

# Байесовский выбор моделей: методы Монте-Карло по схеме марковских цепей (MCMC)

Александр Адуенко

25е ноября 2020

## Содержание предыдущих лекций

- Формула Байеса и формула полной вероятности;
- Определение априорных вероятностей и selection bias;
- (Множественное) тестирование гипотез
- Экспоненциальное семейства. Достаточные статистики.
- Наивный байесовский классификатор. Связь целевой функции и вероятностной модели.
- Линейная регрессия: связь МНК и  $w_{ML}$ , регуляризации и  $w_{MAP}$ .
- Свойство сопряженности априорного распределения правдоподобию.
- Прогноз для одиночной модели:

$$p(\mathbf{y}_{\text{test}} | \mathbf{X}_{\text{test}}, \mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}) = \int p(\mathbf{y}_{\text{test}} | \mathbf{w}, \mathbf{X}_{\text{test}}) p(\mathbf{w} | \mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}) d\mathbf{w}.$$

- Связь апостериорной вероятности модели и обоснованности
- Обоснованность: понимание и связь со статистической значимостью.
- Логистическая регрессия: проблемы ML-оценки  $w$  и связь априорного распределения с отбором признаков.
- EM-алгоритм и отбор признаков в байесовской линейной регрессии.
- Вариационный EM-алгоритм. Смесь моделей лог. регрессии.
- Гауссовские процессы. Учёт эволюции моделей во времени.
- Построение адекватных мультимodelей.

# EM и вариационный EM-алгоритм (воспоминание)

Пусть  $\mathbf{D} = (\mathbf{X}, \mathbf{y})$  – наблюдаемые переменные,  $\mathbf{Z}$  – скрытые переменные.  
 $p(\mathbf{D}, \mathbf{Z}|\Theta) = p(\mathbf{D}|\mathbf{Z}, \Theta)p(\mathbf{Z}|\Theta)$ .

**Вопрос 1:** как решить задачу  $p(\mathbf{D}|\Theta) = \int p(\mathbf{D}, \mathbf{Z}|\Theta)d\mathbf{Z} \rightarrow \max_{\Theta}$ ?

EM-алгоритм

Введем  $F(q, \Theta) = - \int q(\mathbf{Z}) \log q(\mathbf{Z})d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{D}, \mathbf{Z}|\Theta)d\mathbf{Z} = \log p(\mathbf{D}|\Theta) - D_{\text{KL}}(q||p(\mathbf{Z}|\mathbf{D}, \Theta))$ .

**Идея 1:**  $p(\mathbf{D}|\Theta) \rightarrow \max_{\Theta}$  заменим на  $F(q, \Theta) \rightarrow \max_{q, \Theta}$ .

**Идея 2:** Пошагово оптимизируем по  $\Theta$  и  $q$ , то есть

**1** E-шаг:  $q^s = F(q, \Theta^{s-1}) \rightarrow \max_{q \in Q}$ ;

**2** M-шаг:  $\Theta^s = F(q^s, \Theta) \rightarrow \max_{\Theta}$ .

**Вариационный EM-алгоритм:**  $Q = \left\{ q(\mathbf{Z}) : q(\mathbf{Z}) = \prod_{k=1}^K q(\mathbf{z}_k) \right\}$ .

Прогноз:

$p(y|\mathbf{x}, \mathbf{D}, \Theta) = \int p(y, \mathbf{Z}|\mathbf{x}, \mathbf{D}, \Theta)d\mathbf{Z} = \int p(y|\mathbf{x}, \mathbf{Z})p(\mathbf{Z}|\mathbf{D}, \Theta)d\mathbf{Z}$ .

**Вопрос 2:** Как быть с  $p(\mathbf{Z}|\mathbf{D}, \Theta) = \frac{p(\mathbf{D}, \mathbf{Z}|\Theta)}{p(\mathbf{D}|\Theta)}$ ?

# Необходимость сэмплирования

Пусть есть некоторая переменная  $\mathbf{Z}$  с распределением  $p(\mathbf{Z})$ .

- Найти  $P(f(\mathbf{Z}) > 0)$ ;
  - $P(\mathbf{w}^T \mathbf{x} > B)$ , где  $\mathbf{w}$  – веса признаков,  $\mathbf{x}$  – признакововое описание, а  $\mathbf{w}^T \mathbf{x}$  – ожидаемый доход.
- $E f(\mathbf{Z}) = \int f(\mathbf{Z}) p(\mathbf{Z}) d\mathbf{Z} \approx \frac{1}{m} \sum_{i=1}^m f(\mathbf{Z}_i)$ ;
  - $p(\mathbf{y}|\mathbf{X}, \mathbf{A}) = \int p(\mathbf{y}|\mathbf{X}, \mathbf{w}) p(\mathbf{w}|\mathbf{A}) d\mathbf{w}$  в лог. регрессии;
  - $E_{q(\mathbf{Z})} \log p(\mathbf{X}, \mathbf{Z}|\Theta)$  на  $M$  шаге EM-алгоритма.

**Вопрос 1:** что делать, если  $p(\mathbf{Z})$  известно с точностью до константы, то есть  $p(\mathbf{Z}) \propto \tilde{p}(\mathbf{Z})$ ?

**Вопрос 2:** что делать для решения задачи  $p(\mathbf{Z}) \rightarrow \max_{\mathbf{Z}}$ , если  $p(\mathbf{Z})$  известно, но задача максимизации аналитически не решается?

Метод обратной функции ( $p(z)$  известно,  $z \in \mathbb{R}$ )

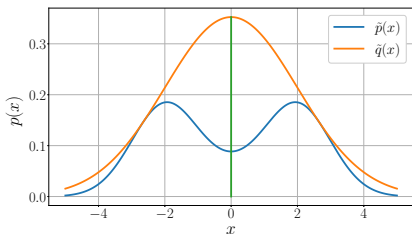
$\xi$  – непрерывная случайная величина, тогда  $\eta = F_\xi(\xi) \sim U[0, 1]$ .

Генерируем  $x_1, \dots, x_m$  как  $x_i = F_\xi^{-1}(y_i)$ , где  $y_i \sim U[0, 1]$ .

Выборка с отклонением (rejection sampling)

**Замечание:**  $p(\mathbf{Z}) \propto \tilde{p}(\mathbf{Z})$  известно с точностью до константы.

Пусть  $q(\mathbf{Z})$  – некоторое предположное (proposal) распределение и  $\tilde{p}(\mathbf{Z}) \leq \tilde{q}(\mathbf{Z}) = \alpha q(\mathbf{Z})$ .



- Сгенерируем выборку  $\mathbf{x}_1, \dots, \mathbf{x}_n$  из  $q(\mathbf{Z})$ ;
- Сгенерируем  $t_1, \dots, t_n \sim U[0, 1]$ ;
- Принимаем те точки выборки, где  $t_i < \frac{\tilde{p}(\mathbf{x}_i)}{\tilde{q}(\mathbf{x}_i)}$ .

**Вопрос 1:** при каких условиях rejection sampling эффективен?

**Вопрос 2:** как сэмплировать из многомерного нормального распределения  $N(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ , если есть генератор из  $N(0, 1)$ ?

Пусть имеется однородная марковская цепь с функцией плотности вероятности перехода между состояниями  $q(\mathbf{Z}_{i+1}|\mathbf{Z}_i)$ .

- Возьмем некоторое  $p_0(\mathbf{Z})$  и сгенерируем  $\mathbf{Z}_0 \sim p_0(\mathbf{Z})$ ;
- Генерируем  $\mathbf{Z}_{i+1} \sim q(\mathbf{Z}_{i+1}|\mathbf{Z}_i)$ ,  $i = 0, 1, \dots$ ;
- Выбрасываем первые  $m_0$  наблюдений (и прореживаем, если нужна НОР (i.i.d) выборка).

**Вопрос:** при каких условиях такая схема приведет к получению выборки из  $p(\mathbf{Z})$  ?

**Условие 1:**  $p(\mathbf{Z})$  инвариантно относительно цепи, то есть

$$p(\mathbf{Z}_{i+1}) = \int p(\mathbf{Z}_i)q(\mathbf{Z}_{i+1}|\mathbf{Z}_i)d\mathbf{Z}_i \text{ (стационарное распределение).}$$

**Достаточное условие:**  $p(\mathbf{Z}_{i+1})q(\mathbf{Z}_i|\mathbf{Z}_{i+1}) = p(\mathbf{Z}_i)q(\mathbf{Z}_{i+1}|\mathbf{Z}_i)$ .

**Условие 2:** цепь эргодична, то есть стационарное распределение не зависит от начальных условий  $\forall p_0(\mathbf{Z}) p_i(\mathbf{Z}_i) \rightarrow p(\mathbf{Z})$  при  $i \rightarrow \infty$ .

**Достаточное условие:**  $\forall s \forall \mathbf{t} : p(\mathbf{t}) \neq 0 q(\mathbf{t}|s) > 0$ .

# Схема Гиббса (Gibbs)

$$p(\mathbf{Z}) \propto \tilde{p}(\mathbf{Z}), \mathbf{Z} \in \mathbb{R}^n.$$

Считаем, что одномерные условные распределения  $p(z_j | \mathbf{Z}_{\setminus j})$  легко нормируемы.

■ Имеем  $\mathbf{Z}_i$ , хотим получить  $\mathbf{Z}_{i+1}$ ;

■  $z_{i+1}^1 \sim p(z^1 | z_i^2, \dots, z_i^n);$

$$z_{i+1}^2 \sim p(z^2 | z_{i+1}^1, z_i^3, \dots, z_i^n);$$

...

$$z_{i+1}^n \sim p(z^n | z_{i+1}^1, z_{i+1}^2, \dots, z_{i+1}^{n-1}).$$

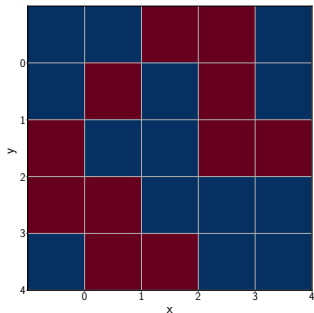
**Упражнение:** доказать инвариантность  $p(\mathbf{Z})$  относительно такой марковской цепи.

**Hint:** доказать по индукции, что сэмплирование одной компоненты сохраняет  $p(\mathbf{Z})$ .

# Модель Изинга (Ising model)

Пусть в каждой точке есть магнитный момент  $x_i \in \{\pm 1\}$  и  $p(\mathbf{x}) = \frac{1}{Z} \exp(-E(\mathbf{x})/T)$ , где  $E(\mathbf{x}) = - \sum_{(i,j) \in \epsilon} x_i x_j - \sum_i h_i x_i$ .

Реализация магнитных моментов



Намагниченность:  $\mu = \left| \frac{1}{N} \sum_i x_i \right|$ ,

где  $N$  – число атомов в решетке.

Вопрос 1: как оценить среднюю

намагниченность:  $E_p \mu$ ?

Вариационное приближение

$$p(\mathbf{x}) \approx q(\mathbf{x}) = \prod_{i=1}^N q_i(x_i).$$

$$\log q_i(x_i) \propto E_{q_{\setminus i}} \log p(\mathbf{x}) = -\frac{1}{T} E_{q_{\setminus i}} E(\mathbf{x}).$$

$$q_i(x_i = 1) = \frac{1}{1 + \exp\left(-\frac{2}{T}(h_i + \sum_{j: (i,j) \in \epsilon} E_{q_j} x_j)\right)}.$$

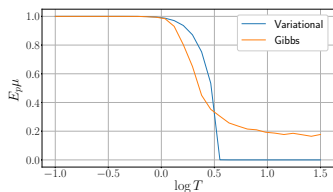
Вопрос 2: насколько хороша вариационная аппроксимация?

Вопрос 3: какую альтернативу можно предложить?

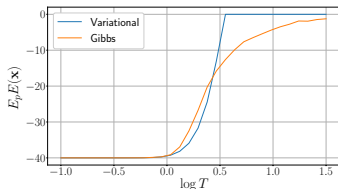


# Сравнение вариационной аппроксимации и схемы Гиббса

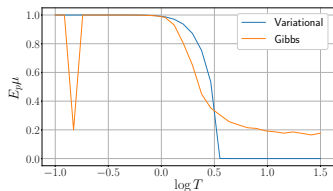
Намагниченность с температурой



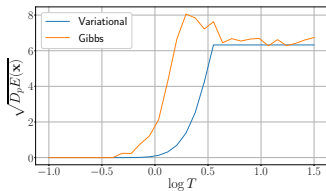
Средняя энергия с температурой



Намагниченность с температурой (проблема)



Стандартная ошибка энергии с температурой

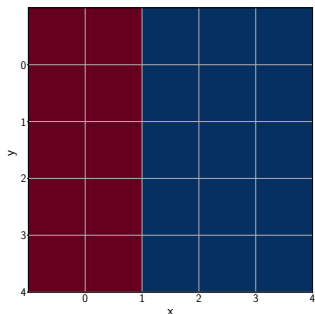


Вопрос: что вызывает провал графика намагнетизации?

## Свойства:

- (+) Подходит и для дискретных, и для непрерывных распределений;
- (+) Нет настраиваемых параметров;
- (-) Неэффективна в пространствах большой размерности;
- (-) Возможна очень долгая сходимость цепи к стационарному распределению.

## Реализация магнитных моментов



**Вопрос:** что произойдет, если начальный элемент MCMC такой и  $T \ll 1$ ?

# Схема Метрополиса-Хастингса (Metropolis-Hastings)

$p(\mathbf{Z}) \propto \tilde{p}(\mathbf{Z})$ ,  $r(\mathbf{Z}|\mathbf{Z}_i)$  – предположеное распределение.

- Имеем  $\mathbf{Z}_i$ , сэмплируем  $\mathbf{Z}^* \sim r(\mathbf{Z}|\mathbf{Z}_i)$ ;
- Вычисляем  $P(\mathbf{Z}^*, \mathbf{Z}_i) = \min\left(1, \frac{\tilde{p}(\mathbf{Z}^*)r(\mathbf{Z}_i|\mathbf{Z}^*)}{\tilde{p}(\mathbf{Z}_i)r(\mathbf{Z}^*|\mathbf{Z}_i)}\right)$
- $\mathbf{Z}_{i+1} = \mathbf{Z}^*$  с вероятностью  $P(\mathbf{Z}^*, \mathbf{Z}_i)$ ,  
 $\mathbf{Z}_{i+1} = \mathbf{Z}_i$  с вероятностью  $1 - P(\mathbf{Z}^*, \mathbf{Z}_i)$ .

Отсюда  $q(\mathbf{Z}_{n+1}|\mathbf{Z}_n) = \begin{cases} r(\mathbf{Z}_{n+1}|\mathbf{Z}_n)P(\mathbf{Z}_{n+1}, \mathbf{Z}_n), & \mathbf{Z}_{n+1} \neq \mathbf{Z}_n, \\ 1 - \int r(\mathbf{Z}^*|\mathbf{Z}_n)P(\mathbf{Z}^*, \mathbf{Z}_n)d\mathbf{Z}^*, & \mathbf{Z}_{n+1} = \mathbf{Z}_n. \end{cases}$

**Достаточное условие эргодичности:**  $\forall s \forall t : \tilde{p}(t) > 0, q(t|s) > 0$ .

**Замечание 1:** для выполнения этого требования достаточно  $r(t|s) > 0 \forall s \forall t$ .

**Достаточное условие инвариантности:**  $\forall s \forall t \tilde{p}(s)q(t|s) = \tilde{p}(t)q(s|t)$ .

**Замечание 2:** Убеждаемся в выполнении условия подстановкой.

Для  $s = t$  очевидно. Пусть  $s \neq t$ , тогда  $\tilde{p}(s)q(t|s) = \tilde{p}(s)r(t|s) \min\left(1, \frac{\tilde{p}(t)r(s|t)}{\tilde{p}(s)r(t|s)}\right) = \min(\tilde{p}(s)r(t|s), \tilde{p}(t)r(s|t)) = \tilde{p}(t)q(s|t)$ .

# Схема Метрополиса-Хастингса (продолжение)

- Имеем  $\mathbf{Z}_i$ , сэмплируем  $\mathbf{Z}^* \sim r(\mathbf{Z}|\mathbf{Z}_i)$ ;
- Вычисляем  $P(\mathbf{Z}^*, \mathbf{Z}_i) = \min\left(1, \frac{\tilde{p}(\mathbf{Z}^*)r(\mathbf{Z}_i|\mathbf{Z}^*)}{\tilde{p}(\mathbf{Z}_i)r(\mathbf{Z}^*|\mathbf{Z}_i)}\right)$
- $\mathbf{Z}_{i+1} = \mathbf{Z}^*$ ,  $P(\mathbf{Z}^*, \mathbf{Z}_i)$ ,  
 $\mathbf{Z}_{i+1} = \mathbf{Z}_i$ ,  $1 - P(\mathbf{Z}^*, \mathbf{Z}_i)$ .

Если  $r(\mathbf{Z}^*|\mathbf{Z}) = r(\mathbf{Z}|\mathbf{Z}^*)$ , то  $P(\mathbf{Z}^*, \mathbf{Z}_i) = \min\left(1, \frac{\tilde{p}(\mathbf{Z}^*)}{\tilde{p}(\mathbf{Z}_i)}\right)$ .

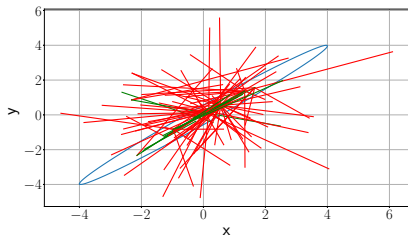
## Пример

$$p(\mathbf{x}) = N(\mathbf{0}, \sigma^2 \begin{pmatrix} 1 & 0.95 \\ 0.95 & 1 \end{pmatrix}),$$

$$\sigma = 2,$$

$$r(\mathbf{Z}^*|\mathbf{Z}) = N(\mathbf{Z}^*|\mathbf{Z}, \sigma^2\mathbf{I}).$$

## Результат сэмплирования



- 1 Bishop, Christopher M. "Pattern recognition and machine learning". Springer, New York (2006). Pp. 523-556.
- 2 MacKay, David JC. Bayesian methods for adaptive models. Diss. California Institute of Technology, 1992.
- 3 MacKay, David JC. "The evidence framework applied to classification networks." *Neural computation* 4.5 (1992): 720-736.
- 4 Gelman, Andrew, et al. Bayesian data analysis, 3rd edition. Chapman and Hall/CRC, 2013.
- 5 Дрейпер, Норман Р. Прикладной регрессионный анализ. Рипол Классик, 2007.