Байесовский выбор моделей: гауссовские процессы для учета эволюции модели

Александр Адуенко

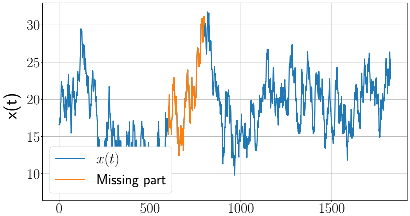
30е октября 2019

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

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

Вопрос 1: как учесть наличие дополнительной информации?



Рассмотрим случайный процесс $x(t)$, $t \in T$.
$m_x(t) = \mathsf{E}x(t)$, $K_x(t, s) = \mathsf{E}x(t)x(s)$, $R_x(t, s) = \mathsf{E}\mathring{x}(t)\mathring{x}(s)$ – 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 востановить пропущенные данные?

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

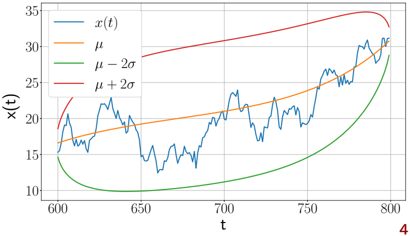
$\mathbf{x}_2|\mathbf{x}_1 \sim N(\mathbf{x}_2|\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}(\mathbf{x}_1 - \boldsymbol{\mu}_1),\ \boldsymbol{\Sigma}_{22}^{-1})$.

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);
- $K(x_1, x_2) = \sigma^2 \exp\left(-2 \frac{\sin^2(\pi r)}{l^2}\right)$ (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?

Байесовская линейная регрессия

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

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

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

Байесовская линейная регрессия с эволюцией

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

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

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

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

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 $v_j(t)$.

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

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

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

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

Получение апостериорного распределения

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

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

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

$$\frac{m}{2}\log\beta - \frac{\beta}{2}\sum_{i=1}^{m}(y_i - \mathbf{w}_i^{\mathsf{T}}\mathbf{x}_i)^2 + \sum_{j=1}^{n}\left[\frac{1}{2}\log\alpha_j + \frac{1}{2}\log\det\mathbf{K} - \frac{\alpha_j}{2}\mathbf{v}_j^{\mathsf{T}}\mathbf{K}\mathbf{v}_j\right].$$

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

$$-\frac{1}{2}\left[\sum_{j=1}^{n}\alpha_j\mathbf{v}_j^{\mathsf{T}}\mathbf{K}\mathbf{v}_j + \beta\sum_{i=1}^{m}\mathbf{w}_i^{\mathsf{T}}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\mathbf{w}_i - 2\beta\sum_{i=1}^{m}y_i\mathbf{w}_i^{\mathsf{T}}\mathbf{x}_i\right].$$

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

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

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

$$\log p(\mathbf{w}_1, \ldots, \mathbf{w}_m, \mathbf{w}_{m+1} | \mathbf{y}, \mathbf{X}, \boldsymbol{t}, \beta, \mathbf{A}, \lambda) \propto -\frac{1}{2}\left[\mathbf{u}^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\mathbf{u} - 2\mathbf{u}^{\mathsf{T}}\mathbf{m}\right] =$$

$$-\frac{1}{2}\left[\sum_{j=1}^{n}\alpha_j \mathbf{v}_j^{\mathsf{T}}\mathbf{K}\mathbf{v}_j + \beta\sum_{i=1}^{m}\mathbf{w}_i^{\mathsf{T}}\mathbf{x}_i\mathbf{x}_i^{\mathsf{T}}\mathbf{w}_i - 2\beta\sum_{i=1}^{m} y_i\mathbf{w}_i^{\mathsf{T}}\mathbf{x}_i\right].$$

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

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

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 $p(\mathbf{y}|\mathbf{X}, \boldsymbol{t}, \beta, \mathbf{A}, \lambda) \to \max\limits_{\beta, \mathbf{A}, \lambda}$.

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

E-шаг. $q(\mathbf{Z}) = p(\mathbf{w}_1, \ldots, \mathbf{w}_m, \mathbf{w}_{m+1}|\mathbf{y}, \mathbf{X}, \boldsymbol{t}, \beta, \mathbf{A}, \lambda)$.
M-шаг. $\mathsf{E}_q \log p(\mathbf{y}, \mathbf{w}_1, \ldots, \mathbf{w}_m, \mathbf{w}_{m+1}|\mathbf{X}, \boldsymbol{t}, \beta, \mathbf{A}, \lambda) \to \max\limits_{\beta, \mathbf{A}, \lambda}$.

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

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

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

Hint: $\alpha_j^{\mathsf{new}} = \dfrac{1-\alpha_j^{\mathsf{old}}\mathrm{tr}(\mathbf{K}\mathsf{E}\mathring{\mathbf{v}}_j\mathring{\mathbf{v}}_j^{\mathsf{T}})}{\mathrm{tr}(\mathbf{K}(\mathsf{E}\mathbf{v}_j)(\mathsf{E}\mathbf{v}_j)^{\mathsf{T}})}$.

М-шаг. $\mathsf{E}_q \log p(\mathbf{y},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_m,\ \mathbf{w}_{m+1}|\mathbf{X},\ \boldsymbol{t},\ \beta,\ \mathbf{A},\ \lambda) \to \max\limits_{\beta,\, \mathbf{A},\, \lambda}.$

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

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

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

Hint: $\alpha_j^{\mathsf{new}} = \frac{1-\alpha_j^{\mathsf{old}}\operatorname{tr}(\mathbf{K}\mathsf{E}\mathring{\mathbf{v}}_j\mathring{\mathbf{v}}_j^{\mathsf{T}})}{\operatorname{tr}(\mathbf{K}(\mathsf{E}\mathbf{v}_j)(\mathsf{E}\mathbf{v}_j)^{\mathsf{T}})}.$

$\mathbf{B} = \sum\limits_{j=1}^{n}\alpha_j\mathsf{E}\mathbf{v}_j\mathbf{v}_j^{\mathsf{T}}$, тогда $f(\lambda) = \frac{n}{2}\log\det\mathbf{K} - \frac{1}{2}\operatorname{tr}(\mathbf{K}\mathbf{B}) \to \max\limits_{\lambda}.$

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

Вопрос: как получить оптимальное λ?

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.