Байесовский выбор моделей: вариационный EM-алгоритм; гауссовские процессы

Александр Адуенко

14е ноября 2018

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 $\mathbf{D} = (\mathbf{X}, \mathbf{y})$ – 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 $p(\mathbf{D}|\mathbf{\Theta}) = \int p(\mathbf{D}, \mathbf{Z}|\mathbf{\Theta})d\mathbf{Z} \to \max\limits_{\mathbf{\Theta}}$?

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

Вопрос: Зачем $q \in Q$? Как Е-шаг был выполнен на прошлой лекции 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. $C \displaystyle\int q(\mathbf{Z}_k) \log \frac{C q(\mathbf{Z}_k)}{e^{\mathsf{E}_{q\setminus k} \log p(\mathbf{D},\, \mathbf{Z}|\boldsymbol{\Theta})}} d\mathbf{Z}_k \to \min_{q(\mathbf{Z}_k)}.$

Полный алгоритм

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

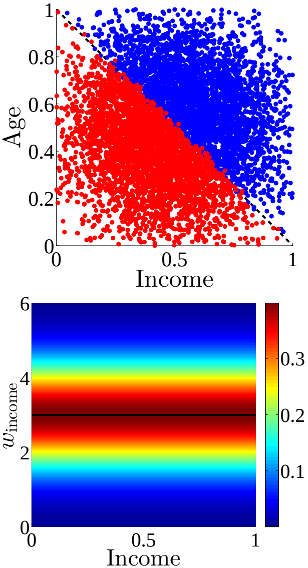
1. E-шаг: $\log q(\mathbf{Z}_k^s) \propto \mathsf{E}_{q\setminus k} \log p(\mathbf{D},\ \mathbf{Z}|\boldsymbol{\Theta}^{s-1})$;
2. M-шаг: $\mathsf{E}_{q^s} \log p(\mathbf{D},\ \mathbf{Z}|\boldsymbol{\Theta}) \to \max\limits_{\boldsymbol{\Theta}}$.

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

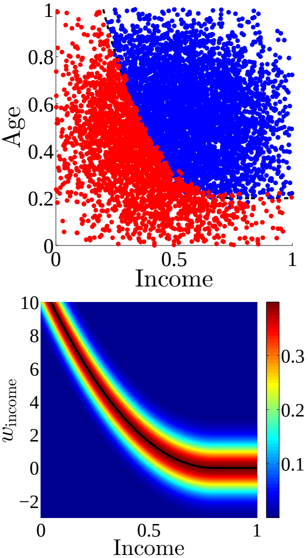
Вопрос 2: как понять, что в конкретной задаче формулы E и M-шагов выписаны верно?

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

$$p(\mathbf{y},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \boldsymbol{\pi}|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K,\ \boldsymbol{\mu}) =$$
$$p(\boldsymbol{\pi}|\boldsymbol{\mu}) \prod_{k=1}^{K} N(\mathbf{w}_k|\mathbf{0},\ \mathbf{A}_k^{-1}) \prod_{i=1}^{m} \left(\sum_{l=1}^{K} \pi_l \sigma(y_i \mathbf{w}_l^{\mathsf{T}} \mathbf{x}_i) \right).$$

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

$$p(\mathbf{y},\ \mathbf{w}_1,\ \ldots,\ \mathbf{w}_K,\ \boldsymbol{\pi},\ \mathbf{Z}|\mathbf{X},\ \mathbf{A}_1,\ \ldots,\ \mathbf{A}_K,\ \boldsymbol{\mu}) =$$
$$p(\boldsymbol{\pi}|\boldsymbol{\mu}) \prod_{k=1}^{K} N(\mathbf{w}_k|\mathbf{0},\ \mathbf{A}_k^{-1}) \prod_{i=1}^{m} \prod_{l=1}^{K} \Big(\pi_l \sigma(y_i \mathbf{w}_l^{\mathsf{T}} \mathbf{x}_i) \Big)^{z_{il}}.$$

Получение апостериорного распределения

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

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

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 $p(\mathbf{w}_1, \ldots, \mathbf{w}_k, \boldsymbol{\pi}, \mathbf{Z}|\mathbf{X}, \mathbf{y})$.

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

$$\Longrightarrow p(z_{ik} = 1) \propto \exp \left(\mathsf{E} \log \pi_k + \mathsf{E} \log \sigma(y_i \mathbf{w}_k^{\mathsf{T}} \mathbf{x}_i) \right).$$

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

Получение апостериорного распределения (продолжение)

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

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

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: $q(\mathbf{w}_k) \approx N(\mathbf{w}_k|\mathbf{w}_k^*, \boldsymbol{\Sigma}_k^{-1})$;
- Ищем $q(\mathbf{w}_k) = N(\mathbf{w}_k|\mathbf{m}_k, \boldsymbol{\Sigma}_k^{-1})$ 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 $\mathbf{A}_1, \ldots, \mathbf{A}_K, \boldsymbol{\mu}$?

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 ЕМ-алгоритмом.

Е-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 $p(\mathbf{w}_1, \ldots, \mathbf{w}_k, \boldsymbol{\pi}, \mathbf{Z}|\mathbf{X}, \mathbf{y})$.

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

Замечание: можно сразу сделать использовать VLB для сигмоидной функции, чтобы получить нижнюю границу обоснованности

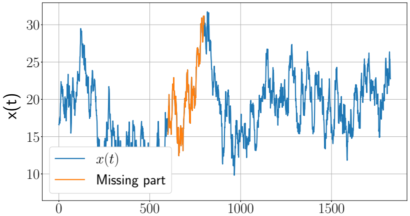
$$p(\mathbf{y}|\mathbf{X}, \mathbf{A}_1, \ldots, \mathbf{A}_K, \boldsymbol{\mu}) \geq \tilde{p}(\mathbf{y}|\mathbf{X}, \mathbf{A}_1, \ldots, \mathbf{A}_K, \boldsymbol{\mu})$$

и тогда на Е-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 $x(t),\ t \in T$.

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

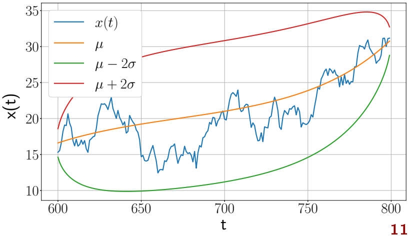
$\mathbf{x}_2|\mathbf{x}_1 \sim N(\mathbf{x}_2|\boldsymbol{\mu}_2 - \boldsymbol{\Sigma}_{22}^{-1}\boldsymbol{\Sigma}_{12}(\mathbf{x}_1 - \boldsymbol{\mu}_1),\ \boldsymbol{\Sigma}_{22}^{-1})$.

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 $R_x(\tau)$.



Литература

1. Bishop, Christopher M. "Pattern recognition and machine learning". Springer, New York (2006). Pp. 113-120, 161-171, 498-505.
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.
6. Chen, Ming-Hui, and Joseph G. Ibrahim. "Conjugate priors for generalized linear models."Statistica Sinica (2003): 461-476.
7. Fahrmeir, Ludwig, and Heinz Kaufmann. "Consistency and asymptotic normality of the maximum likelihood estimator in generalized linear models."The Annals of Statistics (1985): 342-368.
8. Baghishani, Hossein, and Mohsen Mohammadzadeh. "Asymptotic normality of posterior distributions for generalized linear mixed models."Journal of Multivariate Analysis 111 (2012): 66-77.