

Байесовский выбор моделей: гауссовские процессы для учета эволюции модели

Александр Адуенко

11е 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:

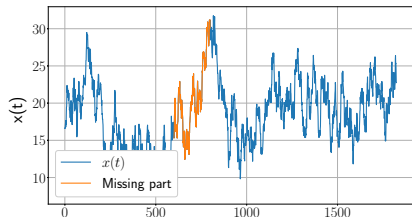
$$p(\mathbf{y}_{test} | \mathbf{X}_{test}, \mathbf{X}_{train}, \mathbf{y}_{train}) = \int p(\mathbf{y}_{test} | \mathbf{w}, \mathbf{X}_{test}) p(\mathbf{w} | \mathbf{X}_{train}, \mathbf{y}_{train}) d\mathbf{w}.$$

- 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. Использование EM-алгоритма для отбора признаков в байесовской линейной регрессии.
- Вариационный EM-алгоритм. Смесь моделей логистической регрессии.

Учет эволюции модели во времени

Пусть у объектов есть еще метка времени, то есть наблюдаем (x_i, y_i, t_i) . Ранее имели модель $p(y, \mathbf{w} | \mathbf{X}, \mathbf{A}) = p(y | \mathbf{X}, \mathbf{w})p(\mathbf{w} | \mathbf{A})$, то есть зависимостью от t пренебрегали.

Вопрос 1: как учесть наличие дополнительной информации?



Рассмотрим случайный процесс $x(t)$, $t \in T$.

$m_x(t) = \mathbb{E}x(t)$, $K_x(t, s) = \mathbb{E}x(t)x(s)$, $R_x(t, s) = \mathbb{E}\dot{x}(t)\dot{x}(s)$ – функция мат. ожидания, ковариационная и корреляционная функция.

Определение. С.п. называется слабо стационарным, если $m_x(t) \equiv m$, $R_x(t, s) = R_x(\tau = |t - s|)$.

Пример. Пусть $x(t)$ – температура в центре Кито.

Вопрос 2: Как восстановить пропущенные данные?

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

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

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

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

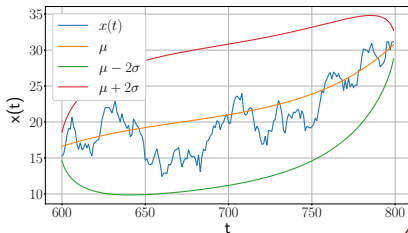
$\mathbf{x}_2 \mid \mathbf{x}_1 \sim N(\mathbf{x}_2 \mid \boldsymbol{\mu}_2 - \Sigma_{22}^{-1} \Sigma_{12}(\mathbf{x}_1 - \boldsymbol{\mu}_1), \Sigma_{22}^{-1})$.

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);
- $K(x_1, x_2) = \sigma^2 \left(1 + \sqrt{3}r/l\right) \exp\left(-\sqrt{3}r/l\right)$ (Matern 3/2);
- $K(x_1, x_2) = \sigma^2 \left(1 + \sqrt{5}r/l + \frac{5}{3}r^2/l^2\right) \exp\left(-\sqrt{5}r/l\right)$ (Matern 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?

Линейная регрессия с эволюцией во времени

Байесовская линейная регрессия

$(\mathbf{X}, \mathbf{y}) = \cup_{i=1}^m (\mathbf{x}_i, y_i)$ – выборка.

$y_i = \mathbf{w}^\top \mathbf{x}_i + \varepsilon_i$, $\varepsilon_i \sim N(\varepsilon_i | 0, \beta^{-1})$, $\mathbf{w} \sim N(\mathbf{w} | \mathbf{0}, \mathbf{A}^{-1})$.

$p(\mathbf{y}, \mathbf{w} | \mathbf{X}, \mathbf{A}, \beta) = p(\mathbf{w} | \mathbf{A}) p(\mathbf{y} | \mathbf{X}, \mathbf{w}, \beta)$.

Байесовская линейная регрессия с эволюцией

$(\mathbf{X}, \mathbf{y}, \mathbf{t}) = \cup_{i=1}^m (\mathbf{x}_i, y_i, t_i)$ – выборка.

Для простоты считаем $t_1 < t_2 < \dots < t_m$.

$y_i = \mathbf{w}_i^\top \mathbf{x}_i + \varepsilon_i$, $\varepsilon_i \sim N(\varepsilon_i | 0, \beta^{-1})$.

Введем матрицу $\mathbf{W} = [\mathbf{w}_1, \dots, \mathbf{w}_m]^\top = [\mathbf{v}_1, \dots, \mathbf{v}_n] \in \mathbb{R}^{m \times n}$.

Вопрос 1: как обучить модель, если у каждого объекта свой индивидуальный вектор параметров \mathbf{w}_i ?

Идея: Априори предположим, что \mathbf{v}_j получен как реализация GP $v_j(t)$.

$\mathbf{v}_j = [v_j(t_1), \dots, v_j(t_m)]^\top$, $K_{v_j}(t_l, t_k) = \alpha_j^{-1} \exp(-\lambda |t_l - t_k|)$.

Тогда $p(\mathbf{w}_1, \dots, \mathbf{w}_m) = \prod_{j=1}^n N(\mathbf{v}_j | \mathbf{0}, (\alpha_j \mathbf{K})^{-1})$, где

$\mathbf{K}^{-1} = \|\exp(-\lambda |t_l - t_k|)\|$.

Вопрос 2: что произойдет при $\lambda \rightarrow 0$?

Получение апостериорного распределения

Пусть дополнительно дана точка для прогноза $(\mathbf{x}_{m+1}, t_{m+1})$.

Найти: $p(\mathbf{w}_1, \dots, \mathbf{w}_m, \mathbf{w}_{m+1} | \mathbf{y}, \mathbf{X}, \mathbf{t}, \beta, \mathbf{A}, \lambda)$.

$$\log p(\mathbf{y}, \mathbf{w}_1, \dots, \mathbf{w}_m, \mathbf{w}_{m+1} | \mathbf{X}, \mathbf{t}, \beta, \mathbf{A}, \lambda) \propto$$
$$\frac{m}{2} \log \beta - \frac{\beta}{2} \sum_{i=1}^m (y_i - \mathbf{w}_i^\top \mathbf{x}_i)^2 + \sum_{j=1}^n \left[\frac{1}{2} \log \alpha_j + \frac{1}{2} \log \det \mathbf{K} - \frac{\alpha_j}{2} \mathbf{v}_j^\top \mathbf{K} \mathbf{v}_j \right].$$

$$\log p(\mathbf{w}_1, \dots, \mathbf{w}_m, \mathbf{w}_{m+1} | \mathbf{y}, \mathbf{X}, \mathbf{t}, \beta, \mathbf{A}, \lambda) \propto$$
$$-\frac{1}{2} \left[\sum_{j=1}^n \alpha_j \mathbf{v}_j^\top \mathbf{K} \mathbf{v}_j + \beta \sum_{i=1}^m \mathbf{w}_i^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{w}_i - 2\beta \sum_{i=1}^m y_i \mathbf{w}_i^\top \mathbf{x}_i \right].$$

Введем $\mathbf{u} = [\mathbf{w}_1, \dots, \mathbf{w}_{m+1}]^\top \in \mathbb{R}^{(m+1)n}$.

$$\mathbf{K}_1 = \begin{pmatrix} \beta \mathbf{x}_1 \mathbf{x}_1^\top & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \beta \mathbf{x}_2 \mathbf{x}_2^\top & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & & & & \\ \mathbf{0} & \mathbf{0} & \dots & \beta \mathbf{x}_m \mathbf{x}_m^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \mathbf{m} = \beta \begin{pmatrix} y_1 \mathbf{x}_1 \\ y_2 \mathbf{x}_2 \\ \vdots \\ y_m \mathbf{x}_m \\ \mathbf{0} \end{pmatrix}.$$

Получение апостериорного распределения

Введем $\mathbf{u} = [\mathbf{w}_1, \dots, \mathbf{w}_{m+1}]^\top \in \mathbb{R}^{(m+1)n}$.

$$\log p(\mathbf{w}_1, \dots, \mathbf{w}_m, \mathbf{w}_{m+1} | \mathbf{y}, \mathbf{X}, \mathbf{t}, \beta, \mathbf{A}, \lambda) \propto -\frac{1}{2} \left[\mathbf{u}^\top \boldsymbol{\Sigma}^{-1} \mathbf{u} - 2\mathbf{u}^\top \mathbf{m} \right] =$$
$$-\frac{1}{2} \left[\sum_{j=1}^n \alpha_j \mathbf{v}_j^\top \mathbf{K} \mathbf{v}_j + \beta \sum_{i=1}^m \mathbf{w}_i^\top \mathbf{x}_i \mathbf{x}_i^\top \mathbf{w}_i - 2\beta \sum_{i=1}^m y_i \mathbf{w}_i^\top \mathbf{x}_i \right].$$

$$\mathbf{K}_1 = \begin{pmatrix} \beta \mathbf{x}_1 \mathbf{x}_1^\top & \mathbf{0} & \dots & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \beta \mathbf{x}_2 \mathbf{x}_2^\top & \dots & \mathbf{0} & \mathbf{0} \\ \vdots & & & & \\ \mathbf{0} & \mathbf{0} & \dots & \beta \mathbf{x}_m \mathbf{x}_m^\top & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} & \mathbf{0} \end{pmatrix}, \quad \mathbf{m} = \beta \begin{pmatrix} y_1 \mathbf{x}_1 \\ y_2 \mathbf{x}_2 \\ \vdots \\ y_m \mathbf{x}_m \\ \mathbf{0} \end{pmatrix},$$

$$\mathbf{K}_2 = \begin{pmatrix} \mathbf{A}K_{11} & \mathbf{A}K_{12} & \dots & \mathbf{A}K_{1, m+1} \\ \mathbf{A}K_{21} & \mathbf{A}K_{22} & \dots & \mathbf{A}K_{2, m+1} \\ \vdots & & & \\ \mathbf{A}K_{m+1, 1} & \mathbf{A}K_{m+1, 2} & \dots & \mathbf{A}K_{m+1, m+1} \end{pmatrix}.$$

Отсюда $\mathbf{u} \sim N(\mathbf{u} | (\mathbf{K}_1 + \mathbf{K}_2)^{-1} \mathbf{m}, (\mathbf{K}_1 + \mathbf{K}_2)^{-1})$.

Отбор признаков и подбор ковариационной функции

Вопрос: как определить \mathbf{A} , β , λ ?

Рассмотрим задачу $p(\mathbf{y}|\mathbf{X}, \mathbf{t}, \beta, \mathbf{A}, \lambda) \rightarrow \max_{\beta, \mathbf{A}, \lambda}$.

Рассмотрим $\mathbf{Z} = (\mathbf{w}_1, \dots, \mathbf{w}_{m+1})$ и воспользуемся EM-алгоритмом.

E-шаг. $q(\mathbf{Z}) = p(\mathbf{w}_1, \dots, \mathbf{w}_m, \mathbf{w}_{m+1}|\mathbf{y}, \mathbf{X}, \mathbf{t}, \beta, \mathbf{A}, \lambda)$.

M-шаг. $E_q \log p(\mathbf{y}, \mathbf{w}_1, \dots, \mathbf{w}_m, \mathbf{w}_{m+1}|\mathbf{X}, \mathbf{t}, \beta, \mathbf{A}, \lambda) \rightarrow \max_{\beta, \mathbf{A}, \lambda}$.

$$\frac{m}{2} \log \beta - \frac{\beta}{2} \sum_{i=1}^m E(y_i - \mathbf{w}_i^\top \mathbf{x}_i)^2 + \frac{1}{2} \sum_{j=1}^n \log \alpha_j + \frac{n}{2} \log \det \mathbf{K} -$$

$$\frac{1}{2} \sum_{j=1}^n \alpha_j E \mathbf{v}_j^\top \mathbf{K} \mathbf{v}_j \rightarrow \max_{\beta, \mathbf{A}, \lambda}$$

$$\beta^{-1} = \frac{1}{m} \sum_{i=1}^m E(y_i - \mathbf{w}_i^\top \mathbf{x}_i)^2; \quad \alpha_j = \frac{1}{E \mathbf{v}_j^\top \mathbf{K} \mathbf{v}_j} = \frac{1}{\text{tr}(\mathbf{K} E \mathbf{v}_j \mathbf{v}_j^\top)}$$

Hint: $\alpha_j^{\text{new}} = \frac{1 - \alpha_j^{\text{old}} \text{tr}(\mathbf{K} E \hat{\mathbf{v}}_j \hat{\mathbf{v}}_j^\top)}{\text{tr}(\mathbf{K} (E \mathbf{v}_j) (E \mathbf{v}_j)^\top)}$.

Отбор признаков и подбор ковариационной функции

M-шаг. $E_q \log p(\mathbf{y}, \mathbf{w}_1, \dots, \mathbf{w}_m, \mathbf{w}_{m+1} | \mathbf{X}, \mathbf{t}, \beta, \mathbf{A}, \lambda) \rightarrow \max_{\beta, \mathbf{A}, \lambda}$.

$$\frac{m}{2} \log \beta - \frac{\beta}{2} \sum_{i=1}^m E(y_i - \mathbf{w}_i^\top \mathbf{x}_i)^2 + \frac{1}{2} \sum_{j=1}^n \log \alpha_j + \frac{n}{2} \log \det \mathbf{K} -$$

$$\frac{1}{2} \sum_{j=1}^n \alpha_j E \mathbf{v}_j^\top \mathbf{K} \mathbf{v}_j \rightarrow \max_{\beta, \mathbf{A}, \lambda}.$$

$$\beta^{-1} = \frac{1}{m} \sum_{i=1}^m E(y_i - \mathbf{w}_i^\top \mathbf{x}_i)^2; \quad \alpha_j = \frac{1}{E \mathbf{v}_j^\top \mathbf{K} \mathbf{v}_j} = \frac{1}{\text{tr}(\mathbf{K} E \mathbf{v}_j \mathbf{v}_j^\top)}.$$

Hint: $\alpha_j^{\text{new}} = \frac{1 - \alpha_j^{\text{old}} \text{tr}(\mathbf{K} E \hat{\mathbf{v}}_j \hat{\mathbf{v}}_j^\top)}{\text{tr}(\mathbf{K} (E \mathbf{v}_j)(E \mathbf{v}_j)^\top)}$.

$$\mathbf{B} = \sum_{j=1}^n \alpha_j E \mathbf{v}_j \mathbf{v}_j^\top, \text{ тогда } f(\lambda) = \frac{n}{2} \log \det \mathbf{K} - \frac{1}{2} \text{tr}(\mathbf{K} \mathbf{B}) \rightarrow \max_{\lambda}.$$

$$2 \frac{df}{d\lambda} = n \text{tr} \left(\frac{d\mathbf{K}}{d\lambda} \mathbf{K}^{-1} \right) - \text{tr} \left(\frac{d\mathbf{K}}{d\lambda} \mathbf{B} \right) = 0.$$

Вопрос: как получить оптимальное λ ?

- 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.