

Байесовский выбор моделей: байесовская оптимизация с помощью гауссовских процессов

Александр Адуенко

8е декабря 2021

Содержание предыдущих лекций

- Формула Байеса и формула полной вероятности;
- Определение априорных вероятностей и selection bias;
- (Множественное) тестирование гипотез
- Экспоненциальное семейства. Достаточные статистики.
- Наивный байес. 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:

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

- 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.
- Сэмплирование. Схема Гиббса и Метрополиса-Хастингса. НМС.

Информация по зачету

- Письменный зачет (150 баллов): 15го декабря во время лекции*;
- Устный зачет (150 баллов): даты будут определены дополнительно;
- Вместо устного зачета можно сделать доклад по одной из четырех тем (по одному человеку на тему)
 - 1 Как выбрать априорное распределение? Неинформативные распределения. Распределение Джеффриса (Jeffreys prior).
 - 2 Методы детектирования разладки (changepoint detection) в характеристиках временного ряда. Построение байесовской модели простейшей разладки.
 - 3 Парадокс Штейна (Stein paradox). Оценка Джеймса-Штейна (Jame-Stein estimator) и ее обобщения.
 - 4 Взаимная информация (Mutual information) как мера связи между случайными величинами. Связь с линейной и ранговой корреляцией. Интерпретация и примеры применения на практике.
- Оценивание: нужное количество баллов на каждую оценку будет объявлено дополнительно.

Гауссовские процессы

$x(t)$ – температура в центре Кито.

Идея: $GP(m_x(t), R_x(\tau))$, где $m_x(t) \equiv m$, $R_x(\tau) = \sigma^2 \exp(-\lambda|\tau|)$.

Рассмотрим t_1, \dots, t_q , тогда для GP имеем

$p(\mathbf{x}) = p(x(t_1), \dots, x(t_q)) = \mathcal{N}(\mathbf{m}, \Sigma)$, где

$\mathbf{m} = [m_x(t_1), \dots, m_x(t_q)]^\top$, $\Sigma = \|\Sigma_{ij}\| = \|R_x(t_i - t_j)\|$.

Упражнение. $\mathbf{x} = [\mathbf{x}_1^\top, \mathbf{x}_2^\top]^\top \sim \mathcal{N}\left(\mathbf{x} | [\boldsymbol{\mu}_1^\top, \boldsymbol{\mu}_2^\top]^\top, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{12}^\top & \Sigma_{22} \end{pmatrix}^{-1}\right)$.

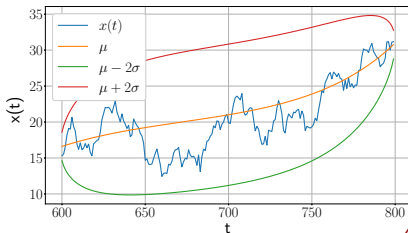
$\mathbf{x}_2 | \mathbf{x}_1 \sim \mathcal{N}(\mathbf{x}_2 | \boldsymbol{\mu}_2 - \Sigma_{22}^{-1} \Sigma_{12}(\mathbf{x}_1 - \boldsymbol{\mu}_1), \Sigma_{22}^{-1})$.

Вопрос 1: что делать, если неизвестно m , где $\boldsymbol{\mu}_1 = m\mathbf{e}_1$, $\boldsymbol{\mu}_2 = m\mathbf{e}_2$?

Вопрос 2: что делать, если неизвестны σ^2 и λ ?

Возможные модификации:

- Непостоянное $m_x(t)$;
- Введение разрывности $R_x(\tau) = \sigma^2(\exp(-\lambda|\tau|) + \kappa * [\tau = 0])$;
- Другая форма $R_x(\tau)$;
- $R_x(\tau) \rightarrow R_x(t_1, t_2)$.



Обозначим $r = \|x_1 - x_2\|$.

- $K(x_1, x_2) = \sigma^2 \exp(-\tau r^2)$ (RBF);
- $K(x_1, x_2) = \sigma^2 \exp(-\tau r)$ (Laplace);
- $K(x_1, x_2) = \sigma^2 \left(1 + \sqrt{3}r/l\right) \exp\left(-\sqrt{3}r/l\right)$ (Mattern 3/2);
- $K(x_1, x_2) = \sigma^2 \left(1 + \sqrt{5}r/l + \frac{5}{3}r^2/l^2\right) \exp\left(-\sqrt{5}r/l\right)$ (Mattern 5/2);
- $K(x_1, x_2) = \sigma^2 \exp\left(-2\frac{\sin^2(\pi r)}{l^2}\right)$ (Periodic);
- $K(x_1, x_2) = \sum_i \sigma_i^2 x_1^i x_2^i$ (Linear).

Вопрос 1: Как выбрать ядро? Какие функции задаёт каждое из вышеперечисленных?

Вопрос 2: Как получить ядро, отличное от вышеперечисленных?

Вопрос 3: Как можно использовать гауссовские процессы?

Требуется: $f(\Theta) \rightarrow \max_{\Theta \in \Theta}$, $\Theta \subseteq \mathbb{R}^n$, если

- Дана функция «черный ящик» $\tilde{f}_i(\Theta) = f(\Theta) + \varepsilon_i$;
- Подсчет функции \tilde{f} – дорог.

Пример: $f(\Theta) = p(\mathbf{y}|\mathbf{X}, \Theta) = \int p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\Theta)d\mathbf{w}$.

Прогноз:

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

$$p(\mathbf{w}|\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}, \Theta) = \frac{p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\Theta)}{p(\mathbf{y}|\mathbf{X}, \Theta)}.$$

Замечание: При известном Θ для получения прогноза используем

- Честный posterior $p(\mathbf{w}|\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}, \Theta)$, когда есть сопряженность;
- Вариационную аппроксимацию;
- Сэмплирование $\mathbf{w}_1, \dots, \mathbf{w}_K \sim p(\mathbf{w}|\mathbf{X}_{\text{train}}, \mathbf{y}_{\text{train}}, \Theta)$.

Вопрос 1: Каковы преимущества вариационной аппроксимации и сэмплирования?

Вопрос 2: Как определить Θ ? Что есть \tilde{f} ?

Требуется: $f(\Theta) \rightarrow \max_{\Theta \in \Theta}$, $\Theta \subseteq \mathbb{R}^n$, если

- Дана функция «черный ящик» $\tilde{f}_i(\Theta) = f(\Theta) + \varepsilon_i$;
- Подсчет функции \tilde{f} – дорог.

Пример: $f(\Theta) = p(\mathbf{y}|\mathbf{X}, \Theta) = \int p(\mathbf{y}|\mathbf{X}, \mathbf{w})p(\mathbf{w}|\Theta)d\mathbf{w}$.

$$\tilde{f}(\Theta) = \frac{1}{N} \sum_{i=1}^N p(\mathbf{y}|\mathbf{X}, \mathbf{w}_i), \mathbf{w}_i \sim p(\mathbf{w}|\Theta).$$

Идея: Предположим, что априори $f(\Theta) \sim \text{GP}(\cdot|\mathbf{0}, K(\cdot, \cdot))$.

$$\tilde{f}(\Theta_i) = f(\Theta_i) + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(\varepsilon_i|0, \sigma^2).$$

Пусть известны $\tilde{\mathbf{f}} = \|\tilde{f}(\Theta_1), \dots, \tilde{f}(\Theta_m)\|$. Рассмотрим Θ^* .

Обозначим $\mathbf{K} = \|K(\Theta_i, \Theta_j)\|$, $\mathbf{k} = \|K(\Theta_i, \Theta^*)\|$, $K^{**} = K(\Theta^*, \Theta^*)$.

$f(\Theta^*)|\tilde{f}(\Theta_1), \dots, \tilde{f}(\Theta_m) \sim \mathcal{N}(f^*|\mu^*, \sigma^{*2})$, где

$$\mu^* = \sum_{i=1}^m \alpha_i K(\Theta^*, \Theta_i), \alpha = (\mathbf{K} + \sigma^2 \mathbf{I}_m)^{-1} \tilde{\mathbf{f}},$$

$$\sigma^{*2} = K^{**} - \mathbf{k}^{*\top} (\mathbf{K} + \sigma^2 \mathbf{I}_m)^{-1} \mathbf{k}^*.$$

Требуется: $f(\Theta) \rightarrow \max_{\Theta \in \Theta} f(\Theta)$, $\Theta \subseteq \mathbb{R}^n$, если

- Дана функция «черный ящик» $\tilde{f}_i(\Theta) = f(\Theta) + \varepsilon_i$;
- Подсчет функции \tilde{f} – дорог.

Схема байесовской оптимизации

- Выбираем исходную точку Θ_0 и вычисляем $\tilde{f}(\Theta_0)$;
- На шаге $l + 1$ строим $\mu(\Theta)$ и $\sigma^2(\Theta)$ с помощью условного нормального распределения по известному $\tilde{\mathbf{f}} = \|\tilde{f}(\Theta_0), \dots, \tilde{f}(\Theta_l)\|$;
- Имеем некоторую функцию полезности $u(\mu^*, \sigma^{*2})$, выбираем следующую Θ_{l+1} из условия $u(\mu^*(\Theta), \sigma^{*2}(\Theta)) \rightarrow \max_{\Theta}$.

Вопрос 1: Чем задача $u(\mu^*(\Theta), \sigma^{*2}(\Theta)) \rightarrow \max_{\Theta}$ лучше исходной?

Вопрос 2: Как выбрать функцию полезности $u(\cdot, \cdot)$?

Схема байесовской оптимизации (продолжение)

- Выбираем исходную точку Θ_0 и вычисляем $\tilde{f}(\Theta_0)$;
- На шаге $l + 1$ строим $\mu(\Theta)$ и $\sigma^2(\Theta)$ с помощью условного нормального распределения по известному $\tilde{\mathbf{f}} = \|\tilde{f}(\Theta_0), \dots, \tilde{f}(\Theta_l)\|$;
- Имеем некоторую функцию полезности $u(\mu^*, \sigma^{*2})$, выбираем следующую Θ_{l+1} из условия $u(\mu^*(\Theta), \sigma^{*2}(\Theta)) \rightarrow \max_{\Theta}$.

Замечания:

- $u(\mu^*(\Theta), \sigma^{*2}(\Theta))$ легко считается вместе с производными;
- При построении $u(\cdot, \cdot)$ требуется соблюсти баланс между устранением неопределенности (exploration) и поиском в районе ожидаемого максимума (exploitation).

Exploration: Выбираем Θ_{l+1} , где $\sigma^{*2}(\Theta_{l+1})$ велико;

Exploitation: Выбираем Θ_{l+1} , где $\mu(\Theta_{l+1})$ велико.

Вопрос 1: Как построить функцию полезности u ?

Вопрос 2: Как откалибровать параметры ядерной функции и дисперсию шума σ^2 ?

Примеры функции полезности

- Верхняя доверительная граница (GP upper confidence band)

$$u(\mu^*(\Theta), \sigma^{*2}(\Theta)) = \mu^*(\Theta) + \xi\sigma^*(\Theta);$$

- Макс. вероятность улучшения (Maximum probability of improvement)

$$u(\mu^*(\Theta), \sigma^{*2}(\Theta)) = P\left(f(\Theta) > \max_{i \in \overline{1, l}} f(\Theta_i)\right).$$

Если $\sigma^2 = 0$, то $u(\mu^*(\Theta), \sigma^{*2}(\Theta)) = \Phi((\mu^* - \max_{i \in \overline{1, l}} f(\Theta_i))/\sigma^*)$;

- Ожидаемое улучшение (expected improvement)

$$u(\mu^*(\Theta), \sigma^{*2}(\Theta)) = E \max\left(0, f(\Theta) - \max_{i \in \overline{1, l}} f(\Theta_i)\right);$$

Если $\sigma^2 = 0$, то

$$u(\mu^*(\Theta), \sigma^{*2}(\Theta)) = \sigma^* E \max\left(0, \xi - \frac{\mu^* - \max_{i \in \overline{1, l}} f(\Theta_i)}{\sigma^*}\right), \xi \sim \mathcal{N}(0, 1).$$

Вопрос 1: Какие еще функции полезности можно предложить?

Вопрос 2: Как откалибровать параметры ядерной функции и дисперсию шума σ^2 ?

Вопрос 3: Какие еще применения (кроме максимизации обоснованности) байесовской оптимизации вы видите?

Калибрация параметров ядерной функции и дисперсии шума

$$f(\Theta) \sim \text{GP}(\cdot | \mathbf{0}, K(\cdot, \cdot)).$$

$$\tilde{f}(\Theta_i) = f(\Theta_i) + \varepsilon_i, \varepsilon_i \sim \mathcal{N}(\varepsilon_i | 0, \sigma^2).$$

$$\tilde{\mathbf{f}} = \|\tilde{f}(\Theta_1), \dots, \tilde{f}(\Theta_m)\|, \mathbf{K} = \|K(\Theta_i, \Theta_j)\|.$$

$$\text{Тогда } \tilde{\mathbf{f}} \sim \mathcal{N}(\tilde{\mathbf{f}} | \mathbf{0}, \mathbf{K} + \sigma^2 \mathbf{I}_m).$$

Обозначим \mathbf{q} – параметры ядерной функции, $\mathbf{C}(\mathbf{q}, \sigma^2) = \mathbf{K} + \sigma^2 \mathbf{I}_m$.

Идея: Найдем \mathbf{q} , σ^2 из максимума правдоподобия наблюдаемых данных.

$$L(\mathbf{q}, \sigma^2) = -2 \log p(\tilde{\mathbf{f}}) = \log \det \mathbf{C}(\mathbf{q}, \sigma^2) + \tilde{\mathbf{f}}^T \mathbf{C}^{-1}(\mathbf{q}, \sigma^2) \tilde{\mathbf{f}} \rightarrow \min_{\mathbf{q}, \sigma^2}.$$

$$\frac{\partial L}{\partial q_i} = \text{tr} \left(\mathbf{C}^{-1} \frac{\partial \mathbf{C}}{\partial q_i} \right) - \tilde{\mathbf{f}}^T \mathbf{C}^{-1} \frac{\partial \mathbf{C}}{\partial q_i} \mathbf{C}^{-1} \tilde{\mathbf{f}} \quad (\text{то же для } \sigma^2).$$

Замечание 1: Для применения ML оценки требуется достаточная по размеру выборка.

Замечание 2: На параметры ядерной функции \mathbf{q} и дисперсию шума σ^2 можно ввести априорное распределение.

Замечание 3: Часто для максимизации обоснованности можно использовать вариационную аппроксимацию (максимизация нижней оценки), а уже для апостериорного распределения – сэмплирование в случае сильной зависимости между переменными.

- 1 Bishop, Christopher M. "Pattern recognition and machine learning". Springer, New York (2006). Pp. 78-88, 303-320.
- 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.