Курс: Байесовские методы машинного обучения, 2011

Вариационный подход как приближенный способ байесовского вывода

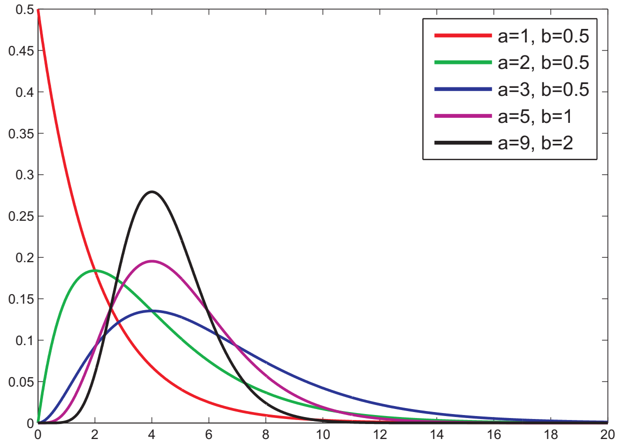
Дата: 9 ноября 2011

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. Часто гамма-распределение используются в качестве априорного распределения для параметра масштаба (например, параметра α в линейной и логистической регрессии). Гамма-распределение является сопряженным для параметра точности в нормальном распределении:

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

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.

Вывод в вероятностных моделях

Пусть имеется некоторая вероятностная модель, задаваемая совместным распределением $p(X,T,Z)$. Здесь X – известные переменные, T – оцениваемые переменные, Z – неизвестные переменные, которые не требуется оценивать. Тогда задача вывода в вероятностной модели соответствует вычислению апостериорного распределения

$$p(T|X) = \frac{p(T,X)}{p(X)} = \frac{\int p(X,T,Z)dZ}{\int p(X,\tilde{T},Z)d\tilde{T}dZ}.$$

В том случае, если нас интересуют точечные оценки, то тогда берется, как правило, мат.ожидание или мода апостериорного распределения:

$$\hat{T} = \arg\max_T p(T|X),$$
$$\hat{T} = \mathbb{E}[T|X].$$

Интегралы, возникающие при получении апостериорного распределения, часто не вычисляются аналитически. Следовательно, для осуществления байесовского вывода требуются приближенные методы. Первый класс методов приближенного байесовского вывода – это методы Монте Карло с марковскими цепями. Другой класс методов – вариационный подход.

Примеры вероятностных моделей

Линейная регрессия.

$$\begin{aligned}
&p(\boldsymbol{t}, \boldsymbol{w}, \alpha, \beta|X) = \prod_{n=1}^{N} p(t_n|\boldsymbol{w}, \beta, \boldsymbol{x}_n) p(\boldsymbol{w}|\alpha) p(\alpha) p(\beta),\\
&p(t_n|\boldsymbol{w}, \beta, \boldsymbol{x}_n) = \mathcal{N}\left(t_n \middle| \sum_{j=1}^{M} w_j \phi_j(\boldsymbol{x}_n), \beta^{-1}\right),\\
&p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1} I),\\
&p(\alpha) = \mathcal{G}(\alpha|a_0, b_0),\\
&p(\beta) = \mathcal{G}(\beta|c_0, d_0).
\end{aligned}$$

Задача вывода (X – $(\boldsymbol{t}, X, \boldsymbol{x}_{new})$, Z – $(\boldsymbol{w}, \beta, \alpha)$, T – t_{new}):

$$p(t_{new}|\boldsymbol{x}_{new}, \boldsymbol{t}, X) = \int p(t_{new}|\boldsymbol{w}, \beta, \boldsymbol{x}_{new}) p(\boldsymbol{w}, \beta, \alpha|\boldsymbol{t}, X) d\boldsymbol{w} d\beta d\alpha.$$

Логистическая регрессия.

$$\begin{aligned}
&p(\boldsymbol{t}, \boldsymbol{w}, \alpha|X) = \prod_{n=1}^{N} p(t_n|\boldsymbol{w}, \boldsymbol{x}_n) p(\boldsymbol{w}|\alpha) p(\alpha),\\
&p(t_n|\boldsymbol{w}, \boldsymbol{x}_n) = \frac{1}{1 + \exp(-t_n \sum_{j=1}^{M} w_j \phi_j(\boldsymbol{x}_n))},\\
&p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{0}, \alpha^{-1} I),\\
&p(\alpha) = \mathcal{G}(\alpha|a_0, b_0).
\end{aligned}$$

Задача вывода (X – $(\boldsymbol{t}, X, \boldsymbol{x}_{new})$, Z – $(\boldsymbol{w}, \alpha)$, T – t_{new}):

$$p(t_{new}|\boldsymbol{x}_{new}, \boldsymbol{t}, X) = \int p(t_{new}|\boldsymbol{w}, \boldsymbol{x}_{new}) p(\boldsymbol{w}, \alpha|\boldsymbol{t}, X) d\boldsymbol{w} d\alpha.$$

Смесь нормальных распределений.

$$\begin{aligned}
&p(X, T, \boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) = \prod_{n=1}^{N} p(\boldsymbol{x}_n|\boldsymbol{t}_n, \boldsymbol{\mu}, \Sigma) p(\boldsymbol{t}_n|\boldsymbol{\pi}) p(\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma),\\
&p(\boldsymbol{x}_n|\boldsymbol{t}_n, \boldsymbol{\mu}, \Sigma) = \prod_{k=1}^{K} [\mathcal{N}(\boldsymbol{x}_n|\boldsymbol{\mu}_k, \Sigma_k)]^{t_{nk}},\\
&p(\boldsymbol{t}_n|\boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{t_{nk}},\\
&p(\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) = \text{const}.
\end{aligned}$$

Задача вывода (X – X, Z – T, T – $(\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma)$):

$$p(\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma|X) \propto p(X|\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma) \to \max_{\boldsymbol{\pi}, \boldsymbol{\mu}, \Sigma}.$$

Нижняя оценка обоснованности

Пусть имеется вероятностная модель $p(X,T)$ и некоторое произвольное распределение $q(T)$. Тогда:

$$\log p(X) = \underbrace{\int \log \frac{p(X,T)}{q(T)} q(T) dT}_{\mathcal{L}(q)} \underbrace{- \int \log \frac{p(T|X)}{q(T)} q(T) dT}_{\mathrm{KL}(q||p(T|X))}.$$

На это равенство можно смотреть с нескольких точек зрения. Допустим, что апостериорное распределение $p(T|X)$ не поддается вычислению (не вычисляется нормировочная константа) и мы хотим найти приближение $q(T)$ для распределения $p(T|X)$. Будем искать это приближение путем минимизации КЛ-дивергенции между распределением $q(T)$ и $p(T|X)$ в некотором семействе распределений $q(T)$. Тогда из равенства выше следует, что

$$\mathrm{KL}(q||p(T|X)) \to \min_q \Leftrightarrow \mathcal{L}(q) \to \max_q.$$

Теперь задача минимизации, которая зависит от недоступного для вычисления распределения $p(T|X)$, сведена к задаче максимизации функционала $\mathcal{L}(q)$, который зависит от известного полного совместного распределения модели $p(X,T)$.

С другой точки зрения, нас может интересовать значение маргинального распределения $p(X)$ (нормировочной константы для распределения $p(T|X)$). Так как $\mathrm{KL}(q||p(T|X)) \geq 0$, то значение функционала $\mathcal{L}(q)$ является нижней границей для $\log p(X)$. Таким образом, решая задачу максимизации функционала $\mathcal{L}(q)$ в некотором семестве распределений $q(T)$, мы одновременно получаем аналитическое приближение апостериорного распределения $p(T|X)$ и нижнюю границу для обоснованности $\log p(X)$.

Рассмотренные выше рассуждения справедливы также и для случая, когда требуется оценить произвольное распределение, известное с точностью до константы. Пусть имеется некоторое распределение

$$p(T) = \frac{1}{Z}\tilde{p}(T),$$

в котором мы умеем вычислять $\tilde{p}(T)$ для произвольного T, но нормировочная константа Z является недоступной. Тогда максимизация функционала

$$\mathcal{L}(q) = \int \log \frac{\tilde{p}(T)}{q(T)} q(T) dT \to \max_q$$

позволяет найти приближение $q(T)$ для распределения $p(T)$, а также оценить значение нормировочной константы $\log Z \geq \mathcal{L}(q)$.

Минимизация в семействе факторизованных распределений

Рассмотрим задачу приближения апостериорного распределения $p(T|X)$ в семействе т.н. факторизованных распределений:

$$q(T) = \prod_{j=1}^{J} q_j(T_j).$$

Здесь множество переменных T разбито на непересекающиеся подмножества T_j, причем $\cup_j T_j = T$, а $q_j(T_j)$ – произвольное распределение в пространстве переменных T_j. Таким образом, мы приходим к следующей задаче оптимизации:

$$\mathcal{L}(q) = \int \log \frac{p(X,T)}{\prod_j q_j(T_j)} \prod_j q_j(T_j) dT_j \to \max_{q_1,\dots,q_J}.$$

Рассмотрим решение этой задачи оптимизации с помощью покоординатного подъема, т.е. зафиксируем все компоненты распределения q, кроме q_i, и рассмотрим оптимизацию $\mathcal{L}(q)$ по отдельной компоненте $q_i(T_i)$. Оказывается, что решение такой (вариационной) задачи оптимизации можно получить аналитически:

$$\mathcal{L}(q) = \int \log p(X,T) \prod_j q_j(T_j) dT_j - \sum_{j=1}^{J} \int \log q_j(T_j) q_j(T_j) dT_j =$$
$$\int \left(\int \log p(X,T) \prod_{j \neq i} q_j(T_j) dT_j \right) q_i(T_i) dT_i - \int \log q_i(T_i) q(T_i) dT_i + \text{const}.$$

Рассмотрим следующее распределение $r(T_i)$:

$$r(T_i) = \frac{1}{Z} \exp\left(\int \log p(X,T) \prod_{j \neq i} q_j(T_j) dT_j \right),$$

где Z – нормировочная константа распределения, не зависящая от $q_i(T_i)$. Тогда $\log r(T_i) + \log Z = \int \log p(X,T) \prod_{j \neq i} q_j(T_j) dT_j$. Подставляя этот результат в выражение для $\mathcal{L}(q)$, получаем:

$$\mathcal{L}(q) = \int \log r(T_i) q_i(T_i) dT_i - \int \log q_i(T_i) q_i(T_i) dT_i + \text{const} =$$
$$\int \log \frac{r(T_i)}{q_i(T_i)} q_i(T_i) dT_i + \text{const} = -\text{KL}(q_i || r) + \text{const}.$$

Таким образом, максимизация $\mathcal{L}(q)$ в данном случае эквивалентна минимизации $\text{KL}(q_i||r)$ по распределению $q_i(T_i)$. Однако, известно, что минимум КЛ-дивергенции достигается в случае тождественных распределений, т.е.

$$q_i(T_i) = \frac{\exp\left(\int \log p(X,T) \prod_{j \neq i} q_j(T_j) dT_j \right)}{\int \exp\left(\int \log p(X,T) \prod_{j \neq i} q_j(T_j) dT_j \right) dT_i}. \tag{1}$$

Последняя формула является основным результатом вариационного подхода. Заметим, что оптимальное распределение $q_i(T_i)$ зависит от всех остальных распределений $q_j(T_j)$ для $j \neq i$. Поэтому при применении вариационного подхода возникает итерационная схема, в которой последовательно пересчитываются отдельные компоненты факторизованного распределения $q(T)$. При этом на каждом шаге итерации происходит монотонное увеличение нижней границы $\mathcal{L}(q)$ для обоснованности $\log p(X)$, и итерационная оптимизация происходит до сходимости по значению $\mathcal{L}(q)$. Заметим также, что в построениях выше не накладывалось никаких ограничений на семейство распределений $q(T)$, кроме факторизации.

В вариационном подходе требуется уметь усреднять логарифм совместного распределения $\log p(X,T)$ по всем компонентам $q(T)$, кроме одного, $\exp\left(\int \log p(X,T) \prod_{j \neq i} q_j(T_j) dT_j \right)$, вычислять нормировочную константу для очередного распределения $\int \exp\left(\int \log p(X,T) \prod_{j \neq i} q_j(T_j) dT_j \right) dT_i$, а также нижнюю границу $\mathcal{L}(q) = \int \log p(X,T) q(T) dT - \int \log q(T) q(T) dT$. Заметим, что все эти величины требуют интегрирования по пространству переменных T. Здесь может создаться впечатление, что вариационный подход никак не облегчает исходную задачу поиска $p(X)$, т.к. $p(X)$ также представляет собой схожий интеграл по пространству T $\int p(X,T) dT$. Тем не менее, во многих реальных вероятностных моделях вариационный подход действительно позволяет решить поставленную задачу. Это связано с тем, что в вариационном подходе происходит интегрирование логарифма совместного распределения $\log p(X,T)$, а не самого исходного распределения $p(X,T)$. Кроме того, интегрирование также облегчает предположение о факторизации $q(T)$, т.е. интеграл часто удается разбить на произведение интегралов от подмножеств переменных T. Конкретные примеры применения вариационного подхода будут рассмотрены ниже.

Заметим, что в общей задаче байесовского вывода (см. пункт 2) мы имеем дело с тремя группами переменных X, T, Z, где X – наблюдаемые переменные, T – переменные, подлежащие оцениванию, и Z – остальные ненаблюдаемые переменные. При применении вариационного подхода происходит поиск приближения апостериорного распределения $p(T,Z|X)$ в семействе факторизованных распределений $q(T,Z)=q_T(T)q_Z(Z)$. Затем

$$p(T|X)=\int p(T,Z|X)dZ\simeq\int q_T(T)q_Z(Z)dZ=q_T(T).$$

Пример применения вариационного подхода для модели линейной регрессии

Рассмотрим задачу регрессии. Пусть имеется некоторая выборка $(\boldsymbol{t},X)=\{t_n,\boldsymbol{x}_n\}_{n=1}^N$, состоящая из N объектов, где каждый объект представлен своим вектором признаков $\boldsymbol{x}_n\in\mathbb{R}^d$ и значением регрессионной переменной $t_n\in\mathbb{R}$. Задача состоит в прогнозе регрессионной компоненты t_{new} для объекта, представленного только своим вектором признаков $\boldsymbol{x}_{new}$.

Для решения этой задачи воспользуемся байесовской моделью линейной регрессии:

$$\begin{aligned}
&p(\boldsymbol{t},t_{new},\boldsymbol{w},\alpha,\beta|X,\boldsymbol{x}_{new})=\prod_{n=1}^N p(t_n|\boldsymbol{w},\beta,\boldsymbol{x}_n)p(t_{new}|\boldsymbol{w},\beta,\boldsymbol{x}_{new})p(\boldsymbol{w}|\alpha)p(\alpha)p(\beta),\\
&p(t_n|\boldsymbol{w},\sigma,\boldsymbol{x}_n)=\mathcal{N}\left(t_n\middle|\sum_{j=1}^M w_j\phi_j(\boldsymbol{x}_n),\beta^{-1}\right),\\
&p(\boldsymbol{w}|\alpha)=\mathcal{N}(\boldsymbol{w}|\boldsymbol{0},\alpha^{-1}I),\\
&p(\alpha)=\mathcal{G}(\alpha|a_0,b_0),\\
&p(\beta)=\mathcal{G}(\beta|c_0,d_0).
\end{aligned}$$

С помощью вариационного подхода будет искать приближение для апостериорного распределения $p(\boldsymbol{w},\alpha,\beta|\boldsymbol{t},X)$ в семействе факторизованных распределений $q_{\boldsymbol{w}}(\boldsymbol{w})q_\alpha(\alpha)q_\beta(\beta)$. Для этого воспользуемся общим результатом (1) и рассмотрим применение этой формулы для каждой компоненты распределения $q_{\boldsymbol{w}}(\boldsymbol{w})$, $q_\alpha(\alpha)$ и $q_\beta(\beta)$.

Компонента $q_{\boldsymbol{w}}(\boldsymbol{w})$.

$$\begin{aligned}
\log q_{\boldsymbol{w}}(\boldsymbol{w})=\int\log p(\boldsymbol{t},\boldsymbol{w},\alpha,\beta|X)q_\alpha(\alpha)q_\beta(\beta)d\alpha d\beta+\text{const}=\\
=\underbrace{\frac{N}{2}(\log\beta-\log 2\pi)}_{\text{не зависит от }\boldsymbol{w}}-\frac{\mathbb{E}\beta}{2}\sum_{n=1}^N(t_n-\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n))^2+\underbrace{\frac{M}{2}(\log\alpha-\log 2\pi)}_{\text{не зависит от }\boldsymbol{w}}-\frac{\mathbb{E}\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}+\\
+\underbrace{\log\mathcal{G}(\alpha|a_0,b_0)+\log\mathcal{G}(\beta|c_0,d_0)}_{\text{не зависит от }\boldsymbol{w}}+\text{const}=-\frac{\mathbb{E}\beta}{2}\sum_{n=1}^N(t_n-\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n))^2-\frac{\mathbb{E}\alpha}{2}\boldsymbol{w}^T\boldsymbol{w}+\text{const}.
\end{aligned}\quad(2)$$

Здесь под $\mathbb{E}\beta$ понимается мат. ожидание по распределению $q_\beta(\beta)$, а под $\mathbb{E}\alpha$ – мат.ожидание по распределению $q_\alpha(\alpha)$. Выражение (2) как функция от $\boldsymbol{w}$ представляет собой квадратичную функцию. Следовательно, распределение $q_{\boldsymbol{w}}(\boldsymbol{w})$ является нормальным распределением со следующими параметрами:

$$\Sigma=\left(\operatorname{diag}(\mathbb{E}\alpha)+\mathbb{E}\beta\sum_n\boldsymbol{\phi}(\boldsymbol{x}_n)\boldsymbol{\phi}(\boldsymbol{x}_n)^T\right)^{-1},\quad(3)$$

$$\boldsymbol{\mu}=\Sigma\left(\sum_n\mathbb{E}\beta t_n\boldsymbol{\phi}(\boldsymbol{x}_n)\right).\quad(4)$$

Компонента $q_{\boldsymbol{\alpha}}(\boldsymbol{\alpha})$.

$$\begin{aligned}
\log q_\alpha(\alpha)=\int\log p(\boldsymbol{t},\boldsymbol{w},\alpha,\beta|X)q_{\boldsymbol{w}}(\boldsymbol{w})q_\beta(\beta)d\boldsymbol{w}d\beta+\text{const}=\\
=\frac{M}{2}\log\alpha+\frac{\alpha}{2}\mathbb{E}\boldsymbol{w}^T\boldsymbol{w}+(a_0-1)\log\alpha-b_0\alpha+\text{const}.
\end{aligned}$$

В этом выражении легко узнается логарифм гамма-распределения с параметрами:

$$a = a_0 + \frac{M}{2}, \tag{5}$$
$$b = b_0 + \mathbb{E}\boldsymbol{w}^T\boldsymbol{w}. \tag{6}$$

Компонента $q_{\boldsymbol{\beta}}(\boldsymbol{\beta})$.

$$\log q_\beta(\beta) = \int \log p(\boldsymbol{t}, \boldsymbol{w}, \alpha, \beta|X) q_{\boldsymbol{w}}(\boldsymbol{w}) q_\alpha(\alpha) d\boldsymbol{w} d\alpha + \text{const} =$$
$$= \frac{N}{2}\log\beta - \frac{\beta}{2}\sum_{n=1}^N \mathbb{E}(t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n))^2 + (c_0 - 1)\log\beta - d_0\beta + \text{const}.$$

В этом выражении также узнается логарифм гамма-распределения с параметрами:

$$c = c_0 + \frac{N}{2}, \tag{7}$$
$$d = d_0 + \frac{1}{2}\sum_n \mathbb{E}(t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n))^2. \tag{8}$$

Итоговый алгоритм.

Таким образом, на каждой итерации вариационного алгоритма компоненты факторизованного распределения $q(\boldsymbol{w}, \alpha, \beta)$ представляют собой

$$q_{\boldsymbol{w}}(\boldsymbol{w}) = \mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \Sigma), \tag{9}$$
$$q_\alpha(\alpha) = \mathcal{G}(\alpha|a, b), \tag{10}$$
$$q_\beta(\beta) = \mathcal{G}(\beta|c, d), \tag{11}$$

где соответствующие параметры распределений пересчитываются по формулам (3)-(4), (5)-(6) и (7)-(8). При этом необходимые статистики распределений для формул пересчета выглядят следующим образом:

$$\mathbb{E}\alpha = \frac{a}{b},$$
$$\mathbb{E}\beta = \frac{c}{d},$$
$$\mathbb{E}\boldsymbol{w}^T\boldsymbol{w} = \text{tr}\mathbb{E}(\boldsymbol{w}\boldsymbol{w}^T),$$
$$\mathbb{E}(t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n))^2 = t_n^2 - 2\mathbb{E}\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n) + \boldsymbol{\phi}(\boldsymbol{x}_n)^T\mathbb{E}\boldsymbol{w}\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n),$$
$$\mathbb{E}\boldsymbol{w} = \boldsymbol{\mu},$$
$$\mathbb{E}\boldsymbol{w}\boldsymbol{w}^T = \Sigma + \boldsymbol{\mu}\boldsymbol{\mu}^T.$$

Заметим, что по аналогии со схемой Гиббса в вариационном подходе все компоненты распределения пересчитываются последовательно. Сначала фиксируются начальные приближения для всех параметров распределений $\boldsymbol{\mu}, \Sigma, a, b, c, d$. Затем, новые значения $\boldsymbol{\mu}$ и Σ вычисляются по формулам (3)-(4). Эти значения используются для вычисления необходимых статистик распределения $\mathbb{E}\boldsymbol{w}$, $\mathbb{E}\boldsymbol{w}\boldsymbol{w}^T$ и др. После этого находятся новые значения a и b по формулам (5)-(6) и новые статистики $\mathbb{E}\alpha$ и $\mathbb{E}\log\alpha$. И так далее.

Как уже было отмечено выше, в вариационном подходе помимо приближения апостериорного распределения $p(\boldsymbol{w}, \alpha, \beta|\boldsymbol{t}, X) \simeq q_{\boldsymbol{w}}(\boldsymbol{w})q_\alpha(\alpha)q_\beta(\beta)$ мы можем получить также нижнюю границу на обоснованность $\log p(\boldsymbol{t}|X)$:

$$\log p(\boldsymbol{t}|X) \geq \mathcal{L}(q) = \mathbb{E}\log p(\boldsymbol{t}, \boldsymbol{w}, \alpha, \beta|X) - \mathbb{E}\log q_{\boldsymbol{w}}(\boldsymbol{w}) - \mathbb{E}\log q_\alpha(\alpha) - \mathbb{E}\log q_\beta(\beta).$$

С учетом результата (9)-(11) нижняя граница $\mathcal{L}(q)$ вычисляется аналитически:

$$\mathbb{E}\log p(\boldsymbol{t}, \boldsymbol{w}, \alpha, \beta|X) = \frac{N}{2}(\mathbb{E}\log\beta - \log 2\pi) - \frac{\mathbb{E}\beta}{2}\sum_n \mathbb{E}(t_n - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_n))^2 + \frac{M}{2}(\mathbb{E}\log\alpha - \log 2\pi) - \frac{\mathbb{E}\alpha}{2}\mathbb{E}\boldsymbol{w}^T\boldsymbol{w} +$$
$$+ (a_0 - 1)\mathbb{E}\log\alpha - b_0\mathbb{E}\alpha + a_0\log b_0 - \log\Gamma(a_0) + (c_0 - 1)\mathbb{E}\log\beta - d_0\mathbb{E}\beta + c_0\log d_0 - \log\Gamma(c_0).$$

$$
\begin{aligned}
&\mathbb{E}\log q_{\boldsymbol{w}}(\boldsymbol{w}) = -\frac{M}{2}(\log 2\pi + 1) - \frac{1}{2}\log\det\Sigma,\\
&\mathbb{E}\log q_{\alpha}(\alpha) = (a-1)\mathbb{E}\log\alpha - b\mathbb{E}\alpha + a\log b - \log\Gamma(a),\\
&\mathbb{E}\log q_{\beta}(\beta) = (c-1)\mathbb{E}\log\beta - d\mathbb{E}\beta + c\log d - \log\Gamma(c).
\end{aligned}
$$

Для вычисления этих выражений необходимо знать ряд дополнительных статистик распределений, которые вычисляются следующим образом (см. ликбез для гамма-распределения):

$$
\begin{aligned}
\mathbb{E}\log\alpha &= \Psi(a) - \log(b),\\
\mathbb{E}\log\beta &= \Psi(c) - \log(d).
\end{aligned}
$$

В ходе итерационного процесса значение $\mathcal{L}(q)$ не убывает. Итерационный процесс заканчивается, когда значение $\mathcal{L}(q)$ стабилизируется.

Для получения прогноза регрессионной компоненты t_{new} для объекта $\boldsymbol{x}_{new}$ необходимо вычислить следующий интеграл:

$$
p(\boldsymbol{t}_{new}|\boldsymbol{x}_{new}, \boldsymbol{t}, X) = \int p(\boldsymbol{t}_{new}|\boldsymbol{x}_{new}, \boldsymbol{w}, \beta)p(\boldsymbol{w}, \alpha, \beta|\boldsymbol{t}, X)d\boldsymbol{w}d\alpha d\beta.
$$

Подставляя вместо апостериорного распределения $p(\boldsymbol{w}, \alpha, \beta|\boldsymbol{t}, X)$ его факторизованное приближение, получаем:

$$
\begin{aligned}
p(\boldsymbol{t}_{new}|\boldsymbol{x}_{new}, \boldsymbol{t}, X) \simeq \hat{p}(t_{new}|\boldsymbol{x}_{new}, \boldsymbol{t}, X) &= \int p(t_{new}|\boldsymbol{x}_{new}, \boldsymbol{w}, \beta)q_{\boldsymbol{w}}(\boldsymbol{w})q_{\beta}(\beta)d\boldsymbol{w}d\beta =\\
&= \int\left[\int \mathcal{N}(t_{new}|\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_{new}), \beta^{-1})\mathcal{N}(\boldsymbol{w}|\boldsymbol{\mu}, \Sigma)d\boldsymbol{w}\right]\mathcal{G}(\beta|c,d)d\beta =\\
&= \int \mathcal{N}(t_{new}|\boldsymbol{\mu}^T\boldsymbol{\phi}(\boldsymbol{x}_{new}), \beta^{-1} + \boldsymbol{\phi}(\boldsymbol{x}_{new})^T\Sigma\boldsymbol{\phi}(\boldsymbol{x}_{new}))\mathcal{G}(\beta|c,d)d\beta.
\end{aligned}
$$

Последний интеграл не берется аналитически. Однако, данный интеграл является одномерным и поэтому может быть эффективно оценен с помощью метода Монте Карло, т.к. мы можем легко получать выборку из гамма-распределения. Кроме того, можно показать, что $\mathbb{E}_{\hat{p}}t_{new} = \boldsymbol{\mu}^T\boldsymbol{\phi}(\boldsymbol{x}_{new})$, а $\mathbb{D}_{\hat{p}}t_{new} \simeq (\mathbb{E}\beta)^{-1} + \boldsymbol{\phi}(\boldsymbol{x}_{new})^T\Sigma\boldsymbol{\phi}(\boldsymbol{x}_{new})$.