Байесовский выбор моделей: Вариационный EM-алгоритм (завершение). Гауссовские Процессы для учета эволюции модели.

Александр Адуенко

18е ноября 2025

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.

EM-алгоритм: воспоминание

Пусть $\mathbf{D} = (\mathbf{X}, \mathbf{y})$ – наблюдаемые переменные, $\mathbf{Z}$ – скрытые переменные.
$p(\mathbf{D}, \mathbf{Z}|\mathbf{\Theta}) = p(\mathbf{D}|\mathbf{Z}, \mathbf{\Theta})p(\mathbf{Z}|\mathbf{\Theta})$.

Вопрос 1: как решить задачу $p(\mathbf{D}|\mathbf{\Theta}) = \int p(\mathbf{D}, \mathbf{Z}|\mathbf{\Theta})d\mathbf{Z} \to \max\limits_{\mathbf{\Theta}}$?

EM-алгоритм

Введем $F(q, \mathbf{\Theta}) = -\int q(\mathbf{Z}) \log q(\mathbf{Z})d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{D}, \mathbf{Z}|\mathbf{\Theta})d\mathbf{Z} =$

$-\int q(\mathbf{Z}) \log q(\mathbf{Z})d\mathbf{Z} + \int q(\mathbf{Z}) \log p(\mathbf{Z}|\mathbf{D}, \mathbf{\Theta})d\mathbf{Z} + \int \log p(\mathbf{D}|\mathbf{\Theta})q(\mathbf{Z})d\mathbf{Z} =$

$\log p(\mathbf{D}|\mathbf{\Theta}) - \int q(\mathbf{Z}) \log \frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{D},\mathbf{\Theta})}d\mathbf{Z} = \log p(\mathbf{D}|\mathbf{\Theta}) - D_{\mathrm{KL}}(q\|p(\mathbf{Z}|\mathbf{D}, \mathbf{\Theta}))$.

Идея 1: $p(\mathbf{D}|\mathbf{\Theta}) \to \max\limits_{\mathbf{\Theta}}$ заменим на $F(q, \mathbf{\Theta}) \to \max\limits_{q,\, \mathbf{\Theta}}$.

Идея 2: Пошагово оптимизируем по $\mathbf{\Theta}$ и q, то есть

1. E-шаг: $q^s = F(q, \mathbf{\Theta}^{s-1}) \to \max\limits_{q \in Q}$;
2. M-шаг: $\mathbf{\Theta}^s = F(q^s, \mathbf{\Theta}) \to \max\limits_{\mathbf{\Theta}}$.

Вопрос: Зачем $q \in Q$? Как E-шаг был выполнен при максимизации обоснованности для модели линейной регрессии?

Вариационный EM-алгоритм

$$F(q,\, \boldsymbol{\Theta}) = \int q(\mathbf{Z}) \log \frac{p(\mathbf{D},\, \mathbf{Z}|\boldsymbol{\Theta})}{q(\mathbf{Z})} d\mathbf{Z} = \log p(\mathbf{D}|\boldsymbol{\Theta}) - D_{\mathrm{KL}}(q \| p(\mathbf{Z}|\mathbf{D},\, \boldsymbol{\Theta})).$$

E-шаг. $\displaystyle\int q(\mathbf{Z}_k) \log \frac{q(\mathbf{Z}_k)}{\frac{1}{C} e^{\mathsf{E}_{q \setminus k} \log p(\mathbf{D},\, \mathbf{Z}|\boldsymbol{\Theta})}} d\mathbf{Z}_k \to \min_{q(\mathbf{Z}_k)}.$

Полный алгоритм

Пошагово оптимизируем по $\boldsymbol{\Theta}$ и $q(\mathbf{Z}_k)$, $k = 1, \ldots, K$, то есть

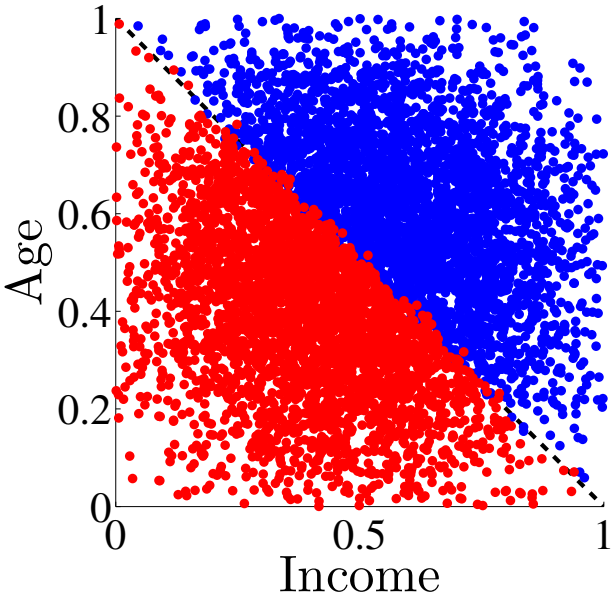
1. E-шаг: $\log q(\mathbf{Z}_k^s) \propto \mathsf{E}_{q \setminus k} \log p(\mathbf{D},\, \mathbf{Z}|\boldsymbol{\Theta}^{s-1})$;
2. M-шаг: $\mathsf{E}_{q^s} \log p(\mathbf{D},\, \mathbf{Z}|\boldsymbol{\Theta}) \to \max\limits_{\boldsymbol{\Theta}}$.

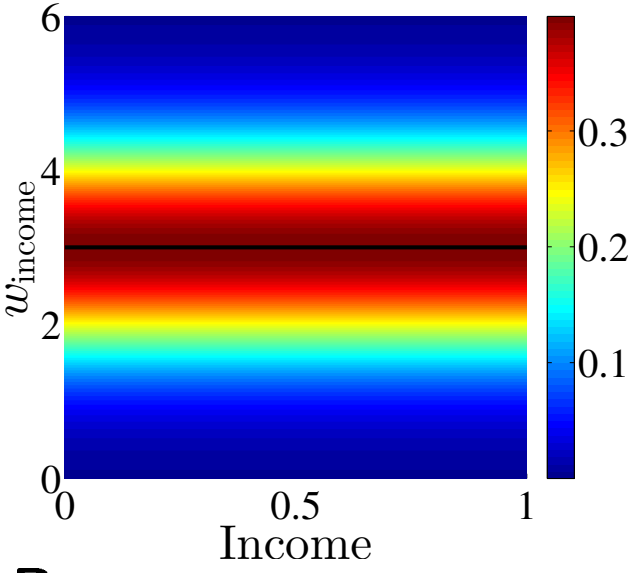
Вопрос 1: Зачем нужна факторизация? Чем полученные итеративные формулы лучше формул полного EM-алгоритма?

Вопрос 2: Как понять, что в конкретной задаче формулы E и M-шагов выписаны верно?

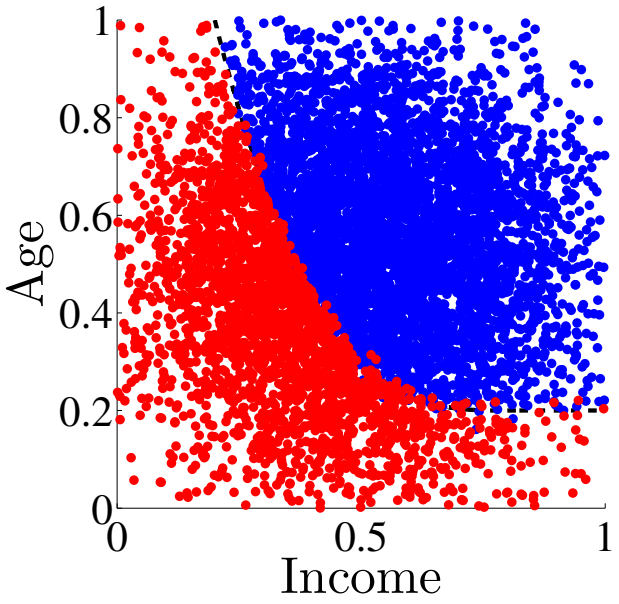
Нарушение свойства $p(\mathbf{w}|\mathbf{x}_i) = p(\mathbf{w})$

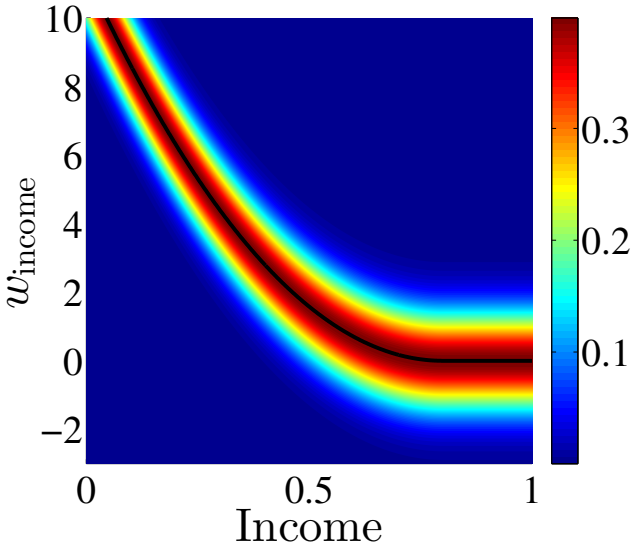
Предполагаемый результат





Реальные данные





Вопрос: как можно учесть указанную нелинейность в модели?

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

$$p(\mathbf{y},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \boldsymbol{\pi}|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K,\ \boldsymbol{\mu}) =$$

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

Введем матрицу скрытых переменных $\mathbf{Z} = \|z_{ik}\|$, где $z_{ik} = 1 \Longleftrightarrow k_i = k$.

$$p(\mathbf{y},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \boldsymbol{\pi},\ \mathbf{Z}|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K,\ \boldsymbol{\mu}) =$$

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

Получение MAP-оценки

Пусть $p(\boldsymbol{\pi}|\boldsymbol{\mu}) = \mathrm{Dir}(\boldsymbol{\mu}) = \dfrac{\Gamma(\sum_k \mu_k)}{\prod_l \Gamma(\mu_l)} \prod_k \pi_k^{\mu_k - 1}$.

$$p(\mathbf{y},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K,\, \boldsymbol{\pi},\, \mathbf{Z}|\mathbf{X},\, \mathbf{A}_1,\, \ldots,\, \mathbf{A}_K,\, \boldsymbol{\mu}) \propto$$

$$\prod_{k=1}^{K} \pi_k^{\mu_k - 1} \prod_{k=1}^{K} \sqrt{\det \mathbf{A}_k} \exp(-\tfrac{1}{2}\mathbf{w}_k^{\mathsf{T}} \mathbf{A}_k \mathbf{w}_k) \prod_{i=1}^{m} \prod_{l=1}^{K} \Big(\pi_l \sigma(y_i \mathbf{w}_l^{\mathsf{T}} \mathbf{x}_i)\Big)^{z_{il}}.$$

$$(\boldsymbol{\pi}^*, \mathbf{w}_1^*, \ldots, \mathbf{w}_K^*) = \arg \max_{\boldsymbol{\pi},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K} p(\mathbf{y}, \mathbf{w}_1, \ldots, \mathbf{w}_K, \boldsymbol{\pi}|\mathbf{X}, \mathbf{A}_1, \ldots, \mathbf{A}_K, \boldsymbol{\mu}).$$

E-шаг. $\log q(\mathbf{Z}) \propto \sum_{i=1}^{m} \sum_{l=1}^{K} z_{il} (\log \pi_l + \log \sigma(y_i \mathbf{w}_l^{\mathsf{T}} \mathbf{x}_i))$, откуда

$$\gamma_{ik} = p(z_{ik} = 1) \propto \pi_k \sigma(y_i \mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i).$$

M-шаг. $\mathsf{E}_q \log p(\mathbf{y},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K,\, \boldsymbol{\pi},\, \mathbf{Z}|\mathbf{X},\, \mathbf{A}_1,\, \ldots,\, \mathbf{A}_K,\, \boldsymbol{\mu}) \to \max_{\boldsymbol{\pi},\, \mathbf{w}_1,\, \ldots,\, \mathbf{w}_K}$.

$$\mathbf{w}_k^* = \arg\max_{\mathbf{w}_k} \left[-\tfrac{1}{2}\mathbf{w}_k^{\mathsf{T}} \mathbf{A}_k \mathbf{w}_k + \sum_{i=1}^{m} \gamma_{ik} \log \sigma(y_i \mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i) \right].$$

$$\boldsymbol{\pi}^* = \arg\max_{\boldsymbol{\pi}} \sum_{k=1}^{K} \log \pi_k \Big(\underbrace{\sum_{i=1}^{m} \gamma_{ik}}_{\gamma_k} + \mu_k - 1 \Big) \Longrightarrow \pi_k \propto \max(0,\, \gamma_k + \mu_k - 1).$$

Получение апостериорного распределения

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

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

Идея: найдем $q(\mathbf{Z}, \mathbf{w}_1, \ldots, \mathbf{w}_K, \boldsymbol{\pi}) = q(\mathbf{Z}) q(\mathbf{w}_1, \ldots, \mathbf{w}_K) q(\boldsymbol{\pi})$, наиболее близкое к $p(\mathbf{w}_1, \ldots, \mathbf{w}_k, \boldsymbol{\pi}, \mathbf{Z} | \mathbf{X}, \mathbf{y})$.

$$\log q(\mathbf{Z}) \propto \mathsf{E}_{q_{\setminus \mathbf{Z}}} \log p(\mathbf{y}, \mathbf{w}_1, \ldots, \mathbf{w}_K, \boldsymbol{\pi}, \mathbf{Z} | \mathbf{X}, \mathbf{A}_1, \ldots, \mathbf{A}_K, \boldsymbol{\mu}) \propto$$
$$\sum_{i=1}^{m} \sum_{k=1}^{K} z_{ik} \left(\mathsf{E} \log \pi_k + \mathsf{E} \log \sigma(y_i \mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i) \right)$$
$$\Longrightarrow p(z_{ik} = 1) \propto \exp \left(\mathsf{E} \log \pi_k + \mathsf{E} \log \sigma(y_i \mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i) \right).$$

$$\log q(\boldsymbol{\pi}) \propto \mathsf{E}_{q_{\setminus \boldsymbol{\pi}}} \log p(\mathbf{y}, \mathbf{w}_1, \ldots, \mathbf{w}_K, \boldsymbol{\pi}, \mathbf{Z} | \mathbf{X}, \mathbf{A}_1, \ldots, \mathbf{A}_K, \boldsymbol{\mu}) \propto$$
$$\sum_{k=1}^{K} \log \pi_k \left(\mu_k - 1 + \sum_{i=1}^{m} \mathsf{E} z_{ik} \right) \Longrightarrow \boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\mu} + \boldsymbol{\gamma}).$$

Получение апостериорного распределения (продолжение)

$$\log q(\mathbf{w}_1, \ldots, \mathbf{w}_K) \propto \mathsf{E}_{q_{\setminus \mathbf{w}}} \log p(\mathbf{y}, \mathbf{w}_1, \ldots, \mathbf{w}_K, \boldsymbol{\pi}, \mathbf{Z} | \mathbf{X}, \mathbf{A}_1, \ldots, \mathbf{A}_K, \boldsymbol{\mu}) \propto$$

$$\sum_{k=1}^{K} \left(-\tfrac{1}{2} \mathbf{w}_k^{\mathsf{T}} \mathbf{A}_k \mathbf{w}_k + \sum_{i=1}^{m} \mathsf{E} z_{ik} \log \sigma(y_i \mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i) \right) = \sum_{k=1}^{K} f_k(\mathbf{w}_k).$$

Вопрос 1: Какую структуру имеет $q(\mathbf{w}_1, \ldots, \mathbf{w}_K)$?

Вопрос 2: Какой вид имеет распределение $q(\mathbf{w}_k)$?

Варианты аппроксимации $q(\mathbf{w}_k)$:

- Аппроксимация Лапласа: $q(\mathbf{w}_k) \approx \mathcal{N}(\mathbf{w}_k^*, \boldsymbol{\Sigma}_k^{-1})$;
- Ищем $q(\mathbf{w}_k) = \mathcal{N}(\mathbf{m}_k, \boldsymbol{\Sigma}_k^{-1})$ такое, что $D_{KL}(q \| C_k f_k(\mathbf{w}_k)) \to \min\limits_{q}$
  - Численно (если число признаков n невелико);
  - С помощью VLB для сигмоиды и соответствующей верхней оценки для $D_{KL}(q \| C_k f_k(\mathbf{w}_k))$.

Вопрос 3: Как определить $\mathbf{A}_1, \ldots, \mathbf{A}_K, \boldsymbol{\mu}$?

Определение гиперпараметров смеси моделей

Максимизация обоснованности

$p(\mathbf{y}|\mathbf{X}, \mathbf{A}_1, \ldots, \mathbf{A}_K, \boldsymbol{\mu}) \to \max_{\mathbf{A}_1, \ldots, \mathbf{A}_K, \{\boldsymbol{\mu}\}}$

Идея: воспользуемся вариационным EM-алгоритмом.

E-шаг: найдем $q(\mathbf{Z}, \mathbf{w}_1, \ldots, \mathbf{w}_K, \boldsymbol{\pi}) = q(\mathbf{Z})q(\mathbf{w}_1, \ldots, \mathbf{w}_K)q(\boldsymbol{\pi})$, наиболее близкое к $p(\mathbf{w}_1, \ldots, \mathbf{w}_k, \boldsymbol{\pi}, \mathbf{Z}|\mathbf{X}, \mathbf{y})$.

M-шаг: $\mathsf{E}_q \log p(\mathbf{y}, \mathbf{w}_1, \ldots, \mathbf{w}_K, \boldsymbol{\pi}, \mathbf{Z}|\mathbf{X}, \mathbf{A}_1, \ldots, \mathbf{A}_K, \boldsymbol{\mu}) \to \max_{\mathbf{A}_1, \ldots, \mathbf{A}_K}$.

$$\sum_{k=1}^{K}\left[\sum_{j=1}^{n} \log \alpha_k^j - \alpha_k^j \mathsf{E} w_{kj}^2\right] \to \max_{\alpha_{11}, \ldots, \alpha_{Kn}}, \text{ где } \mathbf{A}_k = \operatorname{diag}(\alpha_{k1}, \ldots, \alpha_{kn}).$$

$\alpha_{kj} = \dfrac{1}{\mathsf{E} w_{kj}^2}$. **Hint:** $\alpha_{kj} = \dfrac{1 - \alpha_{kj}^{\text{old}} \mathsf{D} w_{kj}}{\mathsf{E}^2 w_{kj}}$.

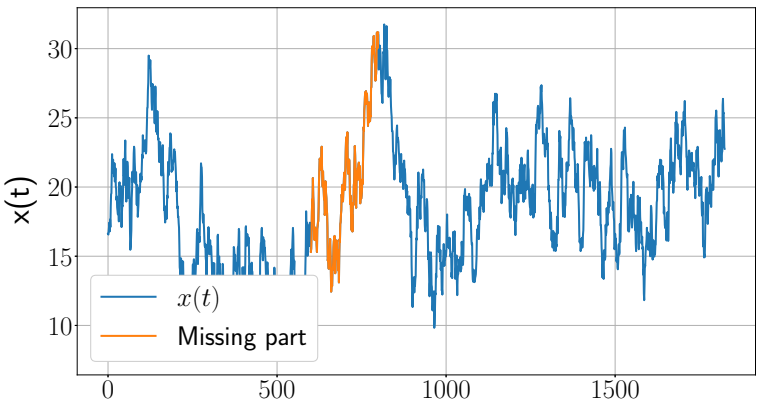
Вопрос: Как получить $\mathsf{E} w_{kj}$ и $\mathsf{D} w_{kj}$?

Замечание: можно сразу сделать использовать VLB для сигмоидной функции, чтобы получить нижнюю границу обоснованности
$p(\mathbf{y}|\mathbf{X}, \mathbf{A}_1, \ldots, \mathbf{A}_K, \boldsymbol{\mu}) \geq \tilde{p}(\mathbf{y}|\mathbf{X}, \mathbf{A}_1, \ldots, \mathbf{A}_K, \boldsymbol{\mu})$
и тогда на E-шаге автоматически $q(\mathbf{w}_k)$ будет нормальным.

Учет эволюции модели во времени

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 $x(t),\ t \in T$.
$m_x(t) = \mathrm{E}x(t),\ K_x(t,\ s) = \mathrm{E}x(t)x(s),\ R_x(t,\ s) = \mathrm{E}\mathring{x}(t)\mathring{x}(s)$ – 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?

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

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

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

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

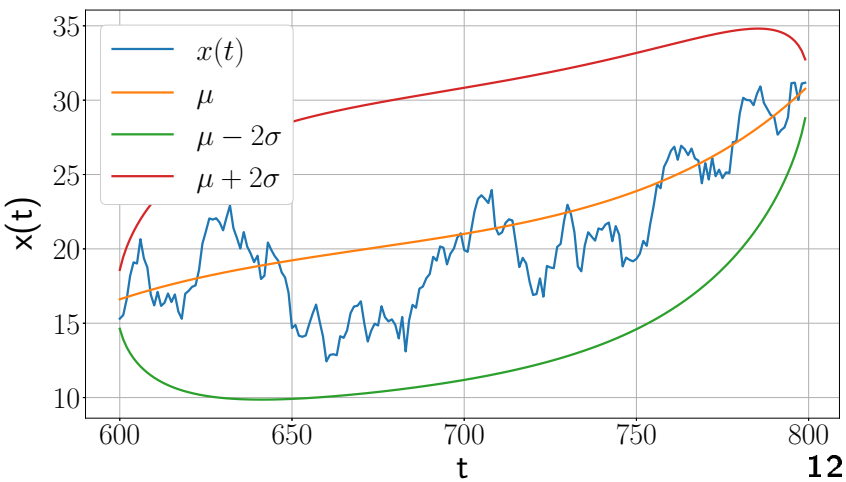
$\mathbf{x}_2|\mathbf{x}_1 \sim \mathcal{N}(\mathbf{x}_2|\boldsymbol{\mu}_2 - \mathbf{\Sigma}_{22}^{-1}\mathbf{\Sigma}_{12}(\mathbf{x}_1 - \boldsymbol{\mu}_1), \mathbf{\Sigma}_{22}^{-1})$.

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?

Линейная регрессия с эволюцией во времени

Байесовская линейная регрессия

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

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

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

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 $t_1 < t_2 < \ldots < t_m$.

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

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

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

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

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

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

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

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

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

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

Получение апостериорного распределения

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

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

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

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

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

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

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

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

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

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

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

Отбор признаков и подбор ковариационной функции

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

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

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

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

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

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

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

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.