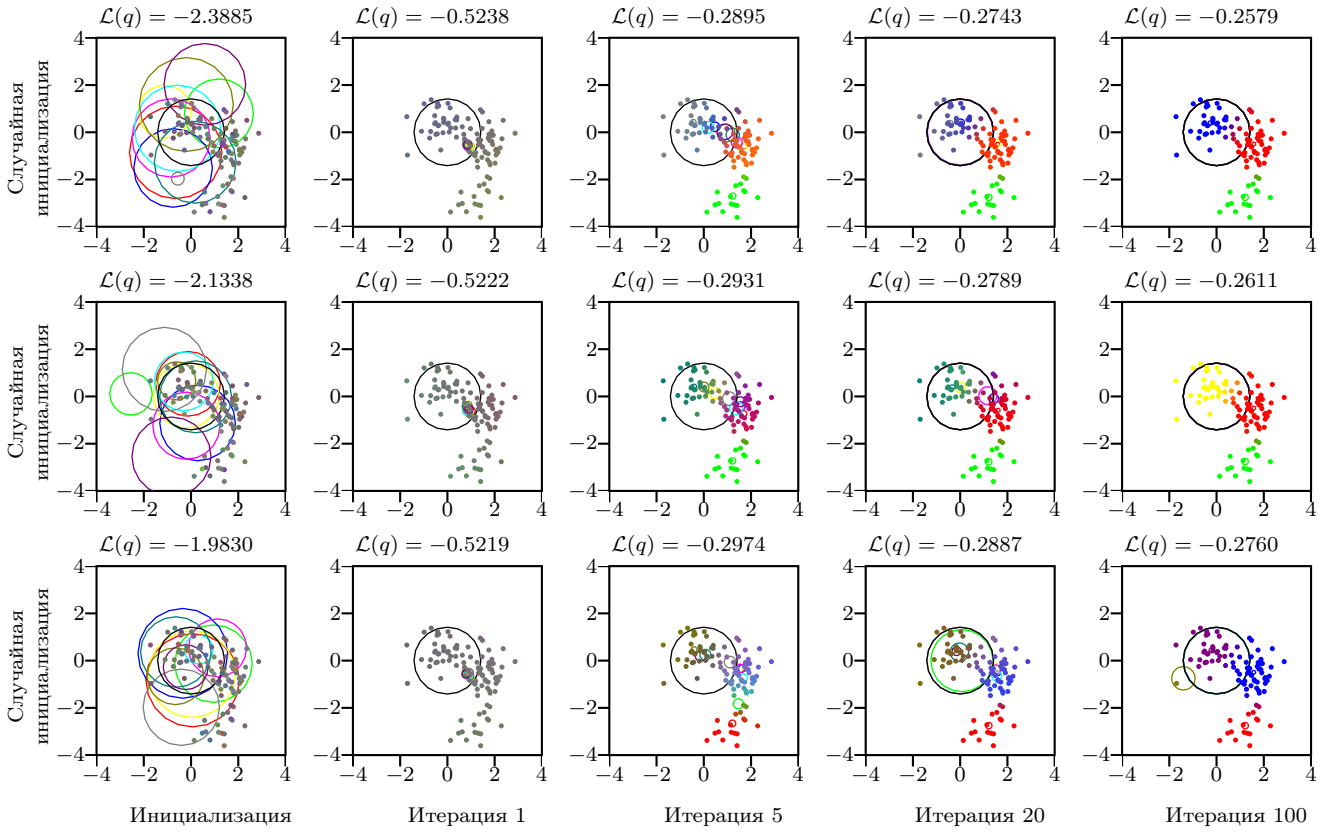


Рис. 4: Примеры работы алгоритма вариационного вывода на модельном примере. Входом алгоритма является набор точек $\{x_i\}_{i=1}^n$ (левая картинка, рис. 2), параметры $\alpha = 1.5$, $\sigma_\mu = 2$, $\sigma_x = 0.3$. Перед запуском алгоритма проводится случайная инициализация $q_t^v(v_t)$, $q_t^\theta(\theta_t)$, $q_i^z(z_i)$ (первая колонка). Далее в каждой строке представлены текущие значения переменных $q_t^\theta(\theta_t)$ и $q_i^z(z_i)$ на некоторых итерациях алгоритма. Цвет каждой точки представляет собой смесь цветов компонент с весами $q_i^z(z_i)$. Цветные окружности показывают распределения центров компонент $q_t^\theta(\theta_t)$. Для каждой диаграммы приведено значение нижней оценки $\mathcal{L}(q)$.

Рекомендации по отладке. Для верификации правильности программы, реализующей вариационный подход, необходимо смотреть на значение нижней оценки $\mathcal{L}(q)$ (24). После каждого пересчёта компонент приближения q значение $\mathcal{L}(q)$ должно строго увеличиваться (рекомендуемая точность – 10^{-10}). Для поиска ошибок полезно смотреть на нижнюю оценку по слагаемым и отслеживать суммы только тех слагаемых, которые изменяются при данном пересчёте.

4 Обобщения процесса Дирихле

Процесс Питмана-Йора (Pitman-Yor process). Как уже упоминалось выше, одним из свойств процесса Дирихле является то, то количество кластеров растёт логарифмически с ростом числа наблюдений. С точки зрения ряда приложений это свойство не является привлекательным, поскольку многие естественные статистики изменяются по-другому. Часто возникает так называемый степенной закон (power law). Примерами статистик, распределённых согласно степенному закону, являются частоты слов в английском языке, размеры городов, количество фолловеров в Twitter и т. д. Процесс Питмана-Йора является модификацией процесса Дирихле, при которой количество кластеров растёт согласно степенному закону: $O(n^d)$, где $d \in [0, 1)$ – дополнительный параметр. Обозначается процесс Питмана-Йора через $\mathrm{PY}(H, \alpha, d)$. Аналогично процессу Дирихле существует как минимум два представления процесса Питмана-Йора: "Китайский ресторан" и "Ломка палки".

Представление "Китайский ресторан" аналогично (5):

$$\theta_{n+1} \mid \theta_1, \ldots, \theta_n \sim \begin{cases} H, & \text{с вероятностью } \frac{\alpha + dm(\theta_1, \ldots, \theta_n)}{\alpha + n}, \\ \delta_c, & \text{с вероятностью } \frac{\sum_{j=1}^{n} [\theta_j = c] - d}{\alpha + n}. \end{cases}$$

Здесь $m(\theta_1, \ldots, \theta_n)$ – количество кластеров в реализации случайных величин $\theta_1, \ldots, \theta_n$.

Представление "Ломка палки" аналогично (6):

$$\begin{aligned} v_i &\sim \mathrm{Beta}(1 - d,\ \alpha + id), \quad i = 1, \ldots, +\infty, \\ p_i &= v_i \prod_{j=1}^{i-1} (1 - v_j), \quad i = 1, \ldots, +\infty, \\ \theta_i &\sim H, \quad i = 1, \ldots, +\infty. \end{aligned}$$

Заметим, что при $d = 0$ процесс Питмана-Йора переходит в процесс Дирихле. Подробнее про процесс Питмана-Йора можно прочитать в [13, п. 2.8].

Иерархический процесс Дирихле (hierarchical Dirichlet process, HDP) [16]. Существуют ситуации, когда, с одной стороны, все данные разделены на группы, и необходимо решить задачу кластеризации в каждой из групп отдельно. С другой же стороны крайне желательно, чтобы параметры кластеров в разных группах совпадали. Примером такой ситуации является задача тематического моделирования (topic models). Здесь отдельные объекты – это слова. Группы объектов – это документы. Кластерами же являются темы. Слова по темам в каждом документе требуется разделять независимо, но должно устанавливаться соответствие между темами в разных документах.

Для решения этой задачи разработана модель иерархического процесса Дирихле . В рамках данной модели для каждой группы j мера G_j генерируется из процесса Дирихле $\mathrm{DP}(G_0, \alpha_0)$, у

которого базовая мера является G_0 является дискретной. При этом все меры G_j также являются дискретными, позиции их атомов выбираются из последовательности позиций атомов G_0. Дискретность базовой меры G_0 обеспечивается при помощи наложения на неё априорного распределения также в виде процесса Дирихле. Математически модель записывается так:

$$G_0 \mid H, \gamma \sim \mathrm{DP}(H, \gamma),$$
$$G_j \mid G_0, \alpha_0 \sim \mathrm{DP}(G_0, \alpha_0), \quad \text{для каждой группы } j.$$

Для модели HDP разработаны как алгоритмы MCMC, так и алгоритмы вариационного вывода. Алгоритмы MCMC предложены в оригинальной статье [16], и с ними авторами ассоциирован термин "Франшиза китайского ресторана" (Chinese restaurant franchise). Алгоритмы вариационного выводы предложены позже, например, в работах [17, 18].

Вложенные процессы Дирихле (nested Chinese restaurant process, nCRP) [5] являются распределениями над иерархическими кластеризациями. nCRP используются в задаче иерархического тематического моделирования (темы образуют иерархию).

Китайский ресторан с расстояниями (distant-dependent CRP, ddCRP) [4] – непараметрическая модель, в которой происходит отказ от взаимозаменяемости (exchangeability) объектов. В модель вводится внешнее понятие расстояния между объектами, которое используется для кластеризации. Расстояние, например, может зависеть от номера объекта, если данные последовательны, или от координат объекта (пикселя) на изображении.

Скрытая марковская модель с бесконечным числом состояний (infinite HMM) [1] – обобщение скрытой марковской модели на случай бесконечного числа состояний.

Процесс "Индийский буфет" (Indian buffet process, IBP) [9] представляет собой распределение над разреженными бинарными матрицами с фиксированным количеством строк и неограниченным количеством столбцов. Фактически данные структуры соответствуют кластеризациям, в которых объектам разрешается принадлежать сразу нескольким кластерам. С понятием IBP тесно связано понятие бета-процесса. IBP и бета-процессы соотносятся также, как CRP и процессы Дирихле.

5 Приложения процессов Дирихле и их расширений

Самым простым приложением процессов Дирихле является задача кластеризации объектов, в которой неизвестно число кластеров. Обратите внимание, что если известно, что число кластеров конечно и не растёт с увеличением объёма данных, то непараметрические модели, предполагающие бесконечное количество кластеров применять не следует. В частности, получаемые оценки на количество кластеров будут завышены.

Одним из наиболее популярных приложений процессов Дирихле и их расширений является тематическое моделирование. Действительно, число тем (топиков) обычно заранее неизвестно, детализация тем, вообще говоря, может расти с ростом объёма данных. Соответственно, непараметрические, модели хорошо подходят для этих приложений. Модель HDP [16], например, позволяет построить непараметрическое обобщение популярной модели LDA [6].

Модели, связанные с процессами Дирихле также используются и в компьютерном зрении: сегментация изображений [14, 8], вычитание фона [10].

Приложения IBP [9] включают в себя восстановление структуры белка, факторизацию матриц, выделение признаков по матрице схожести, и др.

Список литературы

[1] M. J. Beal, Z. Ghahramani, and C. E. Rasmussen, "The infinite hidden Markov model," in *Neural Information Processing Systems (NIPS)*, 2002.

[2] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[3] D. Blei and M. Jordan, "Variational inference for Dirichlet process mixtures," *Journal of Bayesian Analysis*, vol. 1, no. 1, pp. 121–144, 2006.

[4] D. M. Blei and P. I. Frazier, "Distance dependent Chinese restaurant processes," *Journal of Machine Learning Research (JMLR)*, vol. 12, pp. 2383–2410, 2011.

[5] D. M. Blei, T. L. Griffiths, M. I. Jordan, and J. B. Tenenbaum, "Hierarchical topic models and the nested Chinese restaurant process," in *Neural Information Processing Systems (NIPS)*, 2003.

[6] D. M. Blei, A. Y. Ng, and M. I. Jordan, "Latent Dirichlet allocation," *Journal of Machine Learning Research (JMLR)*, vol. 3, pp. 993–1022, 2003.

[7] B. A. Frigyik, A. Kapila, and M. R. Gupta, "Introduction to the Dirichlet distribution and related processes," UWEE, Tech. Rep. UWEETR-2010-0006, 2010.

[8] S. Ghosh, A. Ungureanu, E. Sudderth, and D. Blei, "Spatial distance dependent Chinese restaurant processes for image segmentation," in *Neural Information Processing Systems (NIPS)*, 2011.

[9] T. L. Griffiths and Z. Ghahramani, "The Indian buffet process: An introduction and review," *Journal of Machine Learning Research (JMLR)*, vol. 12, pp. 1185–1224, 2011.

[10] T. S. F. Haines and T. Xiang, "Background subtraction with Dirichlet processes," in *European conference on computer vision (ECCV)*, 2012.

[11] S. N. MacEachern, "Estimating normal means with a conjugate style Dirichlet process prior," *Communications in Statistics - Simulation and Computation*, vol. 23, no. 3, pp. 727–741, 1994.

[12] R. Neal, "Markov chain sampling methods for Dirichlet process mixture models," *Journal of Computational and Graphical Statistics*, vol. 9, pp. 249–265, 2000.

[13] P. Orbanz, *Lecture Notes on Bayesian Nonparametrics*. http://stat.columbia.edu/~porbanz/reports/porbanz_BNP_draft.pdf.

[14] P. Orbanz and J. M. Buhmann, "Nonparametric Bayesian image segmentation," *International Journal of Computer Vision (IJCV)*, vol. 77, pp. 25–45, 2008.

[15] Y. Teh, "Dirichlet processes," in *Encyclopedia of Machine Learning*. Springer, 2010.

[16] Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical Dirichlet processes," *Journal of the American Statistical Association (JASA)*, vol. 101, no. 476, pp. 1566–1581, 2006.

[17] Y. W. Teh, K. Kurihara, and M. Welling, "Collapsed variational inference for HDP," in *Neural Information Processing Systems (NIPS)*, 2007.

[18] C. Wang, J. Paisley, and D. Blei, "Online variational inference for the hierarchical Dirichlet process," in *Artificial Intelligence and Statistics (AISTATS)*, 2011.