

Домашнее задание по вариационному выводу

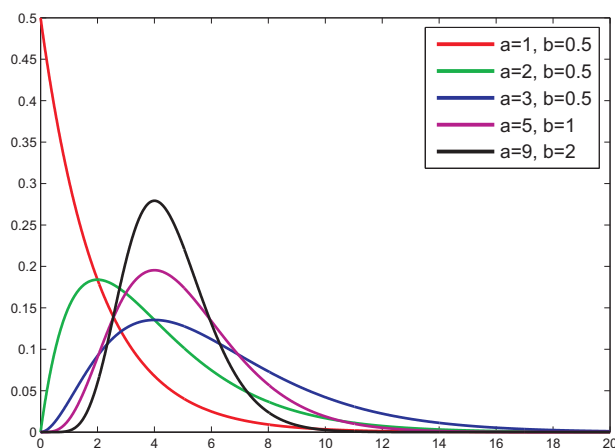
Дата: 5 мая 2013

Ликбез: гамма-распределение

Гамма-распределение является вероятностным распределением для действительной положительной переменной λ и имеет плотность:

$$\mathcal{G}(\lambda|a, b) = \frac{b^a}{\Gamma(a)} \lambda^{a-1} \exp(-b\lambda), \quad a, b > 0.$$

Здесь $\Gamma(a)$ – гамма-функция. Различные виды гамма-распределения:



С помощью гамма-распределения можно задать широкий спектр унимодальных несимметричных распределений на положительной полуоси.

Статистики гамма-распределения:

$$\begin{aligned} \mathbb{E}\lambda &= \frac{a}{b}, \\ \mathbb{D}\lambda &= \frac{a}{b^2}, \\ \mathbb{E} \log \lambda &= \Psi(a) - \log b. \end{aligned}$$

Здесь $\Psi(a) = \frac{d}{da} \log \Gamma(a)$ – дигамма функция.

Можно показать, что при $a = b \rightarrow 0$ гамма-распределение переходит в равномерное распределение на параметр λ в логарифмической шкале.

Ликбез: распределение Дирихле

Случайная величина $\boldsymbol{\theta} \in \mathbb{R}^K$, определенная на симплексе ($\theta_k \geq 0, \sum_{k=1}^K \theta_k = 1$), имеет распределение Дирихле, если ее плотность определяется как

$$p(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \text{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha}) = \frac{\Gamma(\sum_k \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K \theta_k^{\alpha_k - 1}, \quad \alpha_k > 0.$$

Здесь $\Gamma(\cdot)$ – гамма-функция, $\boldsymbol{\alpha}$ – набор параметров распределения. Различные виды распределения Дирихле для случая $K = 3$ показаны на рис. 1. Заметим, что в случае $\alpha_1 = \dots = \alpha_K = 1$ распределение Дирихле переходит в равномерное распределение на симплексе.

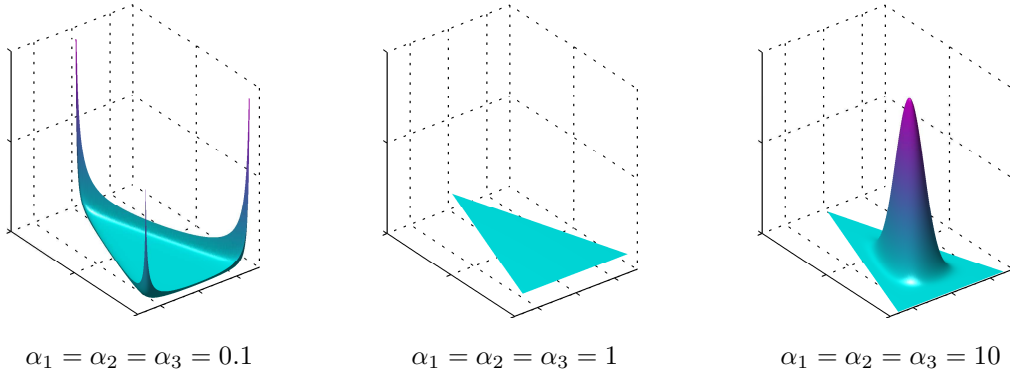


Рис. 1: Различные виды распределения Дирихле

Статистики распределения Дирихле:

$$\begin{aligned} \mathbb{E}_p \theta_i &= \frac{\alpha_i}{\alpha_0}, \\ \text{Cov}(\theta_i, \theta_j) &= \frac{\alpha_i \alpha_0 [i=j] - \alpha_i \alpha_j}{\alpha_0^2 (\alpha_0 + 1)}, \\ \alpha_0 &= \sum_k \alpha_k, \\ \mathbb{E}_p \log \theta_i &= \Psi(\alpha_i) - \Psi\left(\sum_k \alpha_k\right). \end{aligned}$$

Здесь $\Psi(x) = \frac{d}{dx} \log \Gamma(x)$ – дигамма функция.

Распределение Дирихле часто используется в качестве априорного распределения для набора дискретных вероятностей. Рассмотрим дискретную случайную величину, принимающую K значений:

$$\begin{array}{cccc} 1 & 2 & \dots & K \\ \theta_1 & \theta_2 & \dots & \theta_K \end{array}$$

Рассмотрим задачу оценки параметров $\boldsymbol{\theta}$ этой случайной величины по выборке из нее объема N с помощью метода максимального правдоподобия:

$$p(X|\boldsymbol{\theta}) = \prod_{n=1}^N p(x_n|\boldsymbol{\theta}) = \prod_{n=1}^N \theta_{x_n} \rightarrow \max_{\boldsymbol{\theta}: \theta_k \geq 0, \sum_k \theta_k = 1}$$

Данная задача условной оптимизации может быть решена аналитически с помощью функции Лагранжа L :

$$\begin{aligned} L(\boldsymbol{\theta}, \lambda) &= \log p(X|\boldsymbol{\theta}) + \lambda \left(\sum_k \theta_k - 1 \right) = \sum_{n=1}^N \log \theta_{x_n} + \lambda \left(\sum_k \theta_k - 1 \right) = \\ &= \sum_{k=1}^K \log \theta_k \left(\sum_{n=1}^N [x_n = k] \right) + \lambda \left(\sum_{k=1}^K \theta_k - 1 \right). \end{aligned}$$

Приравняв производные функции Лагранжа к нулю и суммируя по k , получаем:

$$\frac{\partial}{\partial \theta_k} L(\boldsymbol{\theta}, \lambda) = \frac{\sum_{n=1}^N [x_n = k]}{\theta_k} + \lambda = 0 \Rightarrow \theta_k = -\frac{1}{\lambda} \sum_{n=1}^N [x_n = k], \Rightarrow \lambda = -\sum_{k=1}^K \sum_{n=1}^N [x_n = k] = -N.$$

Таким образом, оценка максимального правдоподобия для параметров $\boldsymbol{\theta}$ определяется частотами:

$$\theta_k = \frac{\sum_{n=1}^N [x_n = k]}{N}. \quad (1)$$

Введем распределение Дирихле $\text{Dir}(\boldsymbol{\theta}|\boldsymbol{\alpha})$ в качестве априорного распределения для параметров $\boldsymbol{\theta}$ и рассмотрим оценку максимума апостериорного распределения:

$$p(\boldsymbol{\theta}|X, \boldsymbol{\alpha}) \rightarrow \max_{\boldsymbol{\theta}} \Leftrightarrow p(X|\boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \rightarrow \max_{\boldsymbol{\theta}}. \quad (2)$$

Действуя аналогично случаю максимума правдоподобия, получаем следующее решение данной задачи оптимизации:

$$\theta_k = \frac{\alpha_k - 1 + \sum_{n=1}^N [x_n = k]}{\sum_{j=1}^K \alpha_j - K + N}. \quad (3)$$

Заметим, что в случае равномерного априорного распределения ($\alpha_1 = \dots = \alpha_K = 1$) данное решение переходит в оценку максимального правдоподобия (1). При всех $\alpha_k > 1$ решение (3) является менее контрастным, чем решение (1), и, в частности, задает ненулевую вероятность для исходов, ни разу не наблюдавшихся в обучающей выборке. В этом случае происходит сглаживание вероятностей. Напротив, при $\alpha_k < 1$ решение (3) является более контрастным по сравнению с (1), т.к. в этом случае априорное распределение имеет большой вес у границ симплекса. Например, в случае $\alpha_1 = \alpha_2 = \alpha_3 = 0.1$ и выборки из одной единицы, двух двоек и трех троек оценки максимального правдоподобия и максимального апостериорного распределения соответственно равны:

$$\begin{aligned} \theta_{ML,1} &= \frac{1}{6}, \quad \theta_{ML,2} = \frac{1}{3}, \quad \theta_{ML,3} = \frac{1}{2}, \\ \theta_{MP,1} &= \frac{1}{33}, \quad \theta_{MP,2} = \frac{11}{33}, \quad \theta_{MP,3} = \frac{21}{33}. \end{aligned}$$

Пусть для некоторых $k \in \{1, \dots, K\}$ значение $\alpha_k - 1 + \sum_n [x_n = k] \leq 0$. Обозначим множество таких индексов через $K_{\leq 0}$, а множество оставшихся индексов — через $K_{> 0}$. Тогда можно показать, что решение задачи (2) вместо (3) становится следующим:

$$\theta_{MP,k} = \begin{cases} 0, & \text{если } k \in K_{\leq 0}, \\ \frac{\alpha_k - 1 + \sum_n [x_n = k]}{\sum_{j \in K_{> 0}} (\alpha_j - 1 + \sum_n [x_n = j])}, & \text{иначе.} \end{cases}$$

Задача 1

Рассматривается стандартная задача регрессии. Имеется обучающая выборка $(\mathbf{t}, X) = \{t_n, \mathbf{x}_n\}_{n=1}^N$, где $\mathbf{x}_n \in \mathbb{R}^d$ — вектор признаков n -го объекта, а $t_n \in \mathbb{R}$ — его целевая переменная. Необходимо на основе этих данных найти прогноз значения t_{new} для нового объекта, представленного только своим вектором признаков \mathbf{x}_{new} .

Рассмотрим решение этой задачи с помощью модели байесовской линейной регрессии:

$$\begin{aligned} p(\mathbf{t}, \mathbf{w}, \alpha, \beta|X) &= p(\mathbf{t}|X, \mathbf{w}, \beta)p(\mathbf{w}|\alpha)p(\alpha)p(\beta), \\ p(\mathbf{t}|X, \mathbf{w}, \beta) &= \mathcal{N}(\mathbf{t}|X\mathbf{w}, \beta^{-1}I), \\ p(\mathbf{w}|\alpha) &= \mathcal{N}(\mathbf{w}|\mathbf{0}, \alpha^{-1}I), \\ p(\alpha) &= \mathcal{G}(\alpha|a, b), \\ p(\beta) &= \mathcal{G}(\beta|c, d). \end{aligned}$$

Здесь $\mathbf{t} \in \mathbb{R}^N$, $X \in \mathbb{R}^{N \times d}$ — обучающие данные, $\mathbf{w} \in \mathbb{R}^d$, $\alpha, \beta \in \mathbb{R}$ — скрытые переменные, $a, b, c, d \in \mathbb{R}$ — параметры модели. Заметим, что данная модель представляет собой стандартную модель линейной регрессии $\hat{t}(\mathbf{x}) = \mathbf{w}^T \mathbf{x}$ с весами \mathbf{w} в предположении нормального шума с дисперсией β^{-1} , а также с добавлением априорных распределений на веса и на дисперсию шума в виде гамма-распределений.

Необходимо с помощью вариационного вывода найти факторизованное приближение для апостериорного распределения на скрытые переменные:

$$p(\mathbf{w}, \alpha, \beta|\mathbf{t}, X) \approx q_{\mathbf{w}}(\mathbf{w})q_{\alpha}(\alpha)q_{\beta}(\beta).$$

Требуется выписать формулы пересчета для всех компонент факторизованного приближения $q_{\mathbf{w}}(\mathbf{w})$, $q_{\alpha}(\alpha)$, $q_{\beta}(\beta)$, а также вид оптимизируемого функционала $\mathcal{L}\{q\}$.

Задача 2

Рассмотрим модель смеси из K нормальных распределений, в которой добавляется априорное распределение Дирихле на веса компонент смеси:

$$\begin{aligned} p(X, T, \boldsymbol{\theta} | \boldsymbol{\alpha}, \{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K) &= \left[\prod_{n=1}^N p(\mathbf{x}_n | t_n, \{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K) p(t_n | \boldsymbol{\theta}) \right] p(\boldsymbol{\theta} | \boldsymbol{\alpha}), \\ p(\mathbf{x}_n | t_n, \{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K) &= \mathcal{N}(\mathbf{x}_n | \boldsymbol{\mu}_{t_n}, \Sigma_{t_n}), \\ p(t_n | \boldsymbol{\theta}) &= \theta_{t_n}, \\ p(\boldsymbol{\theta} | \boldsymbol{\alpha}) &= \text{Dir}(\boldsymbol{\theta} | \boldsymbol{\alpha}). \end{aligned}$$

Здесь $X = \{\mathbf{x}_n\}_{n=1}^N$, $\mathbf{x}_n \in \mathbb{R}^d$ – наблюдаемые данные, $T = \{t_n\}_{n=1}^N$, $t_n \in \{1, \dots, K\}$ – номера компонент смеси для каждого объекта, $\boldsymbol{\alpha}, \{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K$ – параметры модели.

Требуется записать формулы вариационного EM-алгоритма для решения задачи обучения

$$p(X | \boldsymbol{\alpha}, \{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K) \rightarrow \max_{\{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K} .$$

Здесь, в частности, необходимо рассмотреть факторизованное приближение для апостериорного распределения вида

$$p(T, \boldsymbol{\theta} | X, \boldsymbol{\alpha}, \{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K) \approx q_T(T) q_{\boldsymbol{\theta}}(\boldsymbol{\theta}).$$

Требуется выписать формулы пересчета для компонент факторизованного приближения на E-шаге $q_T(T)$, $q_{\boldsymbol{\theta}}(\boldsymbol{\theta})$, формулы пересчета для параметров $\{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K$ на M-шаге, а также вид оптимизируемого в итерациях функционала $\mathcal{L}\{q\}$.