Вариационный вывод

1. Дивергенция Кульбака-Лейблера: $KL(p||q) = \int p(x) \log \frac{p(x)}{q(x)} dx$

- $KL(p||q) \geq 0$
- $KL(p||q) = 0 \iff p = q$
- $KL(p||q) \neq KL(q||p)$

Мат.ожидание это выпуклая комбинация, поэтому

$$-KL(q||p) = \int q \log \frac{p}{q} dx \leq \{Convexivity\ of\ log\} \leq$$

$$\leq \log \int \frac{pq}{q} dx = \log \int p dx = 0$$

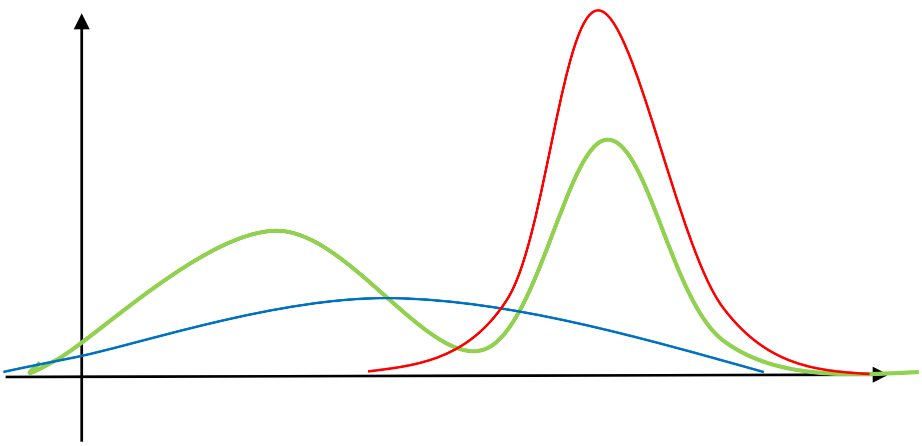
Мера схожести между двумя распределениями. Удобна для приближения «сложного» распреления p(x) более «простым» распределением q(x).

«Простота» многомерных распределений определяется степенью их факторизуемости. Чем на большее число множителей факторизуется совместная плотность, тем распределение «проще»

$$\boxed{p(X) \approx q(X) = \prod_i q_i(x_i)}$$

Минимизация прямой КЛ-дивергенции: $KL(q||p) \rightarrow \min_{q \in \mathcal{Q}}$ Вариационный вывод

Минимизация обратной КЛ-дивергенции: $KL(p||q) \rightarrow \min_{q \in \mathcal{Q}}$ Expectation propagation



2. Пусть $q(X) = \prod_i q_i(x_i) = \prod_i q_i$. Тогда

$$KL(q||p) = \int \prod_i q_i \log \frac{\prod_i q_i}{p} dX = \int \prod_i q_i \sum_i \log q_i dX - \int \prod_i q_i \log p \prod_i dx_i =$$

$$\{\text{minimize wrt } q_j \text{ given all other } q_i \text{ fixed}\} =$$

$$\int q_j \log q_j dx_j + \text{Const} - \int q_j \underbrace{\left(\int \prod_{i \neq j} q_i \log p \prod_{i \neq j} dx_i \right)}_{\log \hat{p}(x_j)} dx_j = KL(q_j||\hat{p}) + \text{Const} \rightarrow \min_{q_j}$$

Отсюда основная формула вар.вывода

$$\boxed{q_j(x_j) = \frac{\exp\left(\mathbb{E}_{q_i: i \neq j} \log p(X)\right)}{Z}}$$

Пусть $f(x) > 0$, $\int f(x) dx < \infty$. Тогда $f(x) = f_0(x) Z$, где $f_0(x)$ – плотность распределения. Отсюда

$$\arg \min_{\int q dx = 1} \int q(x) \log \frac{q(x)}{f(x)} dx =$$

$$\arg \min_{\int q dx = 1} \int q(x) \log \frac{q(x)}{f_0(x)} dx - Z = f_0(x)$$

3. Вариационный ЕМ-алгоритм.

Рассмотрим задачу максимизации неполного правдоподобия $p(X|\theta) \to \max_\theta$

$$\log p(X|\theta) = \mathcal{L}(q,\theta) + KL(q(T)||p(T|X,\theta))$$
$$\text{Е-шаг: } \mathcal{L}(q(T),\theta) \to \max_q \iff KL(q(T)||p(T|X,\theta)) \to \min_q$$
$$\text{М-шаг: } \mathcal{L}(q(T),\theta) \to \max_\theta \iff \mathbb{E}\log p(X,T|\theta) \to \max\theta$$

Если правдоподобие и априорное распределение не образуют пару сопряженных, аналитический подсчет $p(T|X,\theta)$ невозможен и Е-шаг не может быть выполнен точно

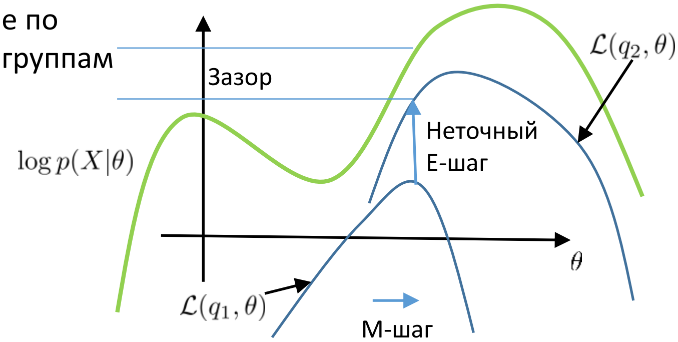
$$\text{Вариационный Е-шаг: } q(T) = \arg\min_{q\in\mathcal{Q}} KL(q(T)||p(T|X,\theta))$$
$$\text{где } \mathcal{Q} = \{q(T)|q(T) = \textstyle\prod_i q_i(t_i)\}$$

Замечание 1. Частным случаем факторизованного семейства является семейство дельта-функций $\mathcal{Q}' = \{q(T)|q(T) = \delta(T-T_0)\}$, поэтому «жесткий» (crisp) ЕМ-алгоритм проводит минимизацию КЛ-дивергенции в более узком семействе чем вариационный ЕМ.

$$\text{Жесткий Е-шаг: } q(T) = \delta(T - T_{MP}), \text{ где } T_{MP} = \arg\max_T p(T|X,\theta) = \arg\max_T p(X,T|\theta)$$

Замечание 2. В обоих вариантах нижняя граница $\mathcal{L}(q,\theta)$ по-прежнему монотонно возрастает, хотя и перестает быть точной после Е-шага.

Замечание 3. Можно проводить факторизацию не по отдельным переменным по непересекающимся группам



4. Распределение Джеффриса.

Представим, что у нас нет никаких предпочтений на масштаб измеряемой величины, тогда

$$P(X\in[c_1,c_2]) = P(X\in[\alpha c_1,\alpha c_2])$$

$$\int_{c_1}^{c_2} p(x)dx = \int_{\alpha c_1}^{\alpha c_2} p(x)dx = \frac{1}{\alpha}\int_{c_1}^{c_2} p\left(\frac{y}{\alpha}\right)dy$$

$$p(x) = \frac{1}{\alpha}p\left(\frac{x}{\alpha}\right) \Rightarrow p(x) \sim \frac{1}{x}$$

Это несобственное распределение Джеффриса, отражающее инвариантность к масштабу. Если нет предпочтений на значения параметра масштаба распределения, в качестве априорного надо брать распределение Джеффриса. Оно же объясняет парадокс Бенфорда.

Парадокс Бенфорда. При измерении любых величин в случайно выбранном масштабе, единица в качестве первой значащей цифры будет встречаться примерно 6 раз чаще девятки!

Легко показать, что это предельный случай гамма-распределения $\mathcal{G}(x|a,b) = \frac{b^a}{\Gamma(a)}x^{a-1}\exp(-bx)$

$$\lim_{a\to 0 b\to 0}\mathcal{G}(x|a,b) \sim \frac{1}{x}$$

5. Вариационная линейная регрессия.

Рассматриваем стандартную задачу восстановления линейной регрессии.

Пусть $(X,T) = \{(x_i,t_i)\}_{i=1}^n$ обучающая выборка, $x\in\mathbb{R}^d$, $t\in\mathbb{R}$.

Дискриминативная вероятностная модель имеет следующий вид

$$p(t,w,\alpha,\beta|x) = p(t|w,x,\beta)p(w|\alpha)p(\alpha)p(\beta) = \mathcal{N}(t|w^Tx,\beta^{-1})\mathcal{N}(w|0,\alpha^{-1})\mathcal{G}(\alpha|a_0,b_0)\mathcal{G}(\beta|c_0,d_0)$$

На этапе обучения необходимо рассчитать апостериорное распределение на всю совокупность скрытых переменных $p(w,\alpha,\beta|X,T)$. Заметим, что хотя все компоненты вероятностной модели лежат в экспоненциальном классе распределений, правдоподобие $p(t|w,x,\beta)$ и априорное распределение $p(w|\alpha)p(\alpha)p(\beta)$ не образуют пару сопряженных распределений, поэтому аналитически посчитать $p(w,\alpha,\beta|X,T)$ не удается.

Заметим, что распределения по каждой отдельной группе переменных при фиксированных остальных становятся сопряженными.

Общий результат. Если совокупность скрытых переменных T можно разбить на непересекающиеся подмножества $T_1, \dots, T_k$ так, что для каждого из них при фиксированных значениях остальных скрытых переменных, правдоподобие и априорные распределения становятся сопряженными, можно получить явные формулы для итерационного пересчета вариационного приближения

$$q(T) = \prod_{l=1}^{k} q_l(T_l) = \arg \min_{q=\prod q_l} KL(q(T)||p(T|X))$$

Рассмотрим факторизованное приближение $q(w,\alpha,\beta) = q(w)q(\alpha)q(\beta)$. Получим итерационные формулы пересчета для каждого из множителей

$$\log q(w) = \int \log p(T,w,\alpha,\beta|X)q(\alpha)q(\beta)d\alpha d\beta + \text{Const} = \int \log p(T|w,\beta)q(\beta)d\beta +$$

$$+ \int \log p(w|\alpha)q(\alpha)d\alpha + \text{Const} = \underbrace{-\frac{1}{2}\sum_{i=1}^{n}(t_i - w^T x_i)^2 \mathbb{E}\beta - \frac{1}{2} w^t w \mathbb{E}\alpha}_{\text{Отрицательно определенная квадратичная форма по w}} + \text{Const} \Rightarrow \ q(w) \sim \mathcal{N}(w|\mu, S)$$

$$\log q(\alpha) = \int \log p(T,w,\alpha,\beta|X)q(w)q(\beta)dwd\beta + \text{Const} = \int \log p(w|\alpha)q(w)dw + \log p(\alpha) + \text{Const} =$$

$$= \underbrace{(a_0 - 1)\log\alpha - b_0\alpha + \frac{d}{2}\log\alpha - \frac{\alpha}{2}\mathbb{E}w^T w}_{\text{По альфа отрицательная линейная функция плюс логарифм}} + \text{Const} \Rightarrow \ q(\alpha) \sim \mathcal{G}\left(\alpha | a_0 + \frac{d}{2}, b_0 + \mathbb{E}w^T w\right)$$

$$\log q(\beta) = \int q(w)q(\alpha)\log p(T,w,\alpha,\beta|X)dwd\alpha + \text{Const} = \int q(w)\log p(T|w,X,\beta)dw + \log p(\beta) +$$

$$+ \text{Const} = (c_0 - 1)\log\beta - d_0\beta - \frac{\beta}{2}\sum_{i=1}^{n}\mathbb{E}(t_i - w^T x_i)^2 + \frac{n}{2}\log\beta + \text{Const} \Rightarrow$$

$$\Rightarrow \ q(\beta) \sim \mathcal{G}\left(\beta | c_0 + \frac{n}{2}, d_0 + \frac{1}{2}\sum_{i=1}^{n} t_i t_i - \mathbb{E}w^T \sum_{i=1}^{n} t_i x_i + \frac{1}{2}\text{tr}\left(\mathbb{E}ww^T \sum_{i=1}^{n} x_i x_i^T\right)\right)$$

Для итерационных формул пересчета нам необходимо знать $\mathbb{E}w, \ \mathbb{E}ww^T, \ \mathbb{E}\beta, \ \mathbb{E}\alpha$, но они легко рассчитываются как достаточные статистики соответствующих распределений $q(.)$

Замечание 4. Вариационная линейная регрессия практически не переобучается и отлична для малых выборок при использовании априорных распределений Джеффриса на альфа и бета.
Замечание 5. Вариационная нижняя оценка $\mathcal{L}(q,\theta)$ может быть использована для выбора Наиболее обоснованной модели, т.к. является нижней оценкой обоснованности.