Байесовский выбор моделей: байесовская оптимизация с помощью гауссовских процессов

Александр Адуенко

12е декабря 2020

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 $\mathbf{w}_{\text{ML}}$, 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 лог. регрессии.
- Гауссовские процессы. Учёт эволюции моделей во времени.
- Построение адекватных мультимоделей.
- Сэмплирование. Схема Гиббса и Метрополиса-Хастингса. HMC.

Информация по зачету

- Письменный зачет (150 баллов): 16го 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 $t_1, \ldots, t_q$, тогда для GP имеем

$p(\mathbf{x}) = p(x(t_1), \ldots, x(t_q)) = \mathrm{N}(\mathbf{m}, \boldsymbol{\Sigma})$, где

$\mathbf{m} = [m_x(t_1), \ldots, m_x(t_q)]^\mathsf{T}$, $\boldsymbol{\Sigma} = \|\boldsymbol{\Sigma}_{ij}\| = \|R_x(t_i - t_j)\|$.

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

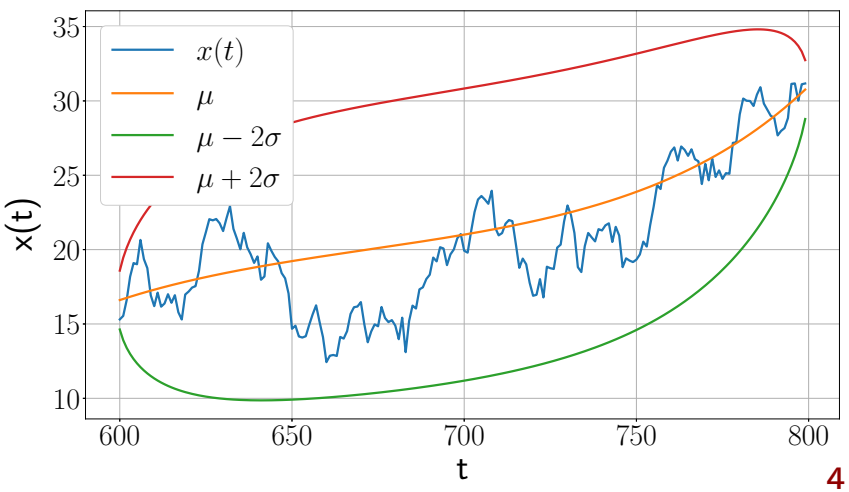
$\mathbf{x}_2|\mathbf{x}_1 \sim N(\mathbf{x}_2|\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}(\mathbf{x}_1 - \boldsymbol{\mu}_1),\, \boldsymbol{\Sigma}_{22}^{-1})$.

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 $R_x(\tau)$;
- $R_x(\tau) \to R_x(t_1, t_2)$.



Примеры ядер

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);
- $K(x_1,\, x_2) = \sigma^2 \exp\left(-2 \dfrac{\sin^2(\pi r)}{l^2}\right)$ (Periodic);
- $K(x_1,\, x_2) = \sum\limits_i \sigma_i^2 x_1^i x_2^i$ (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?

Гауссовские процессы для оптимизации

Требуется: $f(\boldsymbol{\Theta}) \to \max\limits_{\boldsymbol{\Theta} \in \Theta}$, $\Theta \subseteq \mathbb{R}^n$, если

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

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

Прогноз:

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

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

Замечание: При известном $\boldsymbol{\Theta}$ для получения прогноза используем

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

Вопрос 1: Каковы преимущества вариационной аппроксимации и сэмплирования?

Вопрос 2: Как определить $\boldsymbol{\Theta}$? Что есть $\tilde{f}$?

Максимизация обоснованности

Требуется: $f(\boldsymbol{\Theta}) \to \max\limits_{\boldsymbol{\Theta} \in \Theta}$, $\Theta \subseteq \mathbb{R}^n$, если

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

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

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

Идея: Предположим, что априори $f(\boldsymbol{\Theta}) \sim \mathrm{GP}(\cdot|\mathbf{0}, K(\cdot, \cdot))$.
$\tilde{f}(\boldsymbol{\Theta}_i) = f(\boldsymbol{\Theta}_i) + \varepsilon_i$, $\varepsilon_i \sim N(\varepsilon_i|0, \sigma^2)$.
Пусть известны $\tilde{\mathbf{f}} = \|\tilde{f}(\boldsymbol{\Theta}_1), \ldots, \tilde{f}(\boldsymbol{\Theta}_m)\|$. Рассмотрим $\boldsymbol{\Theta}^*$.
Обозначим $\mathbf{K} = \|K(\boldsymbol{\Theta}_i, \boldsymbol{\Theta}_j)\|$, $\mathbf{k} = \|K(\boldsymbol{\Theta}_i, \boldsymbol{\Theta}^*)\|$, $K^{**} = K(\boldsymbol{\Theta}^*, \boldsymbol{\Theta}^*)$.
$f(\boldsymbol{\Theta}^*)|\tilde{f}(\boldsymbol{\Theta}_1), \ldots, \tilde{f}(\boldsymbol{\Theta}_m) \sim N(f^*|\mu^*, \sigma^{*2})$, где

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

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

Схема байесовской оптимизации

Требуется: $f(\mathbf{\Theta}) \to \max\limits_{\mathbf{\Theta} \in \Theta}$, $\Theta \subseteq \mathbb{R}^n$, если

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

Схема байесовской оптимизации

- Выбираем исходную точку $\mathbf{\Theta}_0$ и вычисляем $\tilde{f}(\mathbf{\Theta}_0)$;
- На шаге $l+1$ строим $\mu(\mathbf{\Theta})$ и $\sigma^2(\mathbf{\Theta})$ с помощью условного нормального распределения по известному $\tilde{\mathbf{f}} = \|\tilde{f}(\mathbf{\Theta}_0), \ldots, \tilde{f}(\mathbf{\Theta}_l)\|$;
- Имеем некоторую функцию полезности $u(\mu^*, \sigma^{*2})$, выбираем следующую $\mathbf{\Theta}_{l+1}$ из условия $u(\mu^*(\mathbf{\Theta}), \sigma^{*2}(\mathbf{\Theta})) \to \max\limits_{\mathbf{\Theta}}$.

Вопрос 1: Чем задача $u(\mu^*(\mathbf{\Theta}), \sigma^{*2}(\mathbf{\Theta})) \to \max\limits_{\mathbf{\Theta}}$ лучше исходной?

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

Схема байесовской оптимизации (продолжение)

- Выбираем исходную точку $\boldsymbol{\Theta}_0$ и вычисляем $\tilde{f}(\boldsymbol{\Theta}_0)$;
- На шаге $l+1$ строим $\mu(\boldsymbol{\Theta})$ и $\sigma^2(\boldsymbol{\Theta})$ с помощью условного нормального распределения по известному $\tilde{\mathbf{f}} = \|\tilde{f}(\boldsymbol{\Theta}_0), \ldots, \tilde{f}(\boldsymbol{\Theta}_l)\|$;
- Имеем некоторую функцию полезности $u(\mu^*, \sigma^{*2})$, выбираем следующую $\boldsymbol{\Theta}_{l+1}$ из условия $u(\mu^*(\boldsymbol{\Theta}), \sigma^{*2}(\boldsymbol{\Theta})) \to \max\limits_{\boldsymbol{\Theta}}$.

Замечания:

- $u(\mu^*(\boldsymbol{\Theta}), \sigma^{*2}(\boldsymbol{\Theta}))$ 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 $\boldsymbol{\Theta}_{l+1}$, где $\sigma^{*2}(\boldsymbol{\Theta}_{l+1})$ велико;
Exploitation: Выбираем $\boldsymbol{\Theta}_{l+1}$, где $\mu(\boldsymbol{\Theta}_{l+1})$ 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)
  $u(\mu^*(\boldsymbol{\Theta}),\, \sigma^{*2}(\boldsymbol{\Theta})) = \mu^*(\boldsymbol{\Theta}) + \xi\sigma^*(\boldsymbol{\Theta})$;
- Макс. вероятность улучшения (Maximum probability of improvement)
  $$u(\mu^*(\boldsymbol{\Theta}),\, \sigma^{*2}(\boldsymbol{\Theta})) = \mathsf{P}\left(f(\boldsymbol{\Theta}) > \max_{i\in\overline{1,\,l}} f(\boldsymbol{\Theta}_i)\right).$$
  Если $\sigma^2 = 0$, то $u(\mu^*(\boldsymbol{\Theta}),\, \sigma^{*2}(\boldsymbol{\Theta})) = \Phi((\mu^* - \max\limits_{i\in\overline{1,\,l}} f(\boldsymbol{\Theta}_i))/\sigma^*)$;
- Ожидаемое улучшение (expected improvement)
  $$u(\mu^*(\boldsymbol{\Theta}),\, \sigma^{*2}(\boldsymbol{\Theta})) = \mathsf{E}\max\left(0,\; f(\boldsymbol{\Theta}) - \max_{i\in\overline{1,\,l}} f(\boldsymbol{\Theta}_i)\right);$$
  Если $\sigma^2 = 0$, то
  $$u(\mu^*(\boldsymbol{\Theta}),\, \sigma^{*2}(\boldsymbol{\Theta})) = \sigma^*\mathsf{E}\max\left(0,\; \xi - \frac{\mu^* - \max_{i\in\overline{1,\,l}} f(\boldsymbol{\Theta}_i)}{\sigma^*}\right),\ \xi \sim N(0,\,1).$$

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

$f(\mathbf{\Theta}) \sim \mathrm{GP}(\cdot|\mathbf{0},\, K(\cdot,\,\cdot))$.
$\tilde{f}(\mathbf{\Theta}_i) = f(\mathbf{\Theta}_i) + \varepsilon_i,\ \varepsilon_i \sim N(\varepsilon_i|0,\,\sigma^2)$.
$\tilde{\mathbf{f}} = \|\tilde{f}(\mathbf{\Theta}_1),\,\ldots,\,\tilde{f}(\mathbf{\Theta}_m)\|$, $\mathbf{K} = \|K(\mathbf{\Theta}_i,\,\mathbf{\Theta}_j)\|$.
Тогда $\tilde{\mathbf{f}} \sim N(\tilde{\mathbf{f}}|\mathbf{0},\,\mathbf{K} + \sigma^2\mathbf{I}_m)$.
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.
$L(\mathbf{q},\,\sigma^2) = -2\log p(\tilde{\mathbf{f}}) = \log\det\mathbf{C}(\mathbf{q},\,\sigma^2) + \tilde{\mathbf{f}}^{\mathsf{T}}\mathbf{C}^{-1}(\mathbf{q},\,\sigma^2)\tilde{\mathbf{f}} \to \min\limits_{\mathbf{q},\,\sigma^2}$.

$\frac{\partial L}{\partial q_i} = \mathrm{tr}\left(\mathbf{C}^{-1}\frac{\partial\mathbf{C}}{\partial q_i}\right) - \tilde{\mathbf{f}}^{\mathsf{T}}\mathbf{C}^{-1}\frac{\partial\mathbf{C}}{\partial q_i}\mathbf{C}^{-1}\tilde{\mathbf{f}}$ (то же для σ^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.

Литература

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.