Разделение смеси распределений. EM-алгоритм для классификации и кластеризации

Воронцов Константин Вячеславович
vokov@forecsys.ru
http://www.MachineLearning.ru/wiki?title=User:Vokov

Этот курс доступен на странице вики-ресурса
http://www.MachineLearning.ru/wiki
«Машинное обучение (курс лекций, К.В.Воронцов)»

Видеолекции: http://shad.yandex.ru/lectures

27 апреля 2017

Содержание

Напоминание. Байесовская теория классификации

X — объекты, Y — ответы, $X \times Y$ — в.п. с плотностью $p(x, y)$;

Две подзадачи:

1. **Дано:**
   $X^\ell = (x_i, y_i)_{i=1}^\ell$ — обучающая выборка.
   Найти:
   эмпирические оценки $\hat{P}(y)$ и $\hat{p}(x|y)$, $y \in Y$
   (восстановить плотность каждого класса по выборке).
2. **Дано:**
   априорные вероятности $P(y)$ и плотности $p(x|y)$, $y \in Y$.
   Найти:
   классификатор $a\colon X \times Y$, минимизирующий риск $R(a)$.
   Решение:
   $a(x) = \arg\max\limits_{y \in Y} \lambda_y P(y) p(x|y)$.

Задача восстановления смеси распределений

Порождающая модель смеси распределений:

$$p(x) = \sum_{j=1}^{k} w_j \varphi(x, \theta_j), \qquad \sum_{j=1}^{k} w_j = 1, \qquad w_j \geqslant 0,$$

k — число компонент смеси;
$\varphi(x, \theta_j) = p(x|j)$ — *функция правдоподобия* j-й компоненты;
$w_j = p(j)$ — *априорная вероятность* j-й компоненты.

Задача 1: при фиксированном k,
имея простую выборку $X^m = \{x_1, \ldots, x^m\} \sim p(x)$,
оценить вектор параметров $(w, \theta) = (w_1, \ldots, w_k, \theta_1, \ldots, \theta_k)$.

Задача 2: оценить ещё и k.

Максимизация правдоподобия и EM-алгоритм

Задача максимизации логарифма правдоподобия

$$L(w,\theta) = \ln \prod_{i=1}^{m} p(x_i) = \sum_{i=1}^{m} \ln \sum_{j=1}^{k} w_j \varphi(x_i, \theta_j) \to \max_{w,\theta}.$$

при ограничениях $\sum_{j=1}^{k} w_j = 1;\ \ w_j \geqslant 0.$

Итерационный алгоритм Expectation–Maximization:

1: начальное приближение параметров (w, θ);
2: **повторять**
3: оценка скрытых переменных $G = (g_{ij}),\ \ g_{ij} = p(j|x_i)$:
$G := \text{E-шаг}\,(w, \theta)$;
4: максимизация правдоподобия, отдельно по компонентам:
$(w, \theta) := \text{M-шаг}\,(w, \theta, G)$;
5: **пока** w, θ и G не стабилизируются.

EM-алгоритм как способ решения системы уравнений

Теорема (необходимые условия экстремума)

Точка $(w_j, \theta_j)_{j=1}^k$ локального экстремума $L(w, \theta)$ удовлетворяет системе уравнений относительно w_j, θ_j и g_{ij}:

$$\text{E-шаг:} \quad g_{ij} = \frac{w_j \varphi(x_i, \theta_j)}{\sum_{s=1}^k w_s \varphi(x_i, \theta_s)}, \quad i = 1, \ldots, m, \quad j = 1, \ldots, k;$$

$$\text{M-шаг:} \quad \theta_j = \arg\max_\theta \sum_{i=1}^m g_{ij} \ln \varphi(x_i, \theta), \quad j = 1, \ldots, k;$$

$$w_j = \frac{1}{m} \sum_{i=1}^m g_{ij}, \quad j = 1, \ldots, k.$$

EM-алгоритм — это метод простых итераций для её решения

Вероятностная интерпретация

E-шаг — это формула Байеса:

$$g_{ij} = \mathrm{P}(j|x_i) = \frac{P(j)p(x_i|j)}{p(x_i)} = \frac{w_j\varphi(x_i,\theta_j)}{p(x_i)} = \frac{w_j\varphi(x_i,\theta_j)}{\sum_{s=1}^{k} w_s\varphi(x_i,\theta_s)}.$$

Очевидно, выполнено условие нормировки: $\sum_{j=1}^{k} g_{ij} = 1$.

M-шаг — это максимизация взвешенного правдоподобия, с весами объектов g_{ij} для j-й компоненты смеси:

$$\theta_j = \arg\max_{\theta} \sum_{i=1}^{m} g_{ij} \ln \varphi(x_i,\theta),$$

$$w_j = \frac{1}{m}\sum_{i=1}^{m} g_{ij}.$$

Доказательство. Условия Каруша–Куна–Таккера

Лагранжиан оптимизационной задачи «$L(w,\theta) \to \max$»:

$$\mathscr{L}(w,\theta) = \sum_{i=1}^{m} \ln \underbrace{\left(\sum_{j=1}^{k} w_j \varphi(x_i,\theta_j) \right)}_{p(x_i)} - \lambda \left(\sum_{j=1}^{k} w_j - 1 \right).$$

Приравниваем нулю производные:

$$\frac{\partial \mathscr{L}}{\partial w_j} = 0 \quad \Rightarrow \quad \lambda = m; \quad w_j = \frac{1}{m} \sum_{i=1}^{m} \underbrace{\frac{w_j \varphi(x_i,\theta_j)}{p(x_i)}}_{g_{ij}} = \frac{1}{m} \sum_{i=1}^{m} g_{ij},$$

$$\frac{\partial \mathscr{L}}{\partial \theta_j} = \sum_{i=1}^{m} \underbrace{\frac{w_j \varphi(x_i,\theta_j)}{p(x_i)}}_{g_{ij}} \frac{\partial}{\partial \theta_j} \ln \varphi(x_i,\theta_j) = \frac{\partial}{\partial \theta_j} \sum_{i=1}^{m} g_{ij} \ln \varphi(x_i,\theta_j) = 0.$$

■

EM-алгоритм

Вход: $X^m = \{x_1, \ldots, x_m\}$, k, δ, начальные $(w_j, \theta_j)_{j=1}^k$;

Выход: $(w_j, \theta_j)_{j=1}^k$ — параметры смеси распределений

1: **повторять**

2: E-шаг (expectation):

для всех $i = 1, \ldots, m$, $j = 1, \ldots, k$

$$g_{ij}^0 := g_{ij}; \qquad g_{ij} := \frac{w_j \varphi(x_i, \theta_j)}{\sum_{s=1}^k w_s \varphi(x_i, \theta_s)};$$

3: M-шаг (maximization):

для всех $j = 1, \ldots, k$

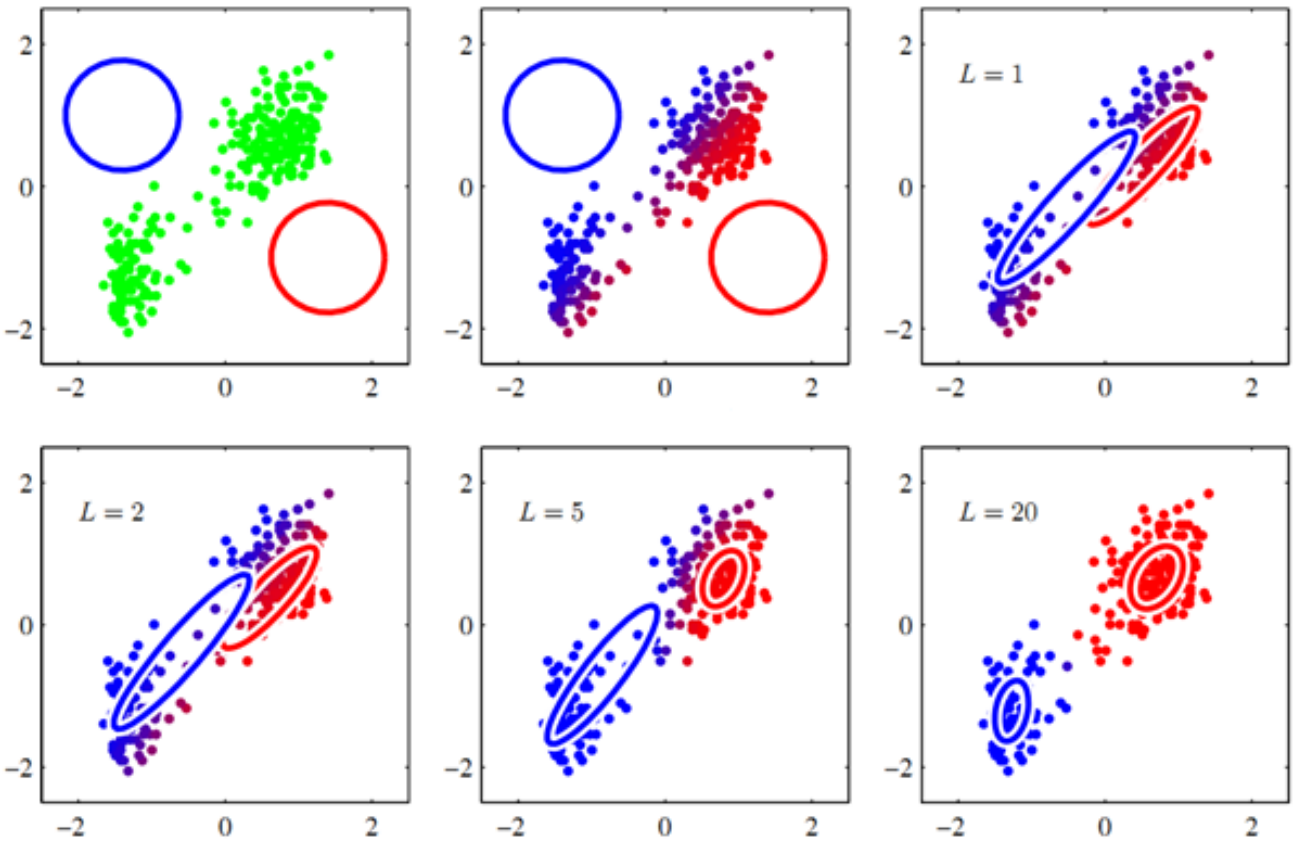
$$\theta_j := \arg\max_{\theta} \sum_{i=1}^m g_{ij} \ln \varphi(x_i, \theta); \qquad w_j := \frac{1}{m} \sum_{i=1}^m g_{ij};$$

4: **пока** $\max\limits_{i,j} |g_{ij} - g_{ij}^0| > \delta$;

5: **вернуть** $(w_j, \theta_j)_{j=1}^k$;

Пример

Две гауссовские компоненты $k = 2$ в пространстве $X = \mathbb{R}^2$.
Расположение компонент в зависимости от номера итерации L:



EM-алгоритм с последовательным добавлением компонент

Проблемы базового варианта EM-алгоритма:

- Как выбирать начальное приближение?
- Как определять число компонент?
- Как ускорить сходимость?

Вход:

выборка $X^m = \{x_1, \ldots, x_m\}$;

R — допустимый разброс правдоподобия объектов;

m_0 — минимальная длина выборки, по которой можно восстанавливать плотность;

δ — параметр критерия останова;

Выход:

k — число компонент смеси;

$(w_j, \theta_j)_{j=1}^k$ — веса и параметры компонент;

EM-алгоритм с последовательным добавлением компонент

1: начальное приближение — одна компонента:
$\theta_1 := \arg\max_\theta \sum_{i=1}^{m} \ln \varphi(x_i, \theta); \quad w_1 := 1; \quad k := 1;$
2: **для всех** $k := 2, 3, \ldots$
3: выделить объекты с низким правдоподобием:
$U := \left\{ x_i \in X^m \;\middle|\; p(x_i) < \frac{1}{R} \max_j p(x_j) \right\};$
4: **если** $|U| < m_0$ **то**
5: **выход** из цикла по k;
6: начальное приближение для k-й компоненты:
$\theta_k := \arg\max_\theta \sum_{x_i \in U} \ln \varphi(x_i, \theta); \quad w_k := \frac{1}{m}|U|;$
$w_j := w_j(1 - w_k), \;\; j = 1, \ldots, k-1;$
7: выполнить EM $(X^m, k, \delta, w, \theta)$;

Регуляризация с априорным распределением Дирихле

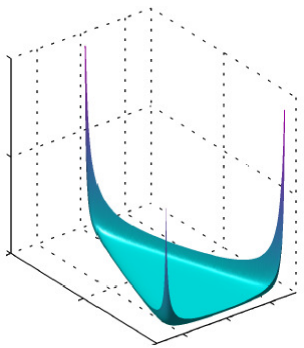
Гипотеза. Вектор весов $w = (w_j)_{j=1}^k$ порождается распределением Дирихле с вектором параметров $\alpha \in \mathbb{R}^k$:

$$\mathrm{Dir}(w|\alpha) = \frac{\Gamma(\alpha_0)}{\prod_j \Gamma(\alpha_j)} \prod_j w_j^{\alpha_j - 1}, \quad \alpha_0 = \sum_{j=1}^k \alpha_j, \quad \alpha_j \geqslant 0;$$

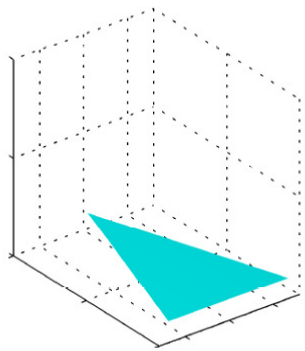
Распределение Дирихле порождает нормированные неотрицательные векторы заданной размерности k

Пример:

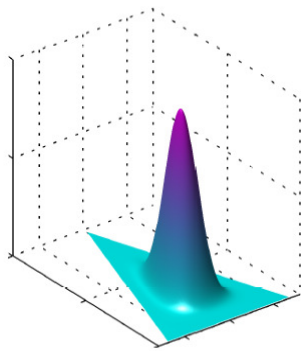
$\mathrm{Dir}(w|\alpha)$
$k = 3$
$w, \alpha \in \mathbb{R}^3$



$\alpha_1 = \alpha_2 = \alpha_3 = 0.1$

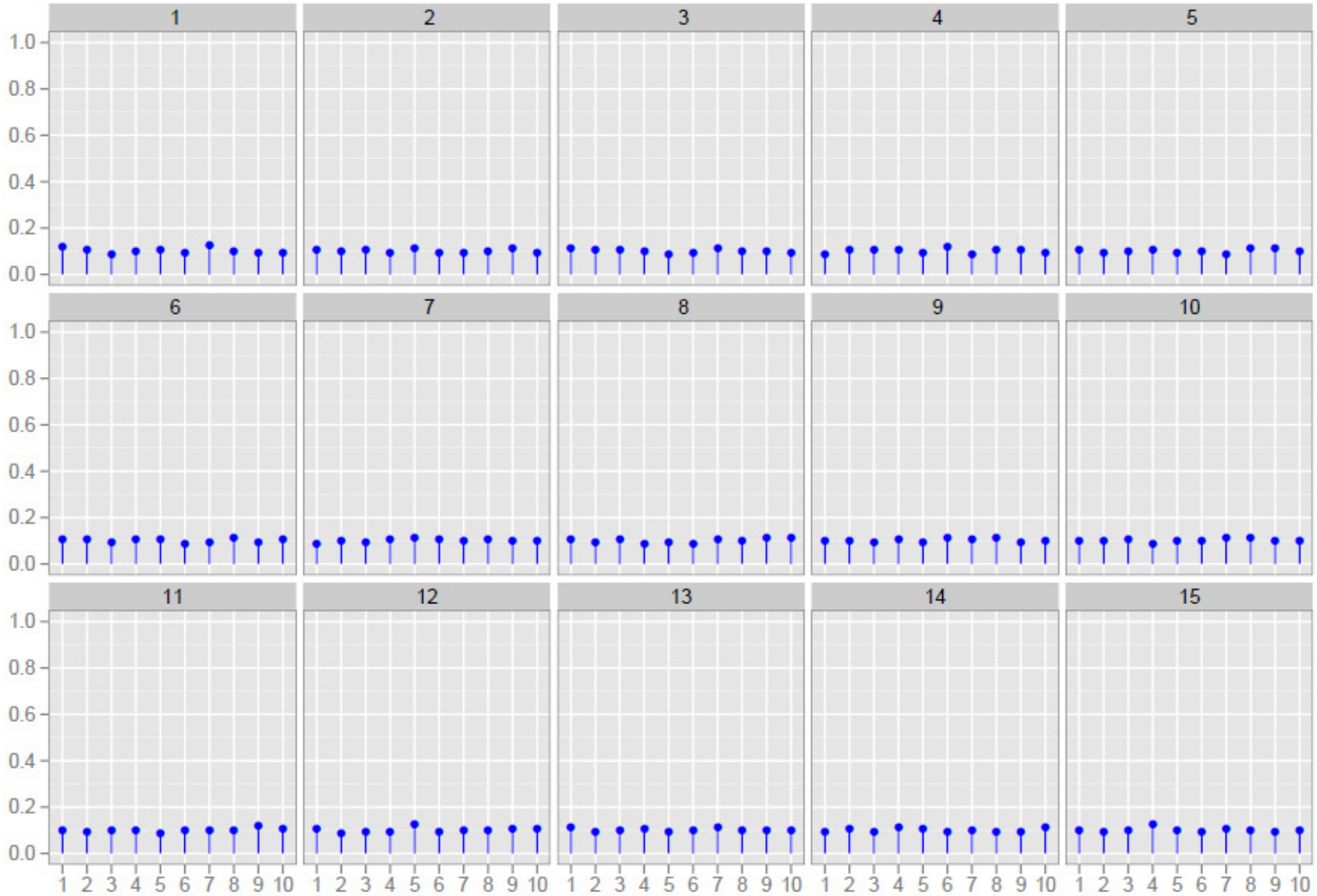


$\alpha_1 = \alpha_2 = \alpha_3 = 1$

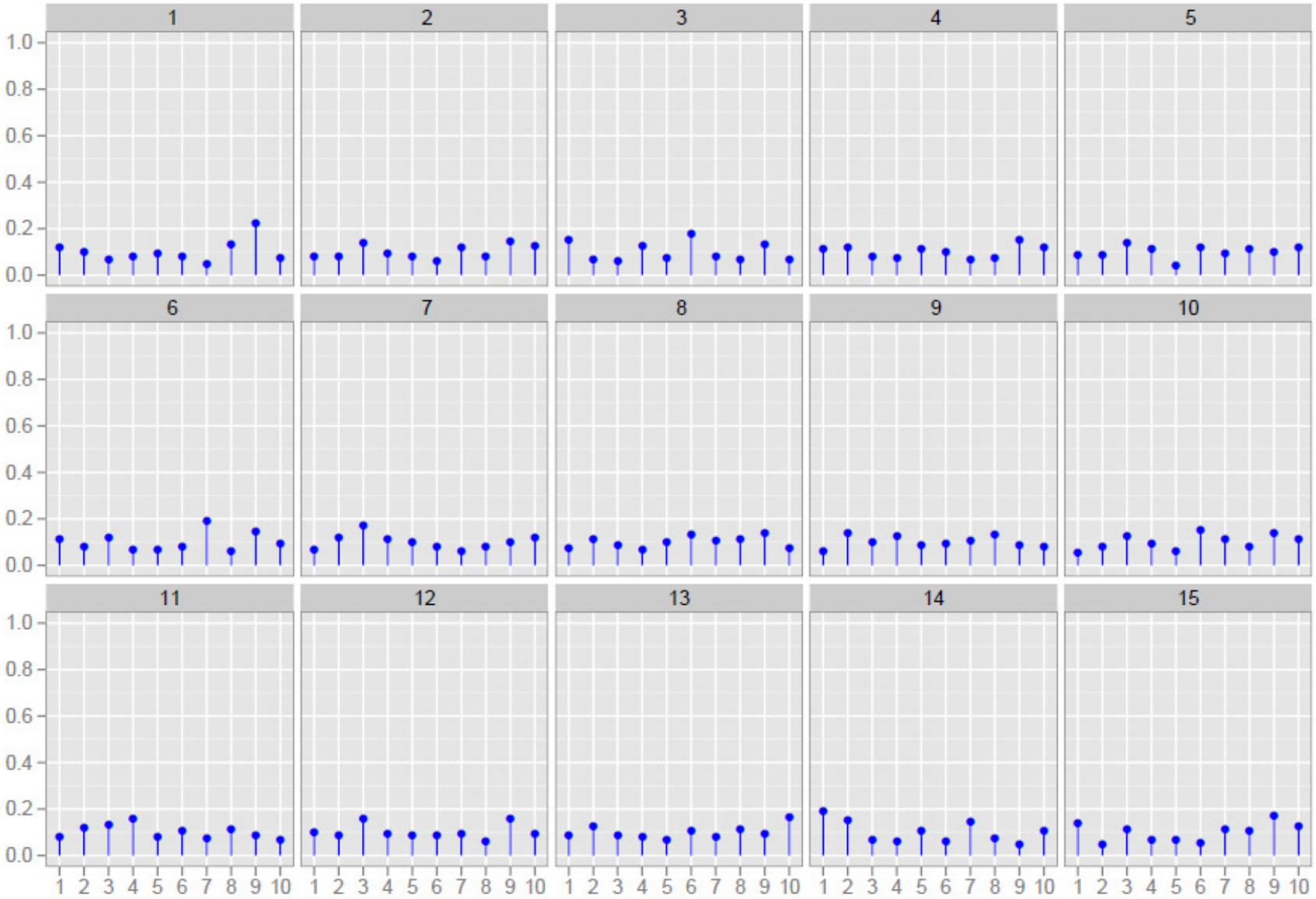


$\alpha_1 = \alpha_2 = \alpha_3 = 10$

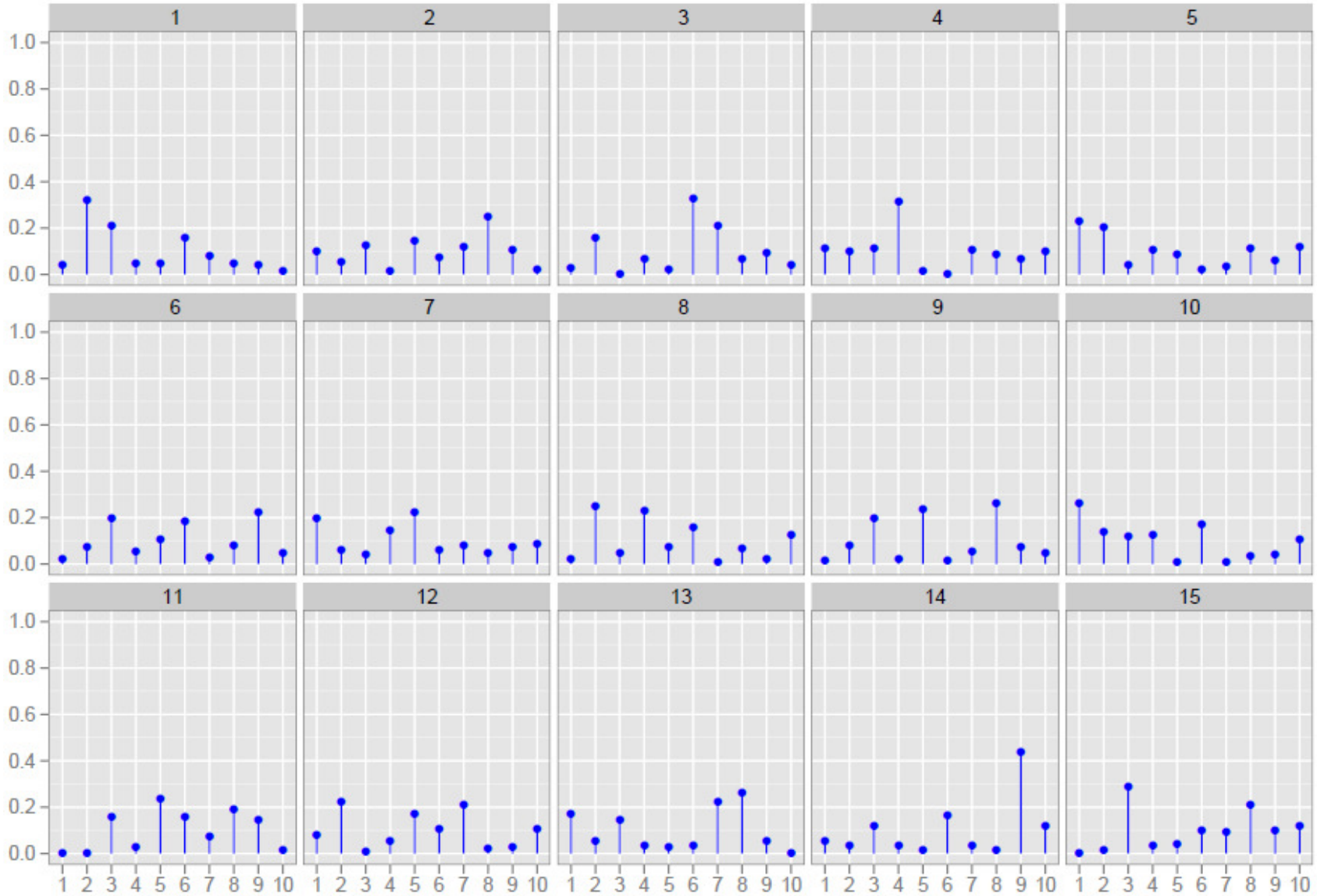
Распределение Дирихле при $\alpha \equiv 100$, 10 компонент



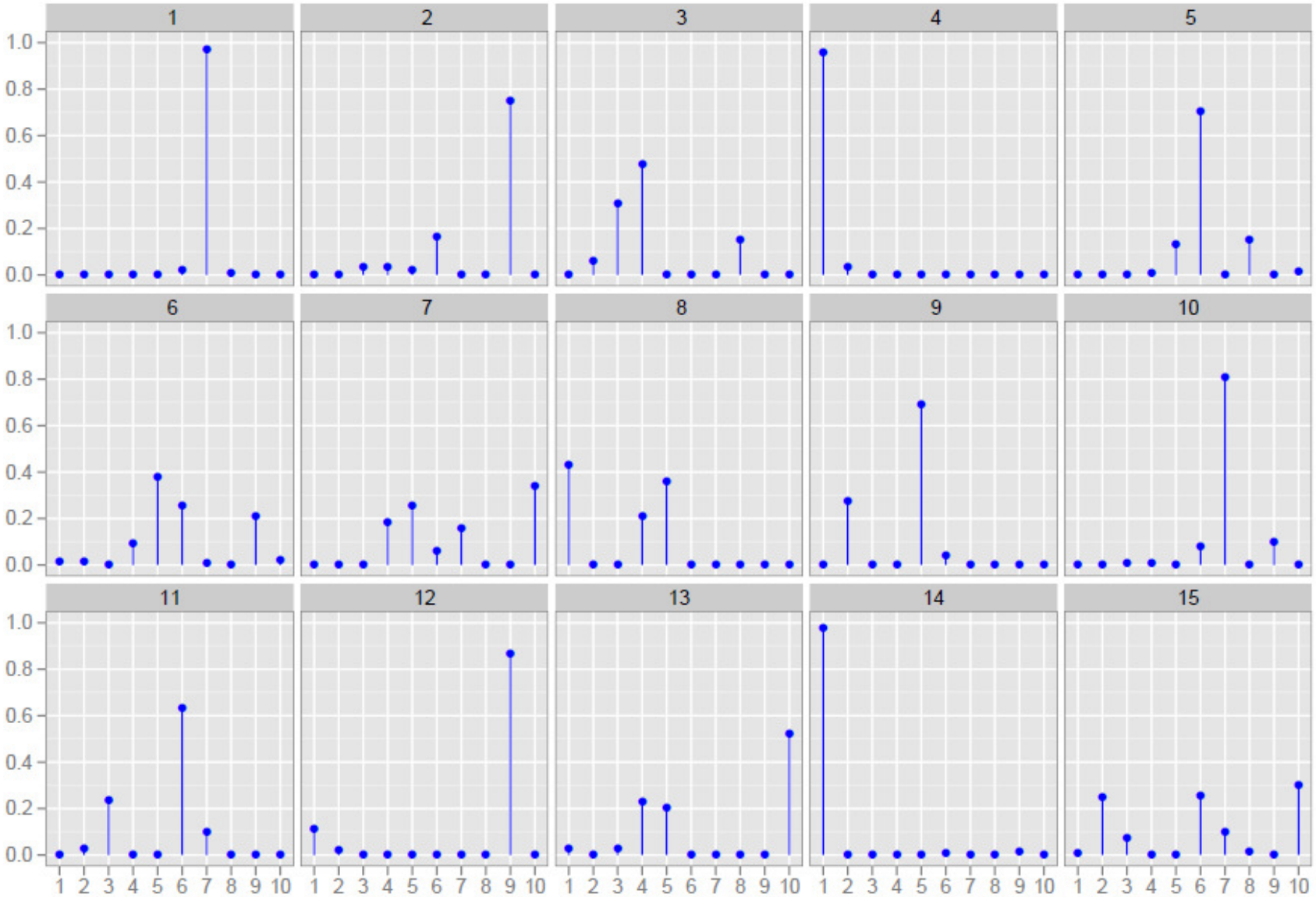
Распределение Дирихле при $\alpha \equiv 10$, 10 компонент



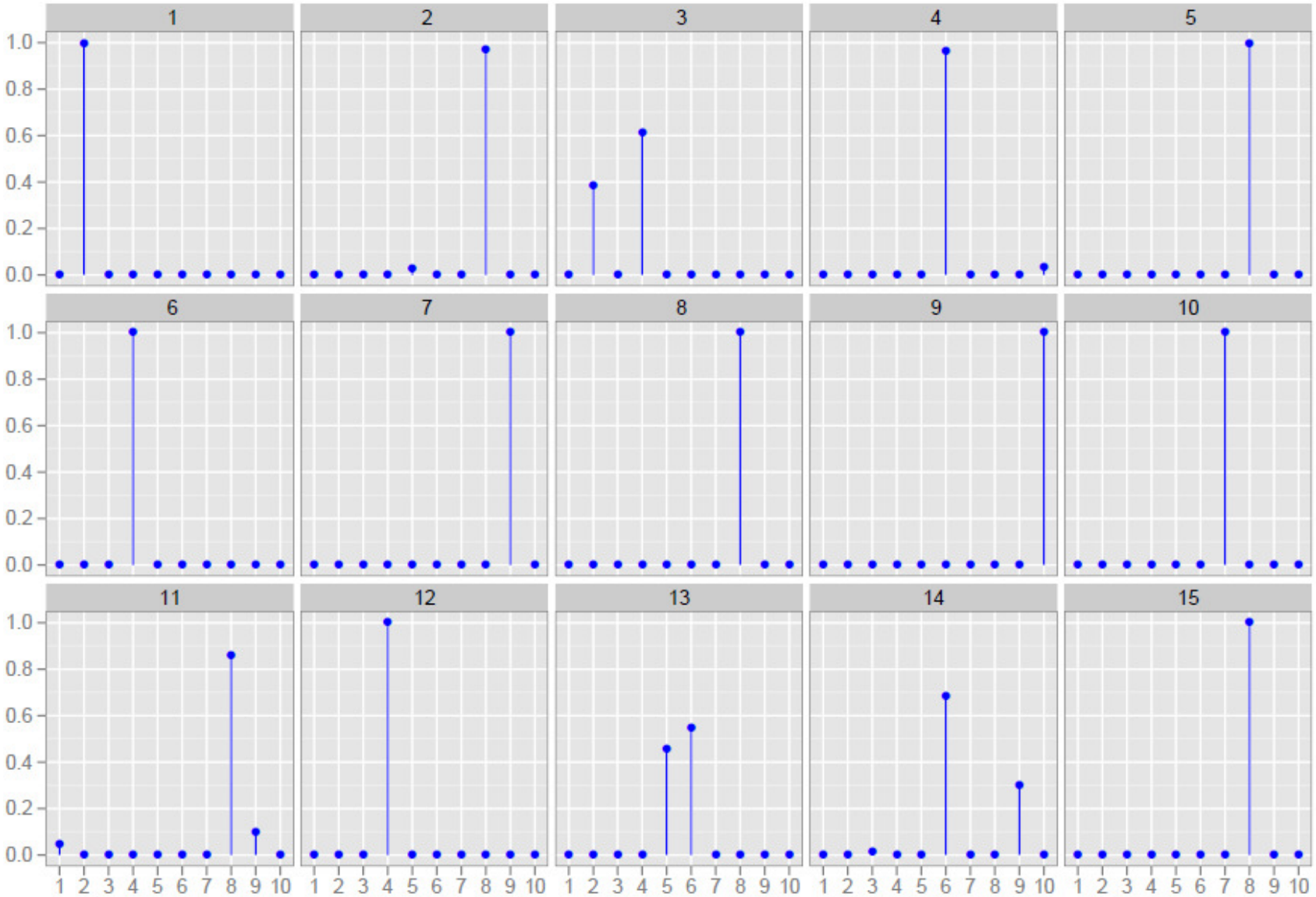
Распределение Дирихле при $\alpha \equiv 1$, 10 компонент



Распределение Дирихле при $\alpha \equiv 0.1$, 10 компонент



Распределение Дирихле при $\alpha \equiv 0.01$, 10 компонент



Принцип максимума апостериорной вероятности

Регуляризация log-правдоподобия:

$$L(w,\theta) + R(w) = \ln \prod_{i=1}^{m} p(x_i) + \ln \operatorname{Dir}(w|\alpha) =$$

$$= \sum_{i=1}^{m} \ln \sum_{j=1}^{k} w_j \varphi(x_i,\theta_j) + \ln \frac{\Gamma(\alpha_0)}{\prod_j \Gamma(\alpha_j)} + \sum_{j=1}^{k} (\alpha_j - 1) \ln w_j \;\to\; \max_{w,\theta}$$

Модификация формулы M-шага в случае разреженного симметричного распределения Дирихле, $\tau = 1 - \alpha_j \in (0,1)$:

$$L(w,\theta) - \tau \sum_{j=1}^{k} \ln w_j \;\to\; \max_{w,\theta};$$

$$w_j \propto \left(\sum_{i=1}^{m} g_{ij} - \tau \right)_+$$

Подбор τ — по максимуму правдоподобия тестовой выборки.

GEM — обобщённый EM-алгоритм

Идея:
Не обязательно добиваться высокой точности на M-шаге.
Достаточно лишь сместиться в направлении максимума,
сделав одну или несколько итераций, и затем выполнить E-шаг.

Преимущества:

- сохраняется свойство слабой локальной сходимости
  (в смысле увеличения правдоподобия на каждом шаге)
- повышается скорость сходимости при сопоставимом
  качестве решения

SEM — стохастический EM-алгоритм

Идея: на M-шаге вместо максимизации

$$\theta_j := \arg\max_{\theta} \sum_{i=1}^{m} g_{ij} \ln \varphi(x_i, \theta)$$

максимизируется обычное, невзвешенное, правдоподобие

$$\theta_j := \arg\max_{\theta} \sum_{x_i \in X_j} \ln \varphi(x_i, \theta),$$

выборки X_j строятся путём сэмплирования объектов из X^m m раз с возвращениями: $i \sim P(i \mid j) = \frac{P(j \mid x_i)P(i)}{P(j)} = \frac{g_{ij}}{mw_j}$.

Преимущества:
ускорение сходимости, предотвращение зацикливаний.

HEM — иерархический EM-алгоритм

Идея:
«Плохо описанные» компоненты расщепляются на две или более *дочерних* компонент.

Преимущество:
автоматически выявляется иерархическая структура каждого класса, которую затем можно интерпретировать содержательно.

Гауссовская смесь с диагональными матрицами ковариации

Гауссовская смесь GMM — Gaussian Mixture Model

Допущения:

1. Функции правдоподобия классов $p(x|y)$ представимы в виде смесей k_y компонент, $y \in Y = \{1, \ldots, M\}$.
2. Компоненты имеют n-мерные гауссовские плотности с некоррелированными признаками:

$\mu_{yj} = (\mu_{yj1}, \ldots, \mu_{yjn})$, $\Sigma_{yj} = \operatorname{diag}(\sigma^2_{yj1}, \ldots, \sigma^2_{yjn})$, $j = 1, \ldots, k_y$:

$$p(x|y) = \sum_{j=1}^{k_y} w_{yj} p_{yj}(x), \quad p_{yj}(x) = \mathcal{N}(x; \mu_{yj}, \Sigma_{yj}),$$

$$\sum_{j=1}^{k_y} w_{yj} = 1, \quad w_{yj} \geqslant 0;$$

Эмпирические оценки средних и дисперсий

Числовые признаки: $f_d : X \to \mathbb{R}, \quad d = 1, \ldots, n.$

Решение задачи M-шага:

для всех классов $y \in Y$ и всех компонент $j = 1, \ldots, k_y$,

$$w_{yj} = \frac{1}{\ell_y} \sum_{i:\, y_i = y} g_{yij}$$

для всех размерностей (признаков) $d = 1, \ldots, n$

$$\hat{\mu}_{yjd} = \frac{1}{\ell_y w_{yj}} \sum_{i:\, y_i = y} g_{yij} f_d(x_i);$$

$$\hat{\sigma}^2_{yjd} = \frac{1}{\ell_y w_{yj}} \sum_{i:\, y_i = y} g_{yij} \big(f_d(x_i) - \hat{\mu}_{yjd}\big)^2;$$

Замечание: компоненты «наивны», но смесь не «наивна».

Байесовский классификатор

Подставим гауссовскую смесь в байесовский классификатор:

$$a(x) = \arg\max_{y \in Y} \underbrace{\lambda_y P_y \sum_{j=1}^{k_y} w_{yj} \underbrace{\mathcal{N}_{yj} \exp\left(-\tfrac{1}{2}\rho_{yj}^2(x, \mu_{yj})\right)}_{p_{yj}(x)}}_{\Gamma_y(x)}$$

$\mathcal{N}_{yj} = (2\pi)^{-\frac{n}{2}}(\sigma_{yj1} \cdots \sigma_{yjn})^{-1}$ — нормировочные множители;
$\rho_{yj}(x, \mu_{yj})$ — *взвешенная евклидова метрика* в $X = \mathbb{R}^n$:

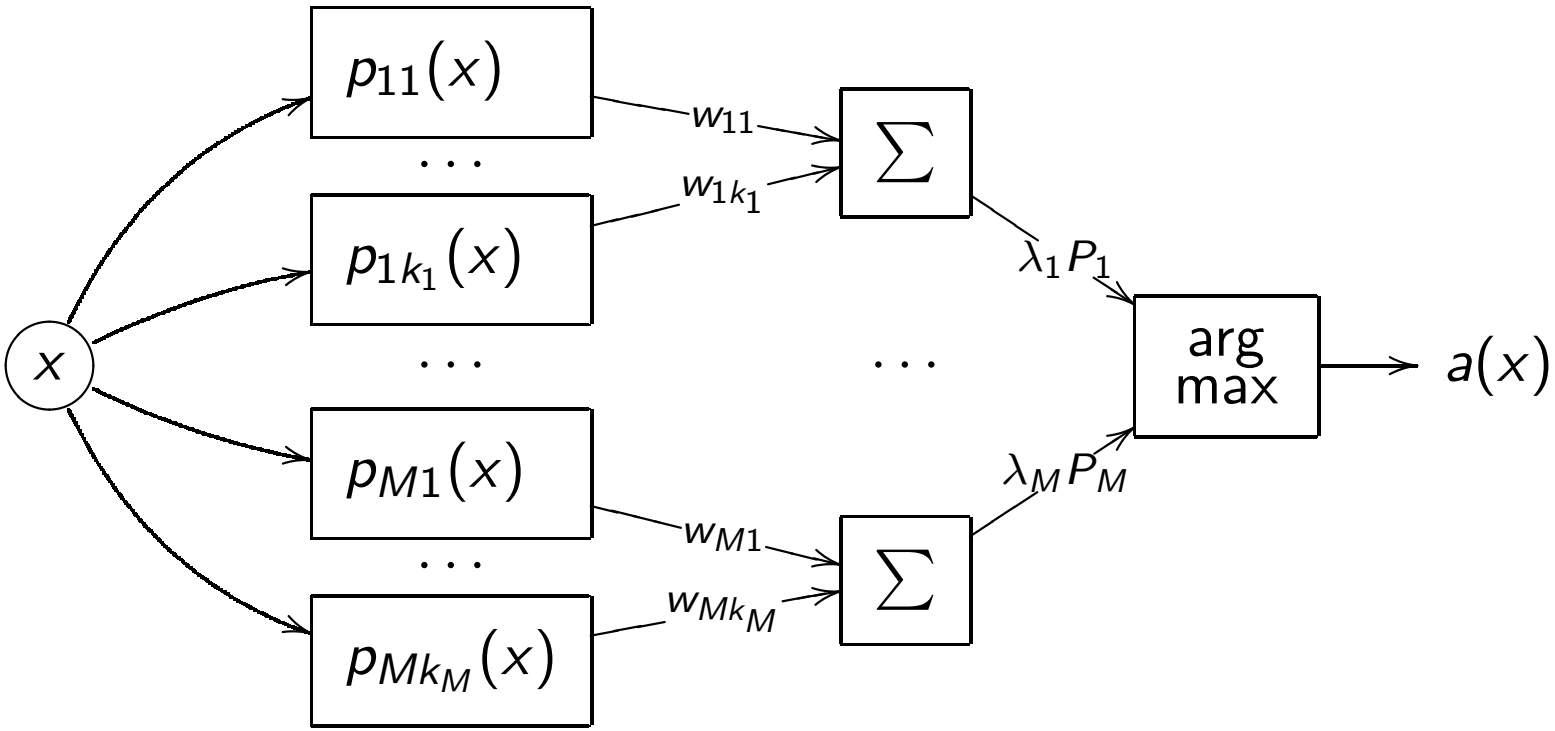
$$\rho_{yj}^2(x, \mu_{yj}) = \sum_{d=1}^{n} \frac{1}{\sigma_{yjd}^2}\left(f_d(x) - \mu_{yjd}\right)^2.$$

Интерпретация — как у метрического классификатора:
$p_{yj}(x)$ — близость объекта x к j-й компоненте класса y;
$\Gamma_y(x)$ — близость объекта x к классу y.

Байесовский классификатор — сеть RBF

Radial Basis Functions (RBF) — трёхуровневая суперпозиция:

$$a(x) = \arg\max_{y \in Y} \lambda_y P_y \sum_{j=1}^{k_y} w_{yj} p_{yj}(x)$$



Преимущества EM-RBF

EM — один из лучших алгоритмов обучения радиальных сетей.

Преимущества EM-алгоритма по сравнению с SVM:

1. EM-алгоритм легко сделать устойчивым к шуму
2. EM-алгоритм довольно быстро сходится
3. автоматически строится *структурное описание* каждого класса в виде совокупности компонент — *кластеров*

Недостатки EM-алгоритма:

1. EM-алгоритм чувствителен к начальному приближению
2. Определение числа компонент — трудная задача (простые эвристики могут плохо работать)

Постановка задачи кластеризации

Дано:

X — пространство объектов;

$X^\ell = \{x_i\}_{i=1}^\ell$ — обучающая выборка;

$\rho\colon X \times X \to [0, \infty)$ — функция расстояния между объектами.

Найти:

Y — множество кластеров,

$a\colon X \to Y$ — алгоритм кластеризации.

Критерии слабо формализованные:

— каждый кластер состоит из близких объектов;

— объекты разных кластеров существенно различны.

Кластеризация — это *обучение без учителя*.

Вероятностная формализация задачи кластеризации

Гипотеза о вероятностной природе данных:
Выборка X^ℓ случайна, независима, из смеси распределений

$$p(x) = \sum_{y \in Y} w_y p_y(x), \qquad \sum_{y \in Y} w_y = 1,$$

$p_y(x)$ — *плотность*, w_y — *априорная вероятность* кластера y.

Гипотеза о нормальности плотностей кластеров:

$$p_y(x) = (2\pi)^{-\frac{n}{2}} (\sigma_{y1} \cdots \sigma_{yn})^{-1} \exp\left(-\tfrac{1}{2}\rho_y^2(x, \mu_y)\right),$$

$x \equiv \big(f_1(x), \ldots, f_n(x)\big) \in \mathbb{R}^n$ — признаковое описание объекта x;

$\mu_y = (\mu_{y1}, \ldots, \mu_{yn})$ — центр кластера y;

σ_{yd}^2 — дисперсия значений признака f_d в классе y;

$\rho_y^2(x, \mu_y) = \sum\limits_{d=1}^{n} \frac{1}{\sigma_{yd}^2} |f_d(x) - \mu_{yd}|^2$ — взвешенная евклидова метрика.

EM-алгоритм (повторение)

1: начальное приближение w_y, μ_y, Σ_y для всех $y \in Y$;
2: **повторять**
3: E-шаг (expectation):

$$g_{iy} := P(y|x_i) \equiv \frac{w_y p_y(x_i)}{\sum_{z \in Y} w_z p_z(x_i)}, \quad y \in Y, \quad i = 1, \ldots, \ell;$$

4: M-шаг (maximization):

$$w_y := \frac{1}{\ell} \sum_{i=1}^{\ell} g_{iy}, \quad y \in Y;$$

$$\mu_{yd} := \frac{1}{\ell w_y} \sum_{i=1}^{\ell} g_{iy} f_d(x_i), \quad y \in Y, \quad d = 1, \ldots, n;$$

$$\sigma^2_{yd} := \frac{1}{\ell w_y} \sum_{i=1}^{\ell} g_{iy} (f_d(x_i) - \mu_{yd})^2, \quad y \in Y, \quad d = 1, \ldots, n;$$

5: $y_i := \arg\max_{y \in Y} g_{iy}, \quad i = 1, \ldots, \ell;$
6: **пока** y_i не перестанут изменяться;

Метод k-средних (k-means)

$X = \mathbb{R}^n$. Упрощённый аналог EM-алгоритма:
— жёсткая кластеризация вместо мягкой,
— метрика фиксирована, дисперсии не настраиваются.

1: начальное приближение центров μ_y, $y \in Y$;
2: **повторять**
3: аналог E-шага:
отнести каждый x_i к ближайшему центру:
$y_i := \arg\min_{y \in Y} \rho(x_i, \mu_y)$, $i = 1, \ldots, \ell$;
4: аналог M-шага:
вычислить новые положения центров:
$$\mu_{yd} := \frac{\sum_{i=1}^{\ell} [y_i = y] f_d(x_i)}{\sum_{i=1}^{\ell} [y_i = y]}, \quad y \in Y, \quad d = 1, \ldots, n;$$
5: **пока** y_i не перестанут изменяться;

Модификации и обобщения

Основные отличия GMM-EM и k-means:

- GMM-EM: мягкая кластеризация: $g_{iy} = \mathrm{P}\{y_i = y\}$
  k-means: жёсткая кластеризация: $g_{iy} = [y_i = y]$
- GMM-EM: кластеры эллиптические, настраиваемые
  k-means: кластеры сферические, не настраиваемые

Гибриды (упрощение GMM-EM — усложнение k-means):

- GMM-EM с жёсткой кластеризацией на E-шаге
- GMM-EM без настройки дисперсий (сферические гауссианы)

Недостатки k-means:

- чувствительность к выбору начального приближения
- медленная сходимость (пользуйтесь k-means++)

Частичное обучение (Semi-supervised learning)

Дано:

Y — множество кластеров;

$\{x_i\}_{i=1}^{\ell}$ — обучающая выборка;

$\{x_i, y_i\}_{i=\ell+1}^{\ell+m}$ — размеченная часть выборки, обычно $m \ll \ell$.

Найти:

$a\colon X \to Y$ — алгоритм кластеризации.

Как приспособить EM-алгоритм:

E-шаг: $g_{iy} := [y = y_i]$, $y \in Y$, $i = \ell{+}1, \ldots, \ell{+}m$;

Как приспособить k-means:

E-шаг: $y_i := \arg\min_{y \in Y} \rho(x_i, \mu_y)$, $i = 1, \ldots, \ell$.

Резюме в конце лекции

- EM-алгоритм — мощный метод восстановления скрытой информации по наблюдаемым данным
- Применяется для разделения смесей распределений, кластеризации, классификации
- Кластеризация k-means — упрощение EM-алгоритма
- Имеет многочисленные модификации, в том числе для определения числа компонент