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

Александр Адуенко

28е ноября 2023

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 $\mathbf{w}_{\mathrm{ML}}$, регуляризации и $\mathbf{w}_{\mathrm{MAP}}$.
- 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 $\mathbf{w}$ 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 мультимоделей.

EM и вариационный EM-алгоритм (воспоминание)

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

Вопрос 1: Как решить задачу $p(\mathbf{D}|\boldsymbol{\Theta}) = \displaystyle\int p(\mathbf{D},\, \mathbf{Z}|\boldsymbol{\Theta})d\mathbf{Z} \to \max_{\boldsymbol{\Theta}}$?

EM-алгоритм

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

Идея 1: $p(\mathbf{D}|\boldsymbol{\Theta}) \to \max\limits_{\boldsymbol{\Theta}}$ заменим на $F(q,\, \boldsymbol{\Theta}) \to \max\limits_{q,\, \boldsymbol{\Theta}}$.

Идея 2: Пошагово оптимизируем по $\boldsymbol{\Theta}$ и q, то есть

1. Е-шаг: $q^s = F(q,\, \boldsymbol{\Theta}^{s-1}) \to \max\limits_{q \in Q}$;
2. М-шаг: $\boldsymbol{\Theta}^s = F(q^s,\, \boldsymbol{\Theta}) \to \max\limits_{\boldsymbol{\Theta}}$.

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

Прогноз:

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

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

Необходимость сэмплирования

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

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

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

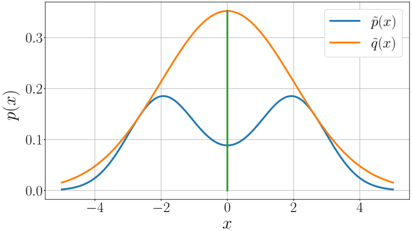
Вопрос 2: Что делать для решения задачи $p(\mathbf{Z}) \to \max\limits_{\mathbf{Z}}$, если $p(\mathbf{Z})$ известно, но задача максимизации аналитически не решается?

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

ξ – непрерывная случайная величина, тогда $\eta = F_\xi(\xi) \sim U[0,\, 1]$.
Генерируем $x_1, \ldots, x_m$ 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 предложное (proposal) распределение и $\tilde{p}(\mathbf{Z}) \leq \tilde{q}(\mathbf{Z}) = \alpha q(\mathbf{Z})$.



- Сгенерируем выборку $\mathbf{x}_1, \ldots, \mathbf{x}_n$ из $q(\mathbf{Z})$;
- Сгенерируем $t_1, \ldots, t_n \sim U[0,\, 1]$;
- 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 $\mathcal{N}(\boldsymbol{\mu},\, \boldsymbol{\Sigma})$, если есть генератор из $\mathcal{N}(0,\, 1)$?

Идея MCMC

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 $\mathbf{Z}_{i+1} \sim q(\mathbf{Z}_{i+1}|\mathbf{Z}_i)$, $i = 0,\ 1,\ \ldots$;
- 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

$p(\mathbf{Z}_{i+1}) = \displaystyle\int p(\mathbf{Z}_i)q(\mathbf{Z}_{i+1}|\mathbf{Z}_i)d\mathbf{Z}_i$ (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 $\forall\, p_0(\mathbf{Z})\ p_i(\mathbf{Z}_i) \to p(\mathbf{Z})$ при $i \to \infty$.

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

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 $\mathbf{Z}_{i+1}$;
- $z_{i+1}^1 \sim p(z^1|z_i^2, \ldots, z_i^n)$;
  $z_{i+1}^2 \sim p(z^2|z_{i+1}^1, z_i^3, \ldots, z_i^n)$;
  $\ldots$
  $z_{i+1}^n \sim p(z^n|z_{i+1}^1, z_{i+1}^2, \ldots, z_{i+1}^{n-1})$.

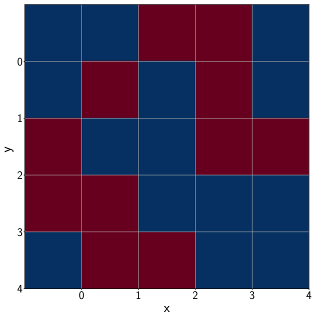
Упражнение: доказать инвариантность $p(\mathbf{Z})$ относительно такой марковской цепи.

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 $E(\mathbf{x}) = -\sum_{(i,\,j)\in\varepsilon} x_i x_j - \sum_i h_i x_i$.

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: $\mathsf{E}_p \mu$?

Вариационное приближение

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

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

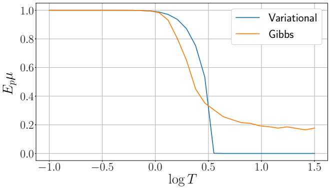
$$q_i(x_i = 1) = \frac{1}{1+\exp\left(-\frac{2}{T}\left(h_i + \sum_{j:\,(i,\,j)\in\varepsilon} \mathsf{E}_{q_j} x_j\right)\right)}.$$

Вопрос 2: Насколько хороша вариационная аппроксимация?

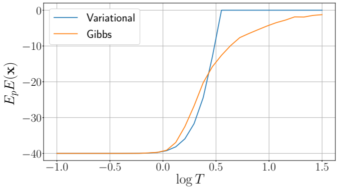
Вопрос 3: Какую альтернативу можно предложить?

Сравнение вариационной аппроксимации и схемы Гиббса

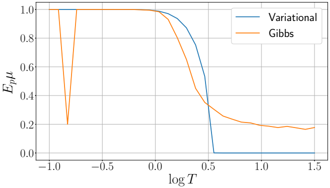
Намагниченность с температурой



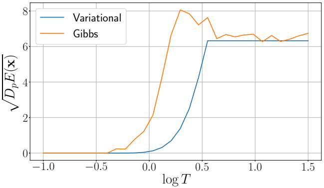
Средняя энергия с температурой



Намагниченность с температурой (проблема)



Стандартная ошибка энергии с температурой



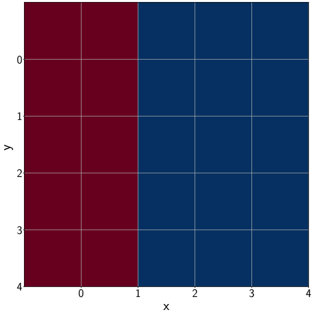
Вопрос: Что вызывает провал графика магнетизации?

Проблемы и преимущества схемы Гиббса

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)

$p(\mathbf{Z}) \propto \tilde{p}(\mathbf{Z})$, $r(\mathbf{Z}|\mathbf{Z}_i)$ – предложное 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: $\forall\, \mathbf{s}\, \forall\, \mathbf{t} : \tilde{p}(\mathbf{t}) > 0,\ q(\mathbf{t}|\mathbf{s}) > 0$.
Замечание 1: для выполнения этого требования достаточно $r(\mathbf{t}|\mathbf{s}) > 0\, \forall\, \mathbf{s} \forall\, \mathbf{t}$.

Достаточное условие инвариантности: $\forall\, \mathbf{s}\, \forall\, \mathbf{t}\ \tilde{p}(\mathbf{s})q(\mathbf{t}|\mathbf{s}) = \tilde{p}(\mathbf{t})q(\mathbf{s}|\mathbf{t})$.
Замечание 2: Убеждаемся в выполнении условия подстановкой.
Для $\mathbf{s} = \mathbf{t}$ очевидно. Пусть $\mathbf{s} \neq \mathbf{t}$, тогда $\tilde{p}(\mathbf{s})q(\mathbf{t}|\mathbf{s}) =$
$\tilde{p}(\mathbf{s})r(\mathbf{t}|\mathbf{s}) \min\left(1,\, \frac{\tilde{p}(\mathbf{t})r(\mathbf{s}|\mathbf{t})}{\tilde{p}(\mathbf{s})r(\mathbf{t}|\mathbf{s})}\right) = \min(\tilde{p}(\mathbf{s})r(\mathbf{t}|\mathbf{s}),\, \tilde{p}(\mathbf{t})r(\mathbf{s}|\mathbf{t})) = \tilde{p}(\mathbf{t})q(\mathbf{s}|\mathbf{t})$.

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

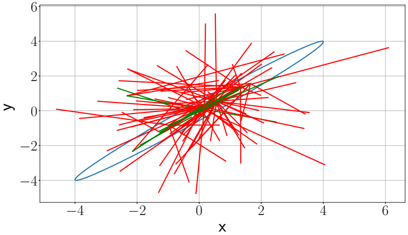
Пример

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

$\sigma = 2$,

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

Результат сэмплирования



Литература

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.
6. Carlo, Chain Monte. "Markov chain monte carlo and gibbs sampling." Lecture notes for EEB 581 (2004): 540.
7. Andrieu, Christophe, Arnaud Doucet, and Roman Holenstein. "Particle markov chain monte carlo methods." Journal of the Royal Statistical Society: Series B (Statistical Methodology) 72.3 (2010): 269-342.