

Байесовский выбор моделей: EM и вариационный EM-алгоритм

Александр Адуенко

3е ноября 2021

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 ; связь априорного распределения с отбором признаков.
- Неопределенность в прогнозе и апостериорное распределение; учёт нелинейных зависимостей, пропусков и выбросов.
- EM-алгоритм. Использование EM-алгоритма для отбора признаков в байесовской линейной регрессии.

EM-алгоритм: воспоминание

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

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

EM-алгоритм

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

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

Идея 2: Пошагово оптимизируем по Θ и q , то есть

1 E-шаг: $q^s = F(q, \Theta^{s-1}) \rightarrow \max_{q \in Q}$;

2 M-шаг: $\Theta^s = F(q^s, \Theta) \rightarrow \max_{\Theta}$.

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

Вариационный EM-алгоритм. E-шаг

$$F(q, \Theta^{s-1}) \rightarrow \max_{q \in Q} \iff D_{\text{KL}}(q \| p(\mathbf{Z} | \mathbf{D}, \Theta)) \rightarrow \min_{q \in Q}$$

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

Пусть $Q = \left\{ q : q(\mathbf{Z}) = \prod_{k=1}^K q(\mathbf{Z}_k) \right\}$, тогда

$$D_{\text{KL}}(q \| p(\mathbf{Z} | \mathbf{D}, \Theta)) \propto \int \prod_{k=1}^K q(\mathbf{Z}_k) \log \frac{\prod_{j=1}^K q(\mathbf{Z}_j)}{p(\mathbf{D}, \mathbf{Z}_1, \dots, \mathbf{Z}_K | \Theta)} d\mathbf{Z}_1 \dots d\mathbf{Z}_K =$$

$$\int q(\mathbf{Z}_k) \log q(\mathbf{Z}_k) \underbrace{\left[\prod_{j \neq k} \int q(\mathbf{Z}_j) d\mathbf{Z}_j \right]}_{=1} d\mathbf{Z}_k - \sum_{j \neq k} C_j \underbrace{\int q(\mathbf{Z}_k) d\mathbf{Z}_k}_{=1}$$

$$\int q(\mathbf{Z}_k) \underbrace{\left[\int \prod_{j \neq k} q(\mathbf{Z}_j) \log p(\mathbf{D}, \mathbf{Z}_1, \dots, \mathbf{Z}_K | \Theta) d\mathbf{Z}_{j \neq k} \right]}_{=1} d\mathbf{Z}_k \propto$$

$$\underbrace{\int q(\mathbf{Z}_k) \log p(\mathbf{D}, \mathbf{Z} | \Theta) d\mathbf{Z}_k}_{E_{q \setminus k} \log p(\mathbf{D}, \mathbf{Z} | \Theta)}$$

$$\int q(\mathbf{Z}_k) \log \frac{q(\mathbf{Z}_k)}{\frac{1}{C} e^{E_{q \setminus k} \log p(\mathbf{D}, \mathbf{Z} | \Theta)}} d\mathbf{Z}_k \rightarrow \min_{q(\mathbf{Z}_k)}$$

Вариационный EM-алгоритм

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

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

Полный алгоритм

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

1 E-шаг: $\log q(\mathbf{Z}_k^s) \propto \mathbb{E}_{q \setminus k} \log p(\mathbf{D}, \mathbf{Z}|\Theta^{s-1})$;

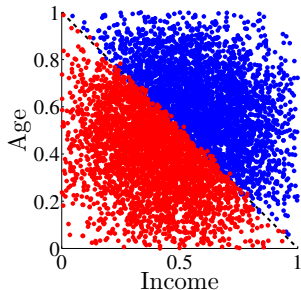
2 M-шаг: $\mathbb{E}_{q^s} \log p(\mathbf{D}, \mathbf{Z}|\Theta) \rightarrow \max_{\Theta}$.

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

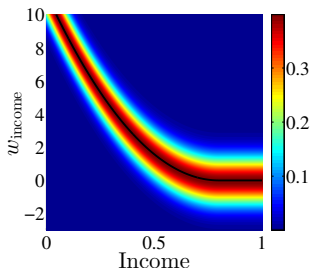
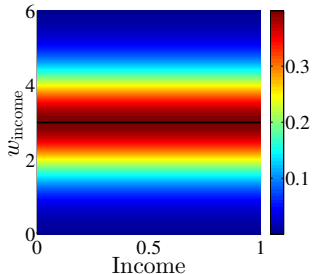
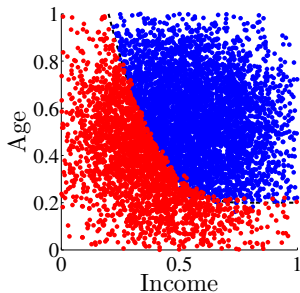
Вопрос 2: как понять, что в конкретной задаче формулы E и M-шагов выписаны верно?

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 π получены из априорного распределения $p(\pi|\mu)$;
- Векторы параметров моделей \mathbf{w}_k получены из нормального распределения $p(\mathbf{w}_k|\mathbf{A}_k) = \mathcal{N}(\mathbf{w}_k|\mathbf{0}, \mathbf{A}_k^{-1})$, $k = 1, \dots, K$;
- 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} : $y_i \sim \text{Be}(\sigma(\mathbf{w}_{k_i}^\top \mathbf{x}_i))$.

Совместное правдоподобие модели

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

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

$$p(\mathbf{y}, \mathbf{w}_1, \dots, \mathbf{w}_K, \pi, \mathbf{Z} | \mathbf{X}, \mathbf{A}_1, \dots, \mathbf{A}_K, \mu) = p(\pi | \mu) \prod_{k=1}^K \mathcal{N}(\mathbf{w}_k | \mathbf{0}, \mathbf{A}_k^{-1}) \prod_{i=1}^m \prod_{l=1}^K \left(\pi_l \sigma(y_i \mathbf{w}_l^\top \mathbf{x}_i) \right)^{\mathbf{z}_{il}}.$$

Получение MAP-оценки

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

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

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

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

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

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

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

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

Получение апостериорного распределения

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

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

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

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

$$\sum_{i=1}^m \sum_{k=1}^K z_{ik} \left(\mathbb{E} \log \pi_k + \mathbb{E} \log \sigma(y_i \mathbf{w}_k^\top \mathbf{x}_i) \right)$$

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

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

$$\sum_{k=1}^K \log \pi_k \left(\mu_k - 1 + \sum_{i=1}^m \mathbb{E} z_{ik} \right) \implies \boldsymbol{\pi} \sim \text{Dir}(\boldsymbol{\pi} | \boldsymbol{\mu} + \boldsymbol{\gamma}).$$

Получение апостериорного распределения (продолжение)

$$\log q(\mathbf{w}_1, \dots, \mathbf{w}_K) \propto E_{q_{\setminus \mathbf{w}}} \log p(\mathbf{y}, \mathbf{w}_1, \dots, \mathbf{w}_K, \boldsymbol{\pi}, \mathbf{Z} | \mathbf{X}, \mathbf{A}_1, \dots, \mathbf{A}_K, \boldsymbol{\mu}) \propto \sum_{k=1}^K \left(-\frac{1}{2} \mathbf{w}_k^\top \mathbf{A}_k \mathbf{w}_k + \sum_{i=1}^m E z_{ik} \log \sigma(y_i \mathbf{w}_k^\top \mathbf{x}_i) \right) = \sum_{k=1}^K f_k(\mathbf{w}_k).$$

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

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

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

- Аппроксимация Лапласа: $q(\mathbf{w}_k) \approx \mathcal{N}(\mathbf{w}_k | \mathbf{w}_k^*, \boldsymbol{\Sigma}_k^{-1})$;
- Ищем $q(\mathbf{w}_k) = \mathcal{N}(\mathbf{w}_k | \mathbf{m}_k, \boldsymbol{\Sigma}_k^{-1})$ такое, что $D_{KL}(q \| C_k f_k(\mathbf{w}_k)) \rightarrow \min_q$
 - 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 $\mathbf{A}_1, \dots, \mathbf{A}_K, \boldsymbol{\mu}$?

Определение гиперпараметров смеси моделей

Максимизация обоснованности

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

Идея: воспользуемся вариационным EM-алгоритмом.

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

M-шаг: $E_q \log p(\mathbf{y}, \mathbf{w}_1, \dots, \mathbf{w}_K, \boldsymbol{\pi}, \mathbf{Z}|\mathbf{X}, \mathbf{A}_1, \dots, \mathbf{A}_K, \boldsymbol{\mu}) \rightarrow \max_{\mathbf{A}_1, \dots, \mathbf{A}_K}$.

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

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

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

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

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

и тогда на E-шаге автоматически $q(\mathbf{w}_k)$ будет нормальным.

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.