Введение в Байесовские методы

Виктор Кантор

План

- Байесовская вероятность и вывод
- Байесовский вывод в ML моделях
- Принцип наибольшей обоснованности
- Методы оценки обоснованности

Байесовская вероятность и вывод

Частотный подход

Случайность – объективная неопределенность

- Единственное возможное средство анализа – проведение серии испытаний
- Вероятность – предел частоты наступления события

Байесовский подход

Случайность – мера нашего незнания. Есть априорное и апостериорное «незнание»

- Параметры – случайные величины
- При размере выборки в 0 объектов – используем априорное распределение параметра, при размере в 1 и больше – можем получить апостериорное

Критика Байесовского подхода

1. Не предлагаются конструктивные методы вывода априорной вероятности

2. Методы долго работают, часто приводят к вычислительно сложным задачам интегрирования в многомерном пространстве

Что нам дают байесовские методы

1. Способ формализации «здравого смысла»
2. Автоматизация подбора параметров и гиперпараметров (при этом возникают новые гиперпараметры, но более абстрактные)
3. Язык для вывода методов ML из простых базовых предположений

Байесовский вывод

$$\boldsymbol{x} = (x_1, \ldots, x_n)$$

Байесовский вывод

$$\boldsymbol{x} = (x_1, \ldots, x_n)$$

$$p(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{\int p(\boldsymbol{x}|\theta)p(\theta)d\theta}$$

Байесовский вывод

$$\boldsymbol{x} = (x_1, \ldots, x_n)$$

$$p(\theta|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\theta)p(\theta)}{\int p(\boldsymbol{x}|\theta)p(\theta)d\theta}$$

Процесс получения апостериорной вероятности из априорной и правдоподобия выборки – и есть байесовский вывод

Точечная оценка параметра

$$\hat{\theta}_B = \int \theta p(\theta|\boldsymbol{x}) d\theta$$

$$\hat{\theta}_{MP} = \arg\max P(\theta|\boldsymbol{x})$$

Пример: оценка вероятности выпадения орла

Задача – по серии из n подбрасываний монетки и m выпадений орла оценить вероятность q выпадения орла

Пример: оценка вероятности выпадения орла

Задача – по серии из n подбрасываний монетки и m выпадений орла оценить вероятность q выпадения орла

$$p(m|n,q) = C_n^m q^m (1-q)^{n-m} \sim \mathcal{B}(m|n,q)$$

Пример: оценка вероятности выпадения орла

Задача – по серии из n подбрасываний монетки и m выпадений орла оценить вероятность q выпадения орла

$$p(m|n,q) = C_n^m q^m (1-q)^{n-m} \sim \mathcal{B}(m|n,q)$$

Вопрос: какую оценку q даст метод максимального правдоподобия?

Пример: оценка вероятности выпадения орла

Байесовская оценка:

$$p(m|n,q) = C_n^m q^m (1-q)^{n-m} \sim \mathcal{B}(m|n,q)$$

$$p(q|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} q^{a-1}(1-q)^{b-1} \sim \mathrm{Beta}(q|a,b)$$

Пример: оценка вероятности выпадения орла

Байесовская оценка:

$$p(m|n,q) = C_n^m q^m (1-q)^{n-m} \sim \mathcal{B}(m|n,q)$$

$$p(q|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} q^{a-1}(1-q)^{b-1} \sim \mathrm{Beta}(q|a,b)$$

$$p(q|\text{«}m\text{ орлов»}) \sim \mathrm{Beta}(q|a+m, b+n-m)$$

Пример: оценка вероятности выпадения орла

Байесовская оценка:

$$p(m|n,q) = C_n^m q^m (1-q)^{n-m} \sim \mathcal{B}(m|n,q)$$

$$p(q|a,b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)} q^{a-1}(1-q)^{b-1} \sim \mathrm{Beta}(q|a,b)$$

$$p(q|\text{«}m\text{ орлов»}) \sim \mathrm{Beta}(q|a+m, b+n-m)$$

$$\boxed{\hat{q}_B = \int_0^1 p(q|\text{«}m\text{ орлов»})q\,dq = \frac{m+1}{n+2}}$$

Сопряженные распределения

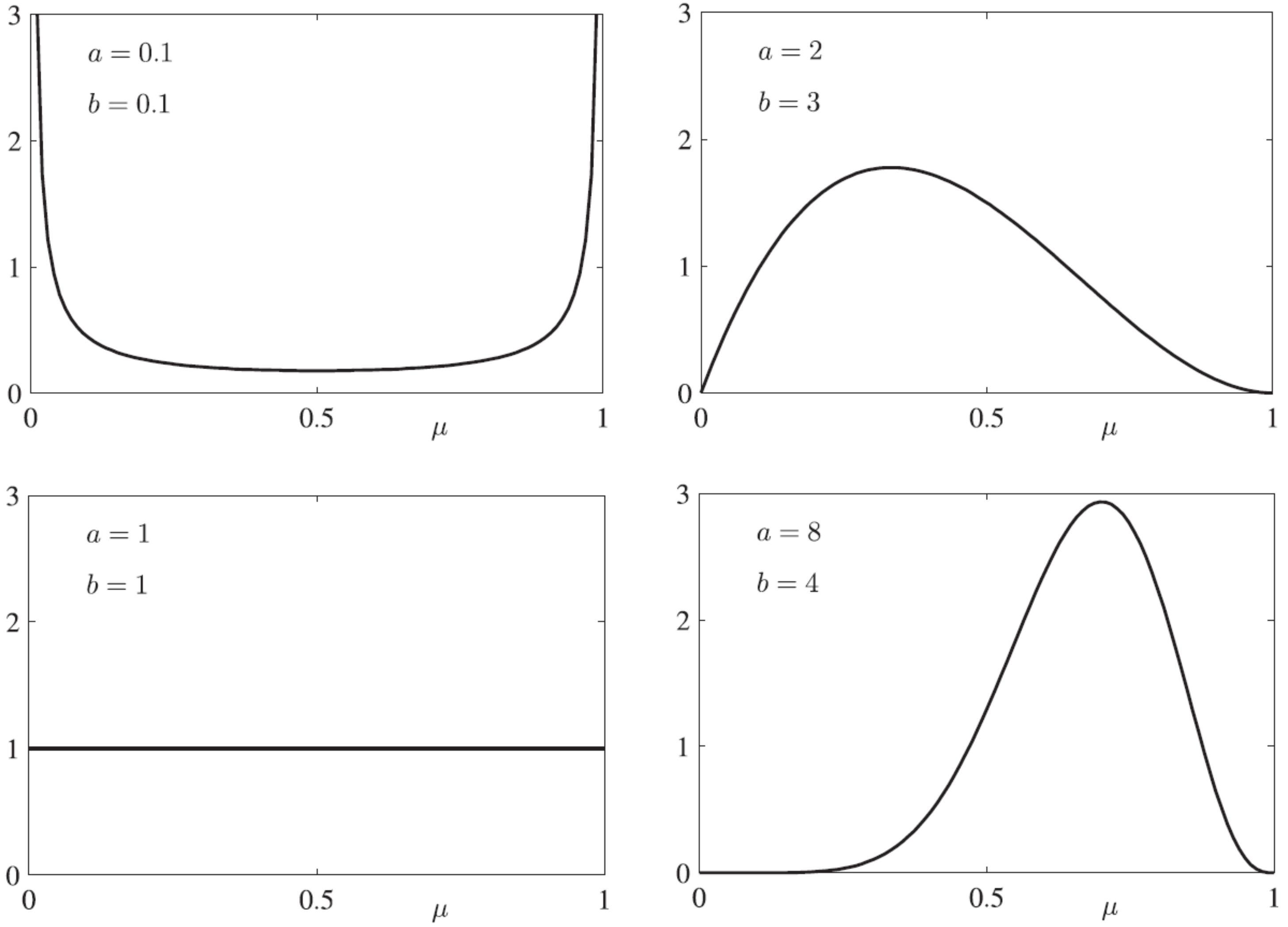
Если из $p(\theta) \sim \mathcal{A}(\alpha_0)$ и $p(\boldsymbol{x}|\theta) \sim \mathcal{B}(\beta)$

следует $p(\theta|\boldsymbol{x}) \sim \mathcal{A}(\alpha_1)$, то говорят, что распределение

$\mathcal{A}$ сопряженное к $\mathcal{B}$

Бета-распределение

Пример: распределение, сопряженное к биномиальному



Еще про сопряженные распределения

- Сопряженное распределение можно выписать явно для любого распределения из экспоненциального семейства:

$$p(\boldsymbol{x}|\boldsymbol{\alpha}) = h(\boldsymbol{x})g(\boldsymbol{\alpha})\exp(\boldsymbol{\alpha}^T u(\boldsymbol{x}))$$

- К этому семейству относятся нормальное, гамма-, бета-, равномерное, Бернулли, Дирихле, хи-квадрат, Пуассоновское и другие распределения

- Именно сопряженное распределение лучше брать как априорное

Байесовский вывод в ML моделях

Оценка параметров модели по MLE

$$\theta_{ML} = \underset{\theta}{\operatorname{argmax}}\, p_\theta(X) = \underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \ln p_\theta(x_i)$$

Байесовская оценка параметров модели

$$\theta_{MP} = \underset{\theta}{\operatorname{argmax}}\, p(\theta|X) = \underset{\theta}{\operatorname{argmax}}\, p(X|\theta)p(\theta) =$$

$$\underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \ln p(x_i|\theta) + \ln p(\theta)$$

Байесовская оценка параметров модели

$$\theta_{MP} = \underset{\theta}{\operatorname{argmax}}\, p(\theta|X) = \underset{\theta}{\operatorname{argmax}}\, p(X|\theta)p(\theta) =$$

$$\underset{\theta}{\operatorname{argmax}} \sum_{i=1}^{n} \ln p(x_i|\theta) + \boxed{\ln p(\theta)}$$

Регуляризатор для функции правдоподобия

Пример: априор в линейной регрессии

Линейная регрессия:

$$(X, \boldsymbol{t}) = \{\boldsymbol{x}_i, t_i\}_{i=1}^n \qquad t = f(\boldsymbol{x}) + \varepsilon \qquad \varepsilon \sim \mathcal{N}(\varepsilon | 0, \sigma^2)$$

$$f(\boldsymbol{x}) = \sum_{j=1}^{m} w_j \phi_j(\boldsymbol{x}) = \boldsymbol{w}^T \boldsymbol{\phi}(\boldsymbol{x})$$

Пример: априор в линейной регрессии

MLE: $$p(\mathbf{t}|X,\boldsymbol{w}) = \prod_{i=1}^{n} p(t_i|\boldsymbol{x}_i,\boldsymbol{w}) = \prod_{i=1}^{n} \mathcal{N}(t_i|f(\boldsymbol{x}_i,\boldsymbol{w}),\sigma^2) =$$

$$\prod_{i=1}^{n} \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(t_i - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2}{2\sigma^2}\right) = \frac{1}{(\sqrt{2\pi}\sigma)^n} \exp\left(-\frac{\sum_{i=1}^{n}(t_i - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2}{2\sigma^2}\right)$$

$$\boxed{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(t_i - \boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i))^2 = -\frac{1}{2\sigma^2}(\boldsymbol{t} - \Phi\boldsymbol{w})^T(\boldsymbol{t} - \Phi\boldsymbol{w}) \to \max_{\boldsymbol{w}}}$$

$$\boxed{\boldsymbol{w}_{ML} = (\Phi^T\Phi)^{-1}\Phi^T\boldsymbol{t}}$$

Пример: априор в линейной регрессии

Байес:

$$w_{MP} = \arg\max_{\boldsymbol{w}} p(\boldsymbol{w}|X, \boldsymbol{t}) = \arg\max_{\boldsymbol{w}} p(\boldsymbol{t}|X, \boldsymbol{w})p(\boldsymbol{w})$$

$$\boxed{p(\boldsymbol{w}|\alpha) = \mathcal{N}(\boldsymbol{w}|0, \alpha^{-1}I)}$$

$$-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(t_i - \phi(\boldsymbol{x}_i))^2 - \frac{\alpha}{2}\|\boldsymbol{w}\|^2 \to \max_{\boldsymbol{w}}$$

$$\boldsymbol{w}_{MP} = (\sigma^{-2}\Phi^T\Phi + \alpha I)^{-1}\sigma^{-2}\Phi^T\boldsymbol{t}$$

Принцип наибольшей обоснованности

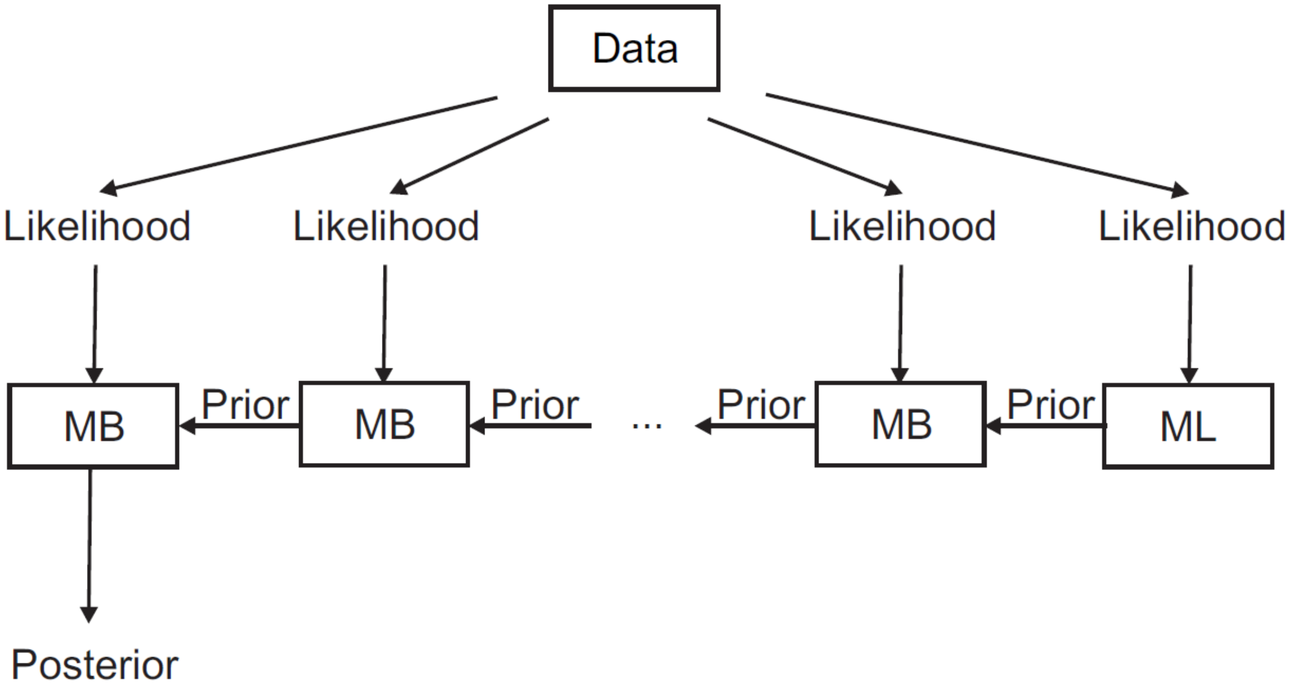
Иерархическая схема Байеса

Распределение параметра тоже можно задать в параметрическом виде: $p(\theta) = p(\theta|\alpha)$

$$p(\alpha|\boldsymbol{x}) = \frac{p(\boldsymbol{x}|\alpha)p(\alpha)}{p(\boldsymbol{x})} = \frac{p(\boldsymbol{x}|\alpha)p(\alpha)}{\int p(\boldsymbol{x}|\alpha)p(\alpha)d\alpha}$$

$p(\boldsymbol{x}|\alpha) = \int p(\boldsymbol{x}|\theta)p(\theta|\alpha)d\theta$ - обоснованность (evidence) модели

Иерархическая схема Байеса



Обычно ограничиваются двухуровневым выводом

Пример: генератор случайных чисел

- Есть генератор случайных чисел от 1 до N и два числа, сгенерированные им: 6 и 8.
- Определить матожидание генерируемых чисел
- Определить, N = 10 или N = 100

Пример: генератор случайных чисел

- Есть генератор случайных чисел от 1 до N и два числа, сгенерированные им: 6 и 8.
- Определить матожидание генерируемых чисел
- Определить, N = 10 или N = 100

Какой ответ на первый вопрос даст классическая максимизация правдоподобия?

Пример: генератор случайных чисел

Байесовский взгляд:

$$D(\boldsymbol{q}|\boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha^0})} q_1^{\alpha_1^0 - 1} \dots q_N^{\alpha_N^0 - 1}, \quad \sum_{i=1}^{N} q_i = 1, \; q_i \geq 0 \qquad p(x_1, x_2|\boldsymbol{q}) = q_8 q_6$$

По формуле Байеса:

$$p(\boldsymbol{q}|x_1, x_2) = \frac{1}{Z} q_1^0 \dots q_5^0 q_6^1 q_7^0 q_8^1 q_9^0 \dots q_N^0 = D(\boldsymbol{q}|\boldsymbol{\alpha^1})$$

Пример: генератор случайных чисел

$$\mathbb{E}q_i = \frac{\alpha_i}{\sum_{j=1}^{N} \alpha_j}$$

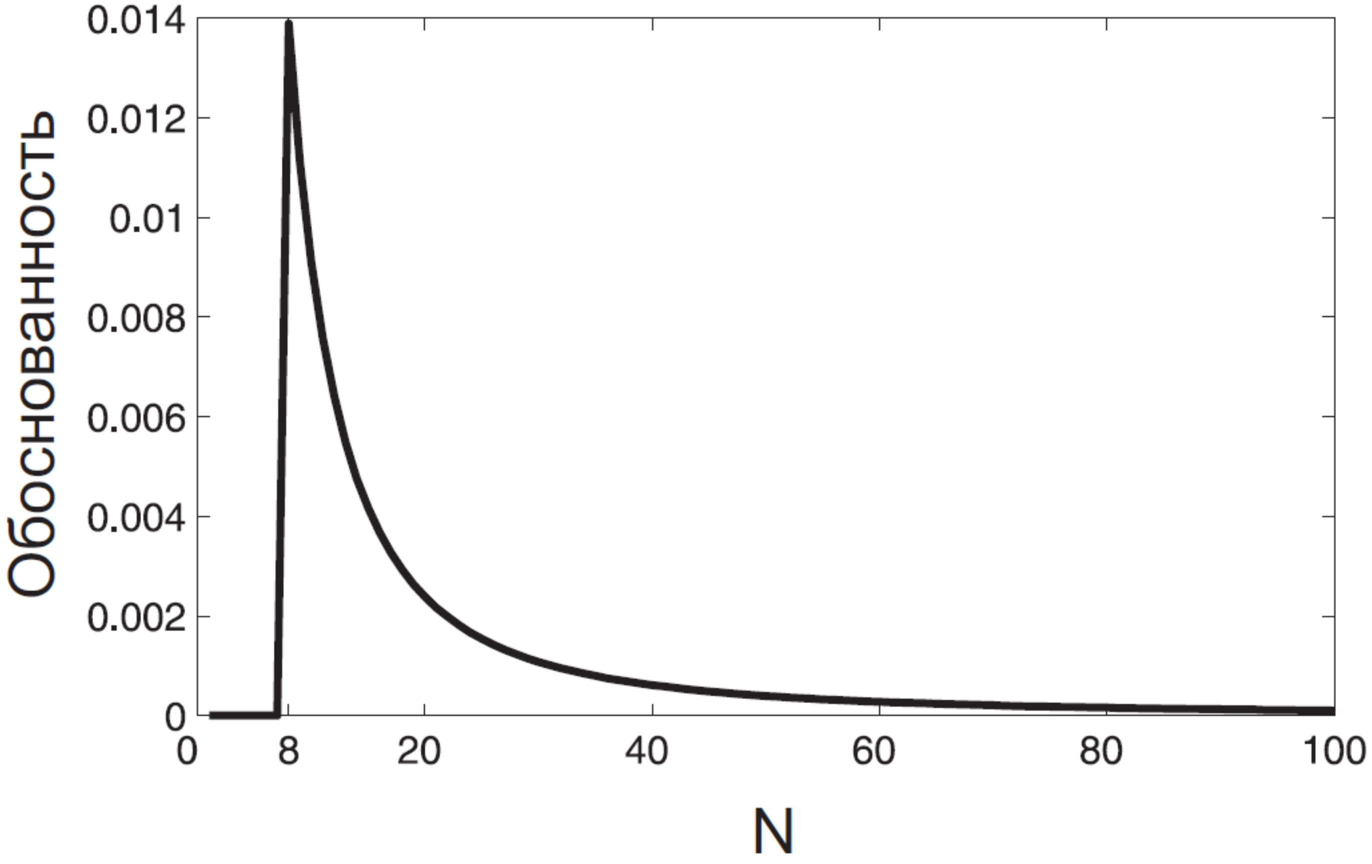
$$q_6 = q_8 = \frac{2}{12} = \frac{1}{6} \approx 0.16, \quad q_i = \frac{1}{12} \approx 0.08 \ \ \forall i \neq 6, 8$$

$$\mu_{MP}(N = 10) = 5.75$$

$$q_6 = q_8 = \frac{2}{102} = \frac{1}{51} \approx 0.02, \quad q_i = \frac{1}{102} \approx 0.01 \ \ \forall i \neq 6, 8$$

$$\mu_{MP}(N = 100) \approx 49.65$$

Пример: генератор случайных чисел



Метод релевантных векторов в регрессии

Более сложное априорное распределение, чем было в Ridge Regression + Байесовский вывод:

$$p(\boldsymbol{w}|\boldsymbol{\alpha}) = \mathcal{N}(0, A^{-1})$$

$$p(\boldsymbol{t}|\boldsymbol{\alpha}, \sigma^2) = \int p(\boldsymbol{t}|X, \boldsymbol{w}, \sigma^2) p(\boldsymbol{w}|\boldsymbol{\alpha}) d\boldsymbol{w} \to \max_{\boldsymbol{\alpha}, \sigma^2}$$

Метод релевантных векторов в регрессии

Обоснованность:

$$p(\boldsymbol{t}|\boldsymbol{\alpha}, \sigma^2) = \int p(\boldsymbol{t}|X, \boldsymbol{w}, \sigma^2)p(\boldsymbol{w}|\boldsymbol{\alpha})d\boldsymbol{w} = \int Q(\boldsymbol{w})d\boldsymbol{w}$$

Метод релевантных векторов в регрессии

Обоснованность:

$$p(\mathbf{t}|\boldsymbol{\alpha}, \sigma^2) = \int p(\mathbf{t}|X, \boldsymbol{w}, \sigma^2) p(\boldsymbol{w}|\boldsymbol{\alpha}) d\boldsymbol{w} = \int Q(\boldsymbol{w}) d\boldsymbol{w}$$

$$L(\boldsymbol{w}) = \log Q(\boldsymbol{w})$$

$$L(\boldsymbol{w}) = L(\boldsymbol{w}_{MP}) + (\nabla_{\boldsymbol{w}} L(\boldsymbol{w}_{MP}))^T (\boldsymbol{w} - \boldsymbol{w}_{MP}) + \frac{1}{2} (\boldsymbol{w} - \boldsymbol{w}_{MP})^T H (\boldsymbol{w} - \boldsymbol{w}_{MP})$$

$$\boldsymbol{w}_{MP} = \arg\max_{\boldsymbol{w}} L(\boldsymbol{w}) \;\Rightarrow\; \nabla_{\boldsymbol{w}} L(\boldsymbol{w}_{MP}) = 0$$

$$H = \nabla\nabla L(\boldsymbol{w}_{MP})$$

Метод релевантных векторов в регрессии

$$L(\boldsymbol{w}) = L(\boldsymbol{w}_{MP}) + (\nabla_{\boldsymbol{w}} L(\boldsymbol{w}_{MP}))^T (\boldsymbol{w} - \boldsymbol{w}_{MP}) + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{MP})^T H (\boldsymbol{w} - \boldsymbol{w}_{MP})$$

$$\boldsymbol{w}_{MP} = \arg\max_{\boldsymbol{w}} L(\boldsymbol{w}) \;\Rightarrow\; \nabla_{\boldsymbol{w}} L(\boldsymbol{w}_{MP}) = 0$$

$$H = \nabla\nabla L(\boldsymbol{w}_{MP})$$

тогда:

$$\int Q(\boldsymbol{w}) d\boldsymbol{w} = \int \exp\left(L(\boldsymbol{w}_{MP}) + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{MP})^T H (\boldsymbol{w} - \boldsymbol{w}_{MP}) \right) d\boldsymbol{w} =$$

$$= Q(\boldsymbol{w}_{MP}) \sqrt{(2\pi)^m} \sqrt{\det((-H)^{-1})} = \sqrt{(2\pi)^m} \frac{Q(\boldsymbol{w}_{MP})}{\sqrt{\det(-H)}}$$

Метод релевантных векторов в регрессии

$$L(\boldsymbol{w}) = L(\boldsymbol{w}_{MP}) + (\nabla_{\boldsymbol{w}} L(\boldsymbol{w}_{MP}))^T (\boldsymbol{w} - \boldsymbol{w}_{MP}) + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{MP})^T H (\boldsymbol{w} - \boldsymbol{w}_{MP})$$

$$\boldsymbol{w}_{MP} = \arg\max_{\boldsymbol{w}} L(\boldsymbol{w}) \;\Rightarrow\; \nabla_{\boldsymbol{w}} L(\boldsymbol{w}_{MP}) = 0$$

$$H = \nabla\nabla L(\boldsymbol{w}_{MP})$$

тогда:

$$\int Q(\boldsymbol{w}) d\boldsymbol{w} = \int \exp\left(L(\boldsymbol{w}_{MP}) + \frac{1}{2}(\boldsymbol{w} - \boldsymbol{w}_{MP})^T H (\boldsymbol{w} - \boldsymbol{w}_{MP})\right) d\boldsymbol{w} =$$

$$= Q(\boldsymbol{w}_{MP})\sqrt{(2\pi)^m}\sqrt{\det((-H)^{-1})} = \sqrt{(2\pi)^m}\frac{Q(\boldsymbol{w}_{MP})}{\sqrt{\det(-H)}}$$

«Взяли» интеграл малыми усилиями, осталось посчитать Q в точке оптимума и оценить нормировочную константу

Метод релевантных векторов в регрессии

Получение MP оценки на w:

$$\beta = \sigma^{-2}$$

$$L(\boldsymbol{w}) = -\frac{1}{2}\beta(\boldsymbol{t} - \Phi\boldsymbol{w})^T(\boldsymbol{t} - \Phi\boldsymbol{w}) - \frac{1}{2}\boldsymbol{w}^T A\boldsymbol{w} - \frac{n}{2}\log(2\pi) -$$

$$-\frac{m}{2}\log(2\pi) + \frac{1}{2}\log\det(A) = -\frac{1}{2}\beta[\boldsymbol{t}^T\boldsymbol{t} - 2\boldsymbol{w}^T\Phi^T\boldsymbol{t} + \boldsymbol{w}^T\Phi^T\Phi\boldsymbol{w}] + C$$

$$\nabla L(\boldsymbol{w}) = -\frac{1}{2}\beta(-2\Phi^T\boldsymbol{t} + 2\Phi^T\Phi\boldsymbol{w}) - A\boldsymbol{w} = 0 \;\Rightarrow\; \boxed{\boldsymbol{w}_{MP} = (\beta\Phi^T\Phi + A)^{-1}\beta\Phi^T\boldsymbol{t}}$$

Метод релевантных векторов в регрессии

Итоговое выражение

$$p(\boldsymbol{t}|X,\boldsymbol{\alpha},\sigma^2) = \int p(\boldsymbol{t}|X,\sigma^2)p(\boldsymbol{w}|\boldsymbol{\alpha})d\boldsymbol{w} = \sqrt{(2\pi)^m}\frac{Q(\boldsymbol{w}_{MP})}{\sqrt{\det(-H)}} =$$

$$\frac{1}{\sqrt{(2\pi)^n}\det(\beta^{-1}I + \Phi A^{-1}\Phi^T)^{1/2}}\exp\left(-\frac{1}{2}\boldsymbol{t}^T(\beta^{-1}I + \Phi A^{-1}\Phi^T)^{-1}\boldsymbol{t}\right)$$

Метод релевантных векторов в регрессии

Итерационные формулы для параметров (получим приравняв производные обоснованности по ним к нулю):

$$\alpha_i^{new} = \frac{\gamma_i}{w_{MP,i}^2} \qquad \gamma_i = 1 - \alpha_i^{old}\Sigma_{ii}$$

$$(\sigma^2)^{new} = \frac{\|\boldsymbol{t} - \Phi\boldsymbol{w}\|^2}{n - \sum_{i=1}^{m}\gamma_i}$$

$$\Sigma = (\beta\Phi^T\Phi + A)^{-1} \qquad \boldsymbol{w}_{MP} = \beta\Sigma\Phi^T\boldsymbol{t}$$

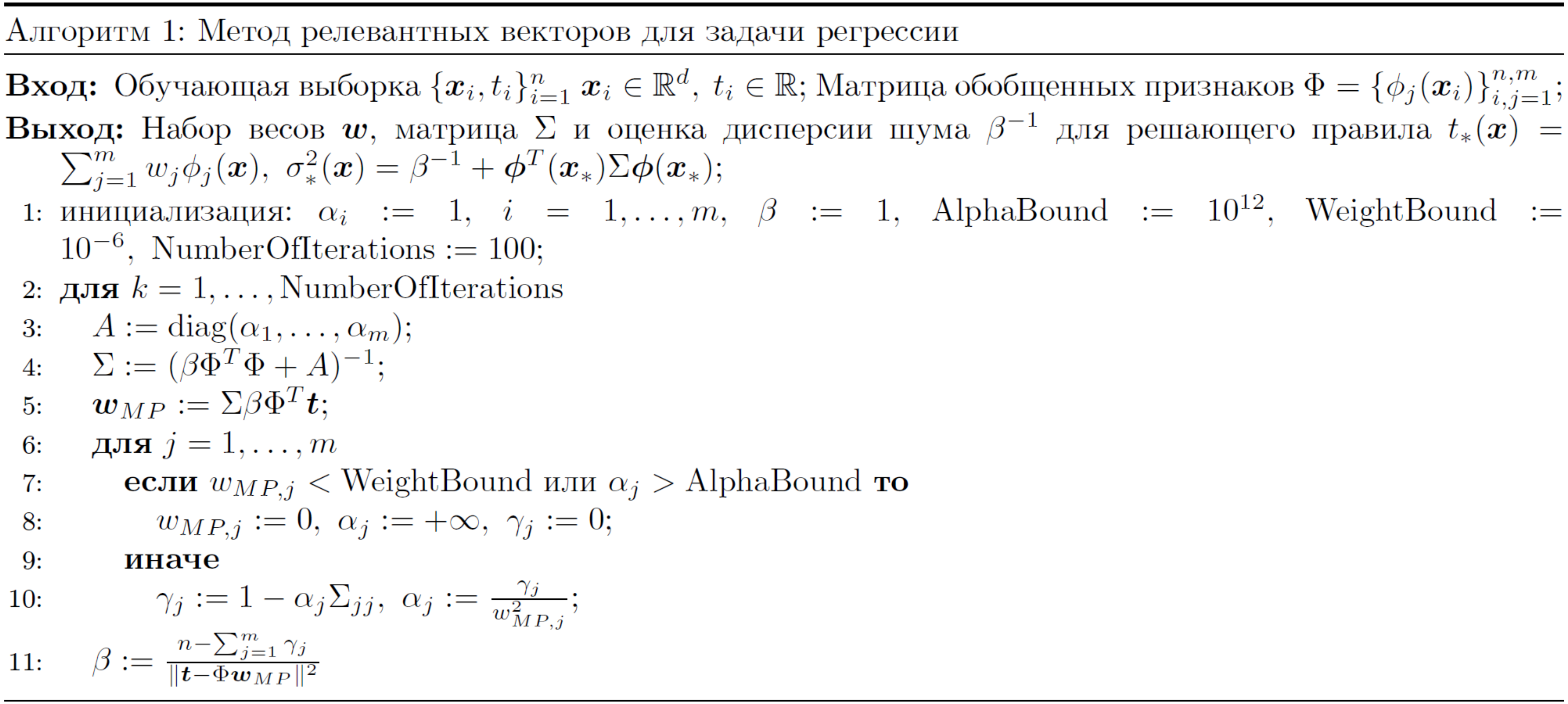
Метод релевантных векторов в регрессии

Получение прогноза:

$$p(t_*|\boldsymbol{x}_*, \boldsymbol{t}, X) = \int p(t_*|\boldsymbol{x}_*, \boldsymbol{w}, \sigma^2_{MP})p(\boldsymbol{w}|\boldsymbol{t}, X, \boldsymbol{\alpha}_{MP}, \sigma^2_{MP})d\boldsymbol{w} = \mathcal{N}(t_*|y_*, \sigma^2_*)$$

$$y_* = \boldsymbol{w}^T_{MP}\boldsymbol{\phi}(\boldsymbol{x}_*)$$
$$\sigma^2_* = \sigma^2_{MP} + \boldsymbol{\phi}(\boldsymbol{x}_*)^T\Sigma\boldsymbol{\phi}(\boldsymbol{x}_*)$$

Метод релевантных векторов в регрессии



Алгоритм 1: Метод релевантных векторов для задачи регрессии

Вход: Обучающая выборка $\{\boldsymbol{x}_i, t_i\}_{i=1}^{n}$ $\boldsymbol{x}_i \in \mathbb{R}^d$, $t_i \in \mathbb{R}$; Матрица обобщенных признаков $\Phi = \{\phi_j(\boldsymbol{x}_i)\}_{i,j=1}^{n,m}$;

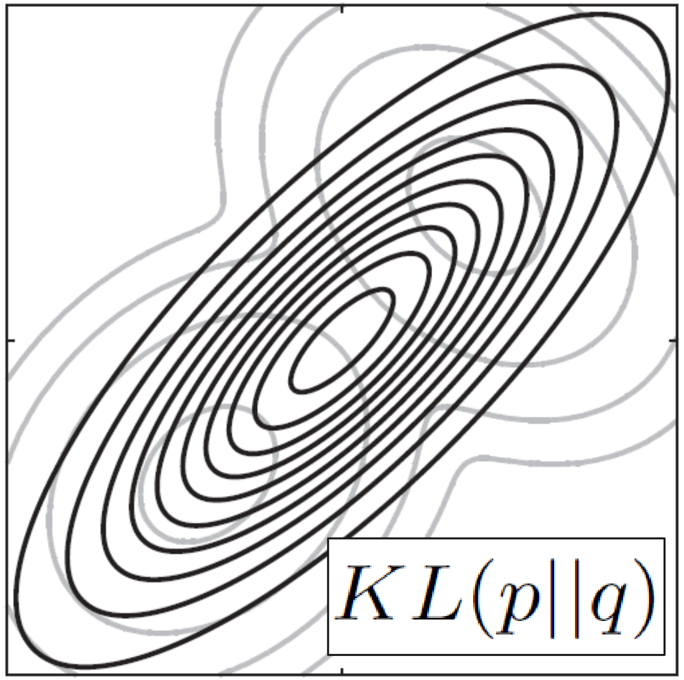
Выход: Набор весов $\boldsymbol{w}$, матрица Σ и оценка дисперсии шума β^{-1} для решающего правила $t_*(\boldsymbol{x}) = \sum_{j=1}^{m} w_j \phi_j(\boldsymbol{x})$, $\sigma_*^2(\boldsymbol{x}) = \beta^{-1} + \boldsymbol{\phi}^T(\boldsymbol{x}_*)\Sigma\boldsymbol{\phi}(\boldsymbol{x}_*)$;

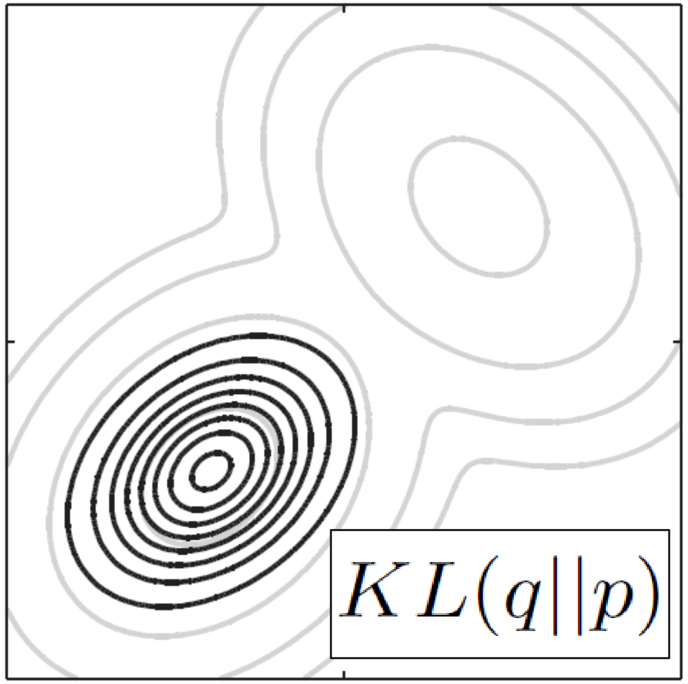
```
инициализация: $\alpha_i := 1,\ i = 1, \ldots, m,\ \beta := 1$, AlphaBound $:= 10^{12}$, WeightBound $:= 10^{-6}$, NumberOfIterations $:= 100$;
**для** $k = 1, \ldots,$ NumberOfIterations
    $A := \mathrm{diag}(\alpha_1, \ldots, \alpha_m)$;
    $\Sigma := (\beta\Phi^T\Phi + A)^{-1}$;
    $\boldsymbol{w}_{MP} := \Sigma\beta\Phi^T\boldsymbol{t}$;
    **для** $j = 1, \ldots, m$
        **если** $w_{MP,j} <$ WeightBound или $\alpha_j >$ AlphaBound **то**
            $w_{MP,j} := 0,\ \alpha_j := +\infty,\ \gamma_j := 0$;
        **иначе**
            $\gamma_j := 1 - \alpha_j\Sigma_{jj},\ \alpha_j := \frac{\gamma_j}{w_{MP,j}^2}$;
    $\beta := \frac{n - \sum_{j=1}^{m}\gamma_j}{\|\boldsymbol{t} - \Phi\boldsymbol{w}_{MP}\|^2}$
```

Методы оценки обоснованности

Дивергенция Кульбака-Лейблера

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





Задача оценки обоснованности

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

$p(X)$ - обоснованность выбранной модели

Прямое интегрирование обычно невозможно, поэтому ограничиваются приближением $p(Z|X)$ некоторым распределением $q(Z)$

Разложение обоснованности

$$\log p(X) = \log p(X) \int q(Z)dZ = \int \log p(X)q(Z)dZ =$$

$$\int \log \frac{p(X,Z)}{p(Z|X)} q(Z)dZ = \int \log \frac{p(X,Z)q(Z)}{q(Z)p(Z|X)} q(Z)dZ =$$

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

Факторизованное приближение

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

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

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

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

Факторизованное приближение

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

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

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

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

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

Факторизованное приближение

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

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

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

Пример: вариационная линейная регрессия

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

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

Будем искать приближение $p(\boldsymbol{w}, \alpha|\boldsymbol{t})$ в виде $q(\boldsymbol{w}, \alpha) = q(\boldsymbol{w})q(\alpha)$

Пример: вариационная линейная регрессия

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

$$p(\alpha) = \mathcal{G}(\alpha|a_0, b_0) \qquad q(\boldsymbol{w}, \alpha) = q(\boldsymbol{w})q(\alpha)$$

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

$$\mathbb{E}_{\boldsymbol{w}} \left(\log p(\boldsymbol{w}|\alpha)p(\alpha)\right) + C = \mathbb{E}_{\boldsymbol{w}} \log p(\boldsymbol{w}|\alpha) + \log p(\alpha) + C =$$

$$\frac{m}{2} \log \alpha - \frac{\alpha}{2}\mathbb{E}\boldsymbol{w}^T\boldsymbol{w} + (a_0 - 1) \log \alpha - b_0\alpha + C_1$$

Пример: вариационная линейная регрессия

$$\alpha \sim \mathcal{G}(\alpha|a_n, b_n) \qquad a_n = a_0 + \frac{m}{2}, \quad b_n = b_0 + \frac{1}{2}\mathbb{E}\boldsymbol{w}^T\boldsymbol{w}$$

Аналогично для w:

$$\log q^*(\boldsymbol{w}) = \mathbb{E}_\alpha \log p(\boldsymbol{t}, \boldsymbol{w}, \alpha) = \mathbb{E}_\alpha \log \left(p(\boldsymbol{t}|\boldsymbol{w})p(\boldsymbol{w}|\alpha)p(\alpha)\right) = \log p(\boldsymbol{t}|\boldsymbol{w}) + \mathbb{E}_\alpha \log p(\boldsymbol{w}|\alpha) + C =$$

$$-\frac{\beta}{2}\sum_{i=1}^{n}(\boldsymbol{w}^T\boldsymbol{\phi}(\boldsymbol{x}_i) - t_i)^2 - \frac{1}{2}\mathbb{E}\alpha \cdot \boldsymbol{w}^T\boldsymbol{w} + C_1 = -\frac{1}{2}\boldsymbol{w}^T(\mathbb{E}\alpha I + \beta\Phi^T\Phi)\boldsymbol{w} + \beta\boldsymbol{w}^T\Phi^T\boldsymbol{t} + C_2$$

В итоге:

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

Другой подход: методы Монте-Карло

$$\int_a^b f(x)dx \approx \frac{b-a}{n}\sum_{i=1}^{n} f(x_i) = \hat{f}, \quad x_i \sim U[a,b]$$

Другой подход: методы Монте-Карло

$$\int_a^b f(x)dx \approx \frac{b-a}{n}\sum_{i=1}^{n} f(x_i) = \hat{f}, \quad x_i \sim U[a,b]$$

$$\int_D f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \approx \frac{|D|}{n}\sum f(\boldsymbol{x}_i)p(\boldsymbol{x}_i), \quad \boldsymbol{x} \sim U(D)$$

Другой подход: методы Монте-Карло

$$\int_a^b f(x)dx \approx \frac{b-a}{n}\sum_{i=1}^{n} f(x_i) = \hat{f}, \quad x_i \sim U[a,b]$$

$$\int_D f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \approx \frac{|D|}{n}\sum f(\boldsymbol{x}_i)p(\boldsymbol{x}_i), \quad \boldsymbol{x} \sim U(D)$$

Такой подход для вероятностных интегралов оказывается не слишком эффективным (как вы думаете, почему?)

Оценка матожидания сэмплированием

$$\int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \approx \frac{1}{n}\sum f(\boldsymbol{x}_i), \quad \boldsymbol{x} \sim p(\boldsymbol{x})$$

Оценка матожидания сэмплированием

$$\int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} \approx \frac{1}{n}\sum f(\boldsymbol{x}_i), \quad \boldsymbol{x} \sim p(\boldsymbol{x})$$

Главная наша проблема теперь – как сэмплировать
из распределения p или хотя бы чего-то похожего

Модификации МС методов

- Сэмплирование из другого распределения

$$\mathbb{E}_p f = \int f(\boldsymbol{x})p(\boldsymbol{x})d\boldsymbol{x} = \frac{1}{Z_p}\int f(\boldsymbol{x})\frac{\tilde{p}(\boldsymbol{x})}{q(\boldsymbol{x})}q(\boldsymbol{x})d\boldsymbol{x} \approx$$

$$\frac{1}{nZ_p}\sum_{i=1}^{n} f(\boldsymbol{x}_i)\frac{\tilde{p}(\boldsymbol{x}_i)}{q(\boldsymbol{x}_i)} = \frac{1}{n\sum_{i=1}^{n} r_i}\sum_{i=1}^{n} f(\boldsymbol{x}_i)r_i, \qquad \boldsymbol{x} \sim q(\boldsymbol{x})$$

- MCMC (Markov Chain Monte Carlo)
- Гибридные МС методы