Ликбез

Вариационный метод

Expectation Propagation

Лекция 9. Приближенные способы вывода. Вариационный подход. Expectation Propagation

Д. П. Ветров[1] Д. А. Кропотов[2]

[1]МГУ, ВМиК, каф. ММП

[2]ВЦ РАН

Спецкурс «Байесовские методы машинного обучения», 2009

План лекции

Ликбез

Вариационный метод

- Идея метода
- Вариационная линейная регрессия

Expectation Propagation

- Идея метода
- Пример использования

Дивергенция Кульбака-Лейблера

- Существует множество способов определить близость между вероятностными распределениями
- Рассмотрим распределения $p(\boldsymbol{x})$ и $q(\boldsymbol{x})$. Дивергенцией Кульбака-Лейблера называется величина

$$KL(q||p) = -\int q(\boldsymbol{x}) \log \frac{p(\boldsymbol{x})}{q(\boldsymbol{x})} d\boldsymbol{x}$$

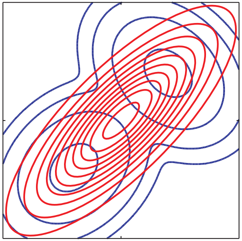
- Заметим, что дивергенция несимметрична

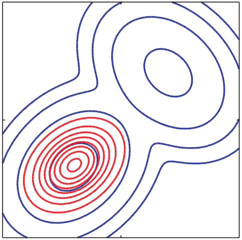
$$KL(q||p) \neq KL(p||q)$$

- Минимизация дивергенции Кульбака-Лейблера часто используется для приближения сложного распределения $p(\boldsymbol{x})$ более простым распределением $q(\boldsymbol{x})$

Геометрический смысл

Пусть $p(\boldsymbol{x})$ — «синее» распределение, а $q(\boldsymbol{x})$ — «красное». Слева показан результат минимизации $KL(p||q)$ по $q(\boldsymbol{x})$, а справа — результат минимизации $KL(q||p)$ по $q(\boldsymbol{x})$





Свойства дивергенции Кульбака-Лейблера

- Неотрицательность: $KL(p||q) \geq 0$ для любых двух распределений
- Дивергенция равна нулю тогда и только тогда, когда $q(\boldsymbol{x}) = p(\boldsymbol{x})$
- Антисимметричность: $KL(p||q) \neq KL(q||p)$

Гамма-распределение

- Гамма-распределение имеет плотность

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

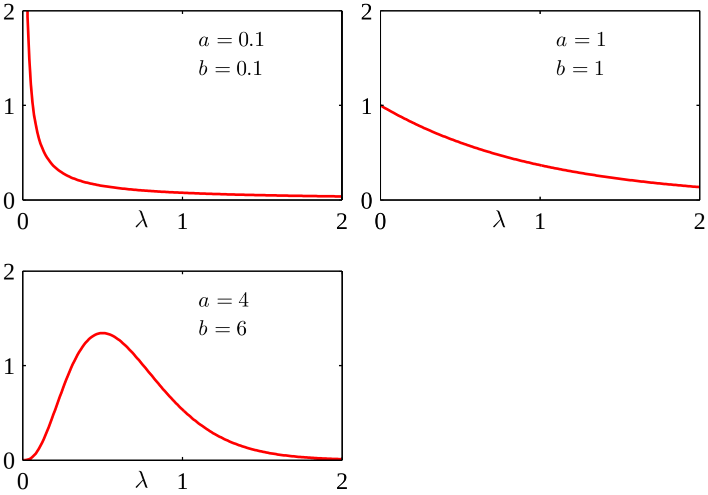
- Характеристики гамма-распределения

$$\mathbb{E}\lambda = \frac{a}{b}, \quad \mathbb{D}\lambda = \frac{a}{b^2}$$

- Гамма-распределение является сопряженным для обратной дисперсии (точности) нормального распределения $\lambda = \sigma^{-2}$, т.к.

$$\mathcal{N}(x|\mu,\sigma^2) = \mathcal{N}(x|\mu,\lambda^{-1}) = \sqrt{\frac{\lambda}{2\pi}} \exp\left(-\frac{\lambda}{2}(x-\mu)^2\right)$$

График гамма-распределения



План лекции

Ликбез

Вариационный метод

Идея метода

Вариационная линейная регрессия

Expectation Propagation

Идея метода

Пример использования

Недостатки приближения Лапласа

Ликбез

Вариационный метод

Идея метода

Вариационная линейная регрессия

Expectation Propagation

- Метод Лапласа хорошо приближает распределение гауссианой в точке максимума, но плохо делает приближение в целом, если распределение сильно отличается от гауссианы
- В частности, математические ожидания и дисперсии распределения и его приближения Лапласа могут сильно отличаться
- Это приводит к сильным смещениям оценки обоснованности

Приближение апостериорного распределения

- Используем общие обозначения, применявшиеся во второй лекции при описании EM-алгоритма. Пусть X — совокупность наблюдаемых переменных, а Z — множество настраиваемых параметров (ненаблюдаемых переменных)
- Вероятностная модель обычно позволяет в явном виде задать совместное распределение $p(X, Z)$. Целью задачи является нахождение (или приближение) обоснованности выбранной модели $p(X) = \int P(X, Z)dZ$ и апостериорного распределения

$$p(Z|X) = \frac{p(X, Z)}{p(X)}$$

- На практике прямое интегрирование выражения $p(X, Z)$ обычно невозможно, поэтому ограничиваются приближением распределения $p(Z|X)$ с помощью некоторого распределения $q(Z)$

Разложение обоснованности

- Справедливо следующее преобразование

$$\log p(X) = \log p(X) \int q(Z)dZ = \int \log p(X)q(Z)dZ =$$
$$\int \log \frac{p(X,Z)}{p(Z|X)} q(Z)dZ = \int \log \frac{p(X,Z)q(Z)}{q(Z)p(Z|X)} q(Z)dZ =$$
$$\int \log \frac{p(X,Z)}{q(Z)} q(Z)dZ - \int \log \frac{p(Z|X)}{q(Z)} q(Z)dZ = \mathcal{L}(q) + KL(q||p)$$

- Величина $\mathcal{L}(q)$ представляет собой нижнюю границу логарифма обоснованности
- Так как $\log p(X)$ не зависит от $q(Z)$, максимизация $\mathcal{L}(q)$ эквивалентна минимизации дивергенции Кульбака-Лейблера $KL(q||p)$ между $q(Z)$ и апостериорным распределением $p(Z|X)$!

Факторизация $q(Z)$

- Очевидно, что максимум $\mathcal{L}(q)$ достигается при $q(Z) = p(Z|X)$. В этом случае второе слагаемое оказывается равным нулю
- Прямое вычисление $p(Z|X)$ обычно невозможно, поэтому необходимо ограничить множество $\{q(Z)\}$, в котором проводится поиск наилучшего приближения, например, классом нормальных распределений, и свести задачу к оптимизации соответствующих параметров
- Альтернативой параметрическому ограничению семейства $\{q(Z)\}$ служит его факторизация

$$q(Z) = \prod_{i=1}^{k} q_i(z_i)$$

Факторизованное приближение

- Подставим $q(Z) = \prod_{i=1}^k q_i(z_i) = \prod_{i=1}^k q_i$ в выражение для $\mathcal{L}(q)$

$$\mathcal{L}(q) = \int \prod_i q_i \left(\log p(X, Z) - \sum_i \log q_i \right) dZ =$$

$$\int q_j \left(\int \log p(X, Z) \prod_{i \neq j} q_i dz_i \right) dz_j - \int q_j \log q_j dz_j + C$$

- Обозначим $\log \tilde{p}(X, z_j) = \mathbb{E}_{i \neq j} \log p(X, Z) = \int \log p(X, Z) \prod_{i \neq j} q_i dz_i$. Тогда

$$\mathcal{L}(q) = \int q_j \log \frac{\tilde{p}(X, z_j)}{q_j} dz_j + C = -KL(q || \tilde{p}) + C$$

Основной результат

- Максимизация $\mathcal{L}(q)$ по q_j эквивалентна минимизации дивергенции между $q_j(z_j)$ и $\tilde{p}(X, z_j)$
- Отсюда оптимальное распределение $q_j^*(z_j) = \tilde{p}(X, z_j)$, т.е.

$$\log q_j^*(z_j) = \mathbb{E}_{i \neq j} \log p(X, Z) + C$$

- Заметим, что нам не пришлось делать каких-либо предположений о функциональной форме распределения $q_j(z_j)$
- Выражение для оптимального $q_j^*(z_j)$ зависит от остальных $q_i(z_i)$, поэтому необходима итерационная оптимизация

План лекции

Лекция 9. Приближенные способы вывода. Вариационный подход. Expectation Propagation

Ветров, Кропотов

Ликбез

Вариационный метод

Идея метода

Вариационная линейная регрессия

Expectation Propagation

Ликбез

Вариационный метод

- Идея метода
- Вариационная линейная регрессия

Expectation Propagation

- Идея метода
- Пример использования

Вероятностная модель линейной регрессии

- Рассмотрим стандартную задачу восстановления регрессии $(X, \boldsymbol{t})$ — обучающая выборка, $t \in \mathbb{R}$. Регрессия имеет вид $y(\boldsymbol{x}) = \sum_{j=1}^{m} w_j \phi_j(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x})$
- Определим следующую вероятностную модель $p(\boldsymbol{t}, \boldsymbol{w}, \alpha) = p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w}|\alpha)p(\alpha)$, где

$$p(\boldsymbol{t}|\boldsymbol{w}) = \prod_{i=1}^{n} \mathcal{N}(t_i|\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i), \beta^{-1}) \qquad p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1}I)$$

$$p(\alpha) = \mathcal{G}(\alpha|a_0, b_0)$$

- В данной модели роль наблюдаемых переменных играет $\boldsymbol{t}$, а в роли Z выступают $\boldsymbol{w}$ и α
- Для простоты предположим, что значение интенсивности белого шума β известно

Вариационный вывод для α

- Будем искать приближение распредедения $p(\boldsymbol{w}, \alpha|\boldsymbol{t})$ в виде
$$q(\boldsymbol{w}, \alpha) = q(\boldsymbol{w})q(\alpha)$$
- Используя основной результат для $q(\alpha)$ получаем
$$\log q^*(\alpha) = \mathbb{E}_{\boldsymbol{w}} \log p(\boldsymbol{t}, \boldsymbol{w}, \alpha) = \mathbb{E}_{\boldsymbol{w}} \left(\log p(\boldsymbol{w}|\alpha)p(\alpha)\right) + C =$$
$$\mathbb{E}_{\boldsymbol{w}} \log p(\boldsymbol{w}|\alpha) + \log p(\alpha) + C =$$
$$\frac{m}{2}\log\alpha - \frac{\alpha}{2}\mathbb{E}\boldsymbol{w}^T\boldsymbol{w} + (a_0 - 1)\log\alpha - b_0\alpha + C_1$$
- Но это в точности логарифм гамма-распределения с параметрами a_n и b_n, т.е. $\alpha \sim \mathcal{G}(\alpha|a_n, b_n)$, причем
$$a_n = a_0 + \frac{m}{2}, \qquad b_n = b_0 + \frac{1}{2}\mathbb{E}\boldsymbol{w}^T\boldsymbol{w}$$

Вариационный вывод для $\boldsymbol{w}$

- Проделаем аналогичную операцию для $q(\boldsymbol{w})$

$$\log q^*(\boldsymbol{w}) = \mathbb{E}_\alpha \log p(\boldsymbol{t}, \boldsymbol{w}, \alpha) = \mathbb{E}_\alpha \log \left(p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w}|\alpha)p(\alpha)\right) =$$
$$\log p(\boldsymbol{t}|\boldsymbol{w}) + \mathbb{E}_\alpha \log p(\boldsymbol{w}|\alpha) + C =$$
$$-\frac{\beta}{2}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i) - t_i)^2 - \frac{1}{2}\mathbb{E}\alpha \cdot \boldsymbol{w}^T\boldsymbol{w} + C_1 =$$
$$-\frac{1}{2}\boldsymbol{w}^T(\mathbb{E}\alpha \boldsymbol{I} + \beta\Phi^T\Phi)\boldsymbol{w} + \beta\boldsymbol{w}^T\Phi^T\boldsymbol{t} + C_2,$$

где $\Phi = (\boldsymbol{\phi}(\boldsymbol{x}_1), \dots, \boldsymbol{\phi}(\boldsymbol{x}_n))$

- Последовательно проведено отбрасывание слагаемых, не зависящих от $\boldsymbol{w}$, раскрытие скобок и приведение подобных слагаемых
- Выделяя полный квадрат, получаем, что $\boldsymbol{w} \sim \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_n, S_n)$, где

$$\boldsymbol{\mu}_n = \beta S_n \Phi^T \boldsymbol{t}, \qquad S_n = (\mathbb{E}\alpha \boldsymbol{I} + \beta\Phi^T\Phi)^{-1}$$

Итерационные формулы

- Окончательные формулы: $q^*(\alpha) = \mathcal{G}(\alpha|a_n, b_n)$, $q^*(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}_n, S_n)$, т.е.

$$\mathbb{E}\alpha = \frac{a_n}{b_n}$$

$$\mathbb{E}\boldsymbol{w}^T\boldsymbol{w} = \mathrm{tr}(\mathbb{E}\boldsymbol{w}\boldsymbol{w}^T) = \mathrm{tr}(\boldsymbol{\mu}_n\boldsymbol{\mu}_n^T + S_n) = \boldsymbol{\mu}_n^T\boldsymbol{\mu}_n + \mathrm{tr}\, S_n$$

- Параметры распределений определяются по итерационным формулам

$$a_n = a_0 + \frac{m}{2}$$

$$b_n = b_0 + \mathbb{E}\boldsymbol{w}^T\boldsymbol{w} = b_0 + \boldsymbol{\mu}_n^T\boldsymbol{\mu}_n + \mathrm{tr}\, S_n$$

$$\boldsymbol{\mu}_n = \beta S_n \Phi^T \boldsymbol{t}$$

$$S_n = \left(\mathbb{E}\alpha I + \beta\Phi^T\Phi\right)^{-1} = \left(\frac{a_n}{b_n} I + \beta\Phi^T\Phi\right)^{-1}$$

Заключительные замечания

- Отметим, что никаких ограничений на форму апостериорных распределений не вводилось, а единственным приближением было предположение о факторизации
- Вариационный метод позволяет получать приближение обоснованности, нижней оценкой которой является выражение

$$\mathcal{L}(q) = \int q(\boldsymbol{w}, \alpha) \log \frac{p(\boldsymbol{t}, \boldsymbol{w}, \alpha)}{q(\boldsymbol{w}, \alpha)} d\boldsymbol{w} d\alpha =$$

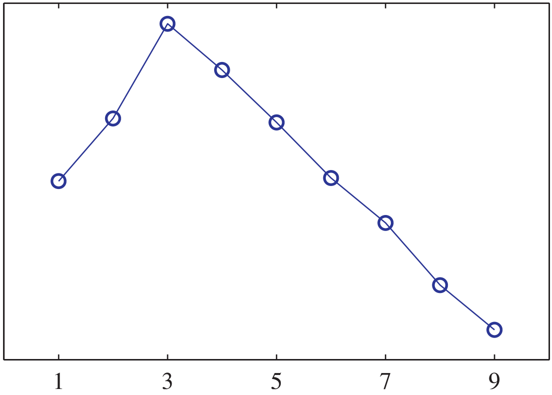
$$\mathbb{E} \log p(\boldsymbol{t}, \boldsymbol{w}, \alpha) - \mathbb{E} \log q(\boldsymbol{w}, \alpha) =$$

$$\mathbb{E}_{\boldsymbol{w}} \log p(\boldsymbol{t}|\boldsymbol{w}) + \mathbb{E}_{\boldsymbol{w},\alpha} \log p(\boldsymbol{w}|\alpha) + \mathbb{E}_{\alpha} \log p(\alpha) - \mathbb{E}_{\boldsymbol{w}} \log q(\boldsymbol{w}) - \mathbb{E}_{\alpha} q(\alpha)$$

- Все эти выражения выписываются в явном виде (Упр.)

Вариационный выбор модели

На рисунке изображена зависимость $\mathcal{L}(q)$ от степени полинома для полиномиальной регрессии, построенной по зашумленной выборке, полученной с помощью кубического многочлена



План лекции

Ликбез

Вариационный метод

- Идея метода
- Вариационная линейная регрессия

Expectation Propagation

- Идея метода
- Пример использования

Альтернативный вариационный подход

- Выше было показано, что вариационный подход эквивалентен минимизации дивергенции Кульбака-Лейблера $KL(q||p)$ по q
- Можно рассмотреть и противоположную ситуацию — минимизацию $KL(p||q)$ по q
- Получаемое приблидение будет обладать иными свойствами
- Такой подход получил называние Expectation Propagation

Факторизованное приближение

- Как и в предыдущем разделе рассмотрим приблидження распределения $p(Z|X)$ факторизованным распределением $q(Z) = \prod_{i=1}^k q_i(z_i) = \prod_{i=1}^k q_i$
- Подставим факторизованное приблидение в выражение для $KL(p||q)$

$$KL(p||q) = -\int p(Z|X) \frac{\log q(Z)}{\log p(Z|X)} dZ = -\int p(Z|X) \sum_{i=1}^k \log q_i dZ + C =$$

$$= -\sum_{i=1}^k \int p(Z|X) \log q_i dZ = -\sum_{i=1}^k \int \left(p(Z|X) \prod_{j\neq i} dz_j \right) \log q_i dz_i$$

- Обозначим $\int p(Z|X) \prod_{j\neq i} dz_j = p(z_i)$. Тогда легко видеть, что минимум дивергенции по q_i достигается при q_i, равном маргинальному распределению

$$q_i^*(z_i) = p(z_i) = \int p(Z|X) \prod_{j\neq i} dz_j$$

Совмещение моментов

- При использовании факторизации минимум $KL(p||q)$ достигается при факторах, равных маргинальному распределению

$$q_i^*(z_i) = p(z_i) = \int p(Z|X) \prod_{j \neq i} dz_j$$

- Во многих случаях аналитический подсчет маргинального распределения невозможен
- Если дополнительно ограничить семейство допустимых факторов нормальными распределениями, достаточно приравнять первые два момента (свойство дивергенции $KL(p||q)$): (Упр.)

$$q_i^*(z_i) \sim \mathcal{N}(z_i|\mu, \sigma^2)$$

$$\mu = \mathbb{E}z_i = \int z_i p(z_i) dz_i, \quad \sigma^2 = \mathbb{D}z_i = \int (z_i - \mu)^2 p(z_i) dz_i$$

- Эти интегралы можно взять численно

Идея Expectation propagation

- В основе метода лежит очень простая идея, последовательного приближения отдельных множителей совместного распределения факторизованными распределениями
- Пусть совместное распределение представимо в виде произведения $p(Z) = \prod_{i=1}^{s} f_i(Z_i)$
- Будем приближать каждый его множитель $f_i(Z_i)$ факторизованным приближением $q_i(Z_i) = \prod_{j=1}^{t_s} q_{ij}(z_{ij})$ Множители $f_i(Z_i)$, вообще говоря, могут не быть распределениями относительно всех своих аргументов
- Приближение осуществляется в контексте: вместо минимизации $KL(p(Z)||\prod_{i,j} q_{ij}(z_{ij}))$ по всем q_{ij} минимизируется локальная дивергенция

$$KL\left(f_k(Z_k)\prod_{i\neq k}\prod_{j=1}^{t_i} q_{ij}(z_{ij}) \middle\| \prod_{j=1}^{t_k} q_{kj}(z_{kj})\prod_{i\neq k}\prod_{j=1}^{t_i} q_{ij}(z_{ij})\right) \to \min_{q_{k1},\dots,q_{kt_k}}$$

Выражения для пересчета q_{kj}

- Минимум выражения

$$KL\left(f_k(Z_k)\prod_{i\neq k}\prod_{j=1}^{t_i}q_{ij}(z_{ij})\,\middle\|\,\prod_{j=1}^{t_k}q_{kj}(z_{kj})\prod_{i\neq k}\prod_{j=1}^{t_i}q_{ij}(z_{ij})\right)\to\min_{q_{k1},\dots,q_{kt_k}}$$

достигается при $q_{kj}(z_{kj})$ равным маргиналам

$$q^*_{kj}(z_{kj})\propto\mathbb{E}_{Z_k\ni z_{tr}\neq z_{kj}\in Z_k}f_k(Z_k)\prod_{i:z_{tr}\in Z_i}\prod_{s:z_{tr}=z_{is}}q_{is}(z_{is})$$

Множители при переменных, которые не входят в Z_k на q_{kj} влияния не оказывают

- Часто q_{kj} дополнительно ограничивают гауссианами, в этом случае вместо аналитического интегрирования достаточно подсчитать три момента: нормировочную константу, мат. ожидание и дисперсию

План лекции

Ликбез

Вариационный метод

Expectation Propagation
- Идея метода
- Пример использования

Ликбез

Вариационный метод
- Идея метода
- Вариационная линейная регрессия

Expectation Propagation
- Идея метода
- Пример использования

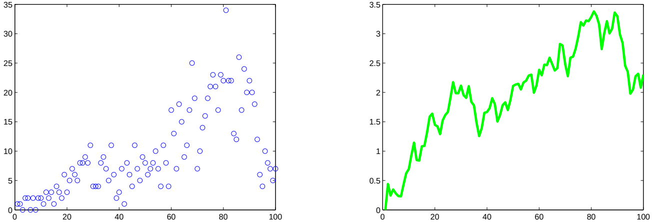
Пуассоновский трекинг

- Рассмотрим следующую задачу: в каждый момент времени мы наблюдаем случайную величину, распределенную по закону Пуассона

$$p(x_n = k) = \lambda_n^k \frac{e^{-\lambda_n}}{k!},$$

параметр $\lambda_n = \exp(z_n)$ которой зависит от ненаблюдаемой компоненты z_n

- z_n представляет собой процесс броуновского движения $z_n \sim \mathcal{N}(z_n|z_{n-1}, \sigma^2)$ с известной и постоянной σ^2



Байесовская сеть сигнала со скрытым процессом

Ликбез

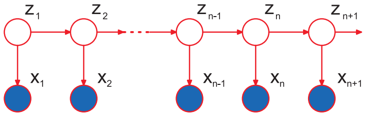
Вариационный метод

Expectation Propagation

Идея метода

Пример использования

- Данная задачи является (пусть и несколько нестандартной) задачей обработки сигнала
- Сигналы со скрытой переменной, определяющей взаимосвязь значений соседних наблюдений удобно представлять в виде байесовской сети



- Совместное распределение задается формулой

$$p(X,Z) = \prod_{n=2}^{N} p(x_n|z_n)p(z_n|z_{n-1})p(z_1) = \prod_{n=2}^{N} f_n(z_n, z_{n-1})g_n(x_n, z_n)g_1(x_1, z$$

Задача найти $Z^* = \arg\max p(Z|X)$

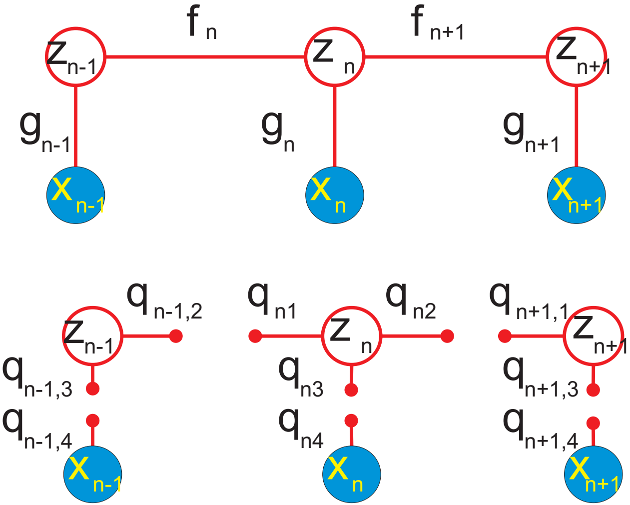
Использование EP

- Будем приближать $f_n(z_n, z_{n-1})$ произведением $q_{n1}(z_n)q_{n-1,2}(z_{n-1})$, а $g_n(x_n, z_n)$ — произведением $q_{n3}(z_n)q_{n4}(x_n)$
- Дополнительно потребуем, чтобы все $q_{nj}(z_n)$ были гауссианами
- Факторизованные приближения будем искать минимизируя соответствующую KL в контексте

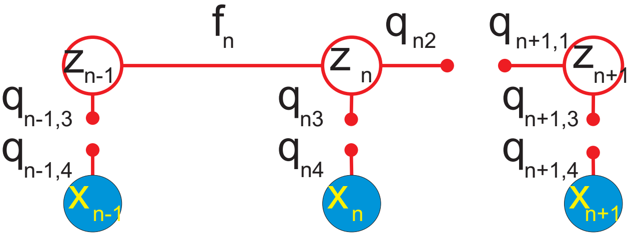
$$q_{k1}(z_k) = \arg\min KL\left(\prod_n q_{n1}(z_{n-1})q_{n-1,2}(z_{n-1})q_{n3}(z_n)\frac{f_k(z_k, z_{k-1})}{q_{k1}(z_k)q_{k-1,2}(z_{k-1})} \middle\| \prod_n q_{n1}(z_{n-1})q_{n-1,2}(z_{n-1})q_{n3}(z_n)\right)$$

$$q_{k3}(z_k) = \arg\min KL\left(\prod_n q_{n1}(z_{n-1})q_{n-1,2}(z_{n-1})q_{n3}(z_n)\frac{g_k(x_k, z_k)}{q_k 3(z_k)} \middle\| \prod_n q_{n1}(z_{n-1})q_{n-1,2}(z_{n-1})q_{n3}(z_n)\right)$$

Графическая иллюстрация факторизации



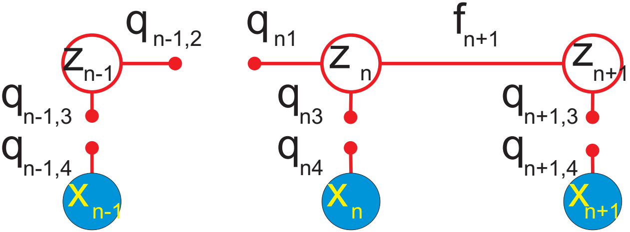
Формулы для пуассоновского трекинга



Учитывая структуру нашей байесовской сети получаем и отбрасывая множители, не зависящие от переменной интегрирования, получим

$$q_{k1}(z_k) \propto \mathbb{E}_{z_{k-1}} f_k(z_k, z_{k-1}) q_{k-1,1}(z_{k-1}) q_{k-1,3}(z_{k-1})$$

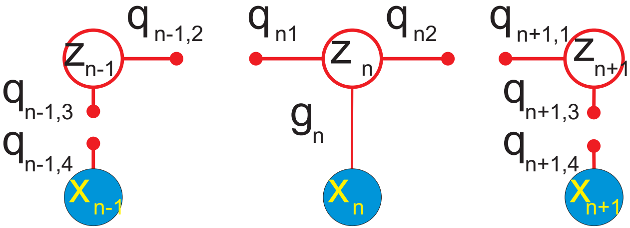
Формулы для пуассоновского трекинга



Учитывая структуру нашей байесовской сети получаем и отбрасывая множители, не зависящие от переменной интегрирования, получим

$$q_{k2}(z_k) \propto \mathbb{E}_{z_{k+1}} f_{k+1}(z_{k+1}, z_k) q_{k+1,2),1}(z_{k+1}) q_{k+1,3}(z_{k+1})$$

Формулы для пуассоновского трекинга

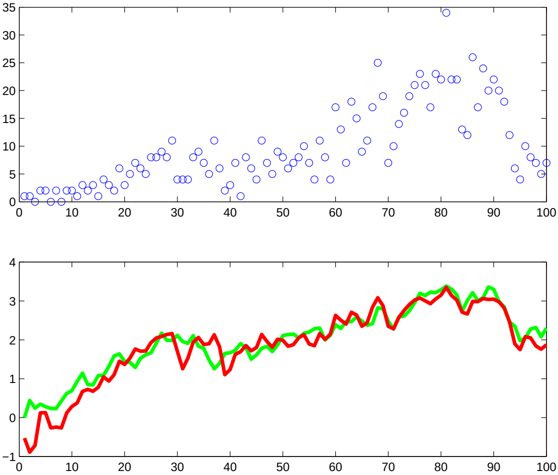


Учитывая структуру нашей байесовской сети получаем и отбрасывая множители, не зависящие от переменной интегрирования, получим

$$q_{k3}(z_k) \propto g_k(x_k, z_k)$$

В последнем случае, мат. ожидание не берется, т.к. значение x_k известно

Результат EP



Красным показана наиболее вероятная реализация скрытого процесса, полученная с помощью EP