Курс: Графические модели, 2014

Задачи для подготовки к контрольной по вариационному выводу

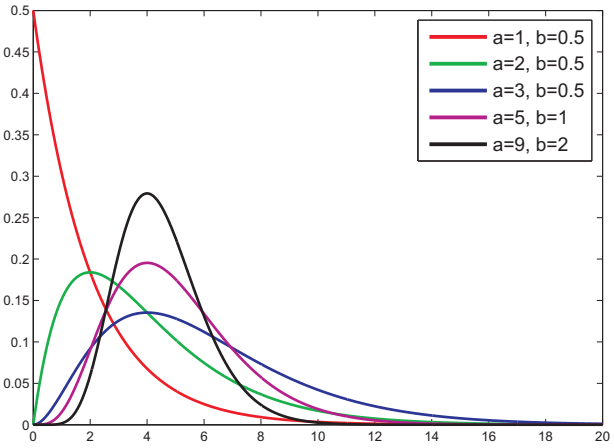
Дата: 8 мая 2014

Ликбез: гамма-распределение

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

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

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:

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

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

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

Ликбез: распределение Стьюдента

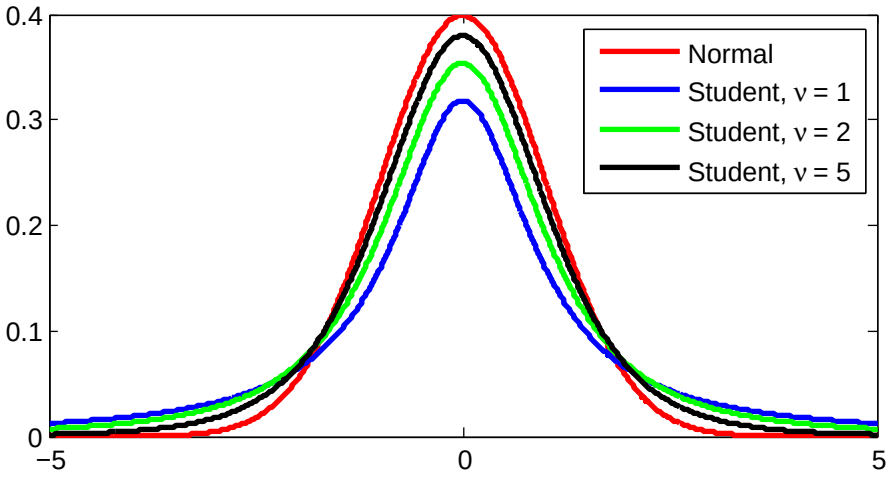
Случайная величина $x \in \mathbb{R}$ имеет распределение Стьюдента с числом степеней свободы $\nu > 0$, если её плотность равна

$$\mathcal{T}(x|\mu,\sigma,\nu) \propto \left[1 + \frac{1}{\nu}\left(\frac{x-\mu}{\sigma}\right)^2\right]^{-\frac{\nu+1}{2}}.$$

Статистики распределения:

$$\mathbb{E}x = \mu, \text{ если } \nu > 1,$$
$$\mathbb{D}x = \frac{\nu\sigma^2}{\nu-2}, \text{ если } \nu > 2,$$
$$\operatorname{Mod} x = \mu.$$

При $\nu \to \infty$ распределение Стьюдента переходит в нормальное распределение с параметрами μ, σ^2. Виды распределения Стьюдента для различных ν:



Видно, что плотность распределения Стьюдента имеет более тяжелые хвосты по сравнению с нормальным распределением. Благодаря этому свойству, восстановление распределения Стьюдента по данным является более робастной процедурой. Можно показать, что

$$\mathcal{T}(x|\mu,\sigma,\nu) = \int \mathcal{N}(x|\mu,\sigma^2/\lambda)\mathcal{G}(\lambda|\nu/2,\nu/2)d\lambda. \tag{1}$$

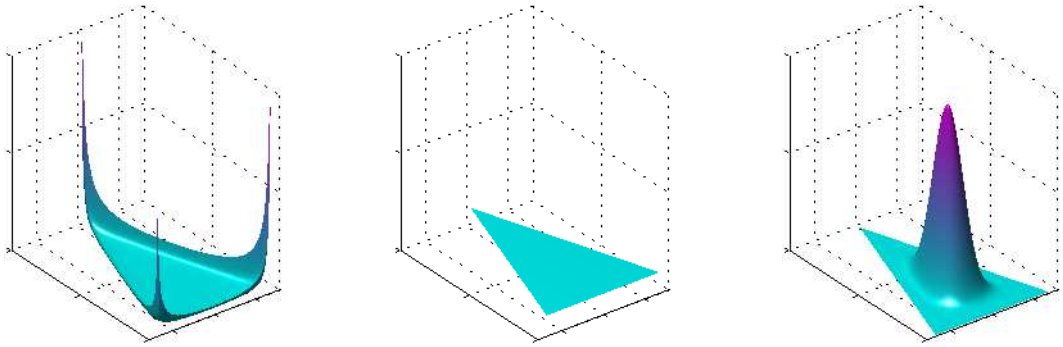
Таким образом, распределение Стьюдента является бесконечной смесью нормальных распределений с весами компонент из гамма-распределения.

Ликбез: распределение Дирихле

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

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

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 $\alpha_1 = \cdots = \alpha_K = 1$ распределение Дирихле переходит в равномерное распределение на симплексе.



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

Рис. 1: Различные виды распределения Дирихле

Статистики распределения Дирихле:

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

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:

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

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

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

Приравнивая 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:

$$\theta_k = \frac{\sum_{n=1}^{N} [x_n = k]}{N}. \tag{2}$$

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

$$p(\boldsymbol{\theta}|X, \boldsymbol{\alpha}) \to \max_{\boldsymbol{\theta}} \;\Leftrightarrow\; p(X|\boldsymbol{\theta}) p(\boldsymbol{\theta}|\boldsymbol{\alpha}) \to \max_{\boldsymbol{\theta}}. \tag{3}$$

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

$$\theta_k = \frac{\alpha_k - 1 + \sum_{n=1}^{N} [x_n = k]}{\sum_{j=1}^{K} \alpha_j - K + N}. \tag{4}$$

Заметим, что в случае равномерного априорного распределения ($\alpha_1 = \dots \alpha_K = 1$) данное решение переходит в оценку максимального правдоподобия (2). При всех $\alpha_k > 1$ решение (4) является менее контрастным, чем решение (2), 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 (4) является более контрастным по сравнению с (2), 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:

$$\theta_{ML,1} = \frac{1}{6},\ \theta_{ML,2} = \frac{1}{3},\ \theta_{ML,3} = \frac{1}{2},$$
$$\theta_{MP,1} = \frac{1}{33},\ \theta_{MP,2} = \frac{11}{33},\ \theta_{MP,3} = \frac{21}{33}.$$

Пусть для некоторых $k \in \{1, \ldots, K\}$ 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 (3) вместо (4) 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

Рассмотрим одномерную смесь распределений Стьюдента:

$$p(x) = \sum_{k=1}^{K} \pi_k \mathcal{T}(x|\mu_k, \sigma_k, \nu),\ \pi_k > 0, \sum_{k=1}^{K} \pi_k = 1. \tag{5}$$

Пусть имеется независимая выборка $X = \{x_n\}_{n=1}^N$, $x_n \in \mathbb{R}$ из модели (5). Тогда, используя (1) и стандартное представление смеси вероятностных распределений через скрытые дискретные переменные, модель (5) можно эквивалентно переписать в виде следующей вероятностной модели со скрытыми переменными:

$$p(X, \Lambda, Z|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \nu) = \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\mathcal{N}(x_n|\mu_k, \sigma_k^2/\lambda_k)\mathcal{G}(\lambda_k|\nu/2, \nu/2)\right]^{z_{nk}} \pi_k^{z_{nk}}.$$

Здесь переменные $z_{nk} \in \{0, 1\}$ обозначают принадлежность объекта x_n к k-ой компоненте смеси.

Пусть количество степеней свободы ν известно. Требуется выписать формулы вариационного EM-алгоритма для решения задачи обучения

$$p(X|\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}, \nu) \to \max_{\boldsymbol{\pi}, \boldsymbol{\mu}, \boldsymbol{\sigma}},$$

где на E-шаге используется следующее факторизованное приближение для апостериорного распределения:

$$p(Z, \Lambda|X, \boldsymbol{\pi}, \boldsymbol{\mu}, \sigma, \nu) \approx q(Z)q(\Lambda).$$

Необходимо выписать формулы пересчёта компонент $q(Z)$, $q(\Lambda)$ на E-шаге, формулы пересчёта для параметров π_k, μ_k, σ_k на M-шаге, а также вид оптимизируемого функционала $\mathcal{L}(q)$.

Задача 2

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

$$p(X, Z, \boldsymbol{\pi}|\{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K, \boldsymbol{\alpha}) = \mathrm{Dir}(\boldsymbol{\pi}|\boldsymbol{\alpha}) \prod_{n=1}^{N} \prod_{k=1}^{K} \left[\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)\pi_k\right]^{z_{nk}}.$$

Здесь $X = \{\boldsymbol{x}_n\}_{n=1}^N$, $\boldsymbol{x}_n \in \mathbb{R}^d$ – наблюдаемые данные, $Z = \{z_{nk}\}_{n,k=1}^{N,K}$, $z_{nk} \in \{0, 1\}$ обозначает принадлежность объекта $\boldsymbol{x}_n$ к k-ой компоненте смеси, $\boldsymbol{\alpha}, \{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K$ – параметры модели.

Требуется записать формулы вариационного EM-алгоритма для решения задачи обучения

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

где на E-шаге используется факторизованное приближение вида

$$p(Z, \boldsymbol{\pi} | X, \boldsymbol{\alpha}, \{\boldsymbol{\mu}_k, \Sigma_k\}_{k=1}^K) \approx q(Z)q(\boldsymbol{\pi}).$$

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