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

Александр Адуенко

28е ноября 2018

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 $\mathbf{w}_{\text{ML}}$, регуляризации и $\mathbf{w}_{\text{MAP}}$.
- 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 логистической регрессии.
- Гауссовские процессы. Учёт эволюции моделей во времени.

Смесь моделей логистической регрессии

Вероятностная модель генерации данных

- Веса моделей в смеси $\boldsymbol{\pi}$ получены из априорного распределения $p(\boldsymbol{\pi}|\mu)$;
- Векторы параметров моделей $\mathbf{w}_k$ получены из нормального распределения $p(\mathbf{w}_k|\mathbf{A}_k) = \mathcal{N}(\mathbf{w}_k|\mathbf{0},\ \mathbf{A}_k^{-1})$, $k = 1, \ldots, K$;
- Для каждого объекта $\mathbf{x}_i$ выбрана модель f_{k_i}, которой он описывается, причем $p(k_i = k) = \pi_k$;
- Для каждого объекта $\mathbf{x}_i$ класс y_i определен в соответствии с моделью f_{k_i}: $y_i \sim \mathrm{Be}\big(\sigma(\mathbf{w}_{k_i}^{\mathsf{T}}\mathbf{x}_i)\big)$.

Совместное правдоподобие модели

$$p(\mathbf{y},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \boldsymbol{\pi}|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K,\ \boldsymbol{\mu}) =$$
$$p(\boldsymbol{\pi}|\boldsymbol{\mu}) \prod_{k=1}^{K} N(\mathbf{w}_k|\mathbf{0},\ \mathbf{A}_k^{-1}) \prod_{i=1}^{m} \left(\sum_{l=1}^{K} \pi_l \sigma(y_i \mathbf{w}_l^{\mathsf{T}} \mathbf{x}_i) \right).$$

Вопрос: как получить $p(\mathbf{w}_1,\ \ldots,\ \mathbf{w}_k,\ \boldsymbol{\pi}|\mathbf{X},\ \mathbf{y},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K,\ \boldsymbol{\mu})$?

$$p(\mathbf{y},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \boldsymbol{\pi}|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K,\ \boldsymbol{\mu}) \propto$$
$$\prod_{k=1}^{K} \pi_k^{\mu_k - 1} \prod_{k=1}^{K} \sqrt{\det \mathbf{A}_k} \exp(-\tfrac{1}{2}\mathbf{w}_k^{\mathsf{T}} \mathbf{A}_k \mathbf{w}) \prod_{i=1}^{m} \left(\sum_{l=1}^{K} \pi_l \sigma(y_i \mathbf{w}_l^{\mathsf{T}} \mathbf{x}_i) \right).$$

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

- Найти $\mathsf{P}(f(\mathbf{Z}) > 0)$;
  - $\mathsf{P}(\mathbf{w}^\top \mathbf{x} > B)$, где $\mathbf{w}$ – веса признаков, $\mathbf{x}$ – признаковое описание, а $\mathbf{w}^\top \mathbf{x}$ – ожидаемый доход.
- $\mathsf{E}f(\mathbf{Z}) = \displaystyle\int f(\mathbf{Z})p(\mathbf{Z})d\mathbf{Z} \approx \frac{1}{m}\sum_{i=1}^{m} f(\mathbf{Z}_i)$;
  - $p(\mathbf{y}|\mathbf{X},\ \mathbf{A}) = \displaystyle\int p(\mathbf{y}|\mathbf{X},\ \mathbf{w})p(\mathbf{w}|\mathbf{A})d\mathbf{w}$ в лог. регрессии;
  - $\mathsf{E}_{q(\mathbf{Z})} \log p(\mathbf{X},\ \mathbf{Z}|\boldsymbol{\Theta})$ на 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 $p(\mathbf{Z}) \to \max\limits_{\mathbf{Z}}$, если $p(\mathbf{Z})$ известно, но задача максимизации аналитически не решается?

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

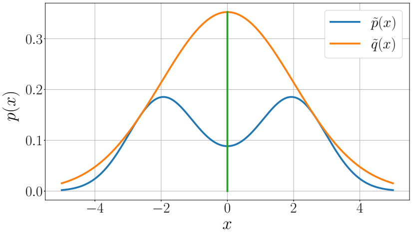
ξ – непрерывная случайная величина, тогда $\eta = F_\xi(\xi) \sim U[0,\ 1]$.

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 распределеня $N(\boldsymbol{\mu},\ \boldsymbol{\Sigma})$, если есть генератор из $N(0,\ 1)$?

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

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

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

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 Метрополиса-Хастингса

$p(\mathbf{Z}) \propto \tilde{p}(\mathbf{Z})$, $r(\mathbf{Z}|\mathbf{Z}_i)$ – 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 $\tilde{p}(\mathbf{s})q(\mathbf{t}|\mathbf{s}) = \tilde{p}(\mathbf{s})q(\mathbf{t}|\mathbf{s}) =$
$\tilde{p}(\mathbf{s})r(\mathbf{t}|\mathbf{s})\min\left(1, \frac{\tilde{p}(\mathbf{t})r(\mathbf{s}|\mathbf{t})}{\tilde{p}(\mathbf{s})r(\mathbf{t}|\mathbf{s})}\right) = \min(\tilde{p}(\mathbf{s})r(\mathbf{t}|\mathbf{s}), \tilde{p}(\mathbf{t})r(\mathbf{s}|\mathbf{t})) = \tilde{p}(\mathbf{t})q(\mathbf{s}|\mathbf{t})$.

$p(\mathbf{Z}) \propto \tilde{p}(\mathbf{Z})$, $\mathbf{Z} \in \mathbb{R}^n$.
Считаем, что одномерные условные распределения $p(z_j|\mathbf{Z}_{\backslash j})$ 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 $p(\mathbf{Z})$.

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.