Байесовский выбор моделей II: введение

Александр Адуенко

20е февраля 2022

Байесовский выбор моделей II: план курса

- Воспоминания из части 1:
    - Априорное распределение и неинформативные распределения (Jeffreys prior);
    - EM-алгоритм и вариационный EM-алгоритм;
    - Методы сэмплирования (MCMC);
- Понятие графической модели. Байесовские и марковские сети.
- Ациклические графические модели и алгоритм Belief Propagation.
- Скрытые марковские модели (HMM) и их расширения.
- Линейные динамические системы. Фильтр Калмана и его расширения.
- Циклические графические модели и вывод в них.
- Сегментация изображений: Алгоритмы на основании разрезов графов и алгоритм α – расширение.
- Приближенный вывод в графических моделях, алгоритм Expectation Propagation.
- Понятие разладки и методы ее нахождения.

Система оценивания

- 12 лекций + 2-3 небольших теста на них (суммарно до 100 баллов);
- 6 заданий:
  - 2 теоретических по 100 баллов;
  - 2 практических по 100 баллов;
  - 2 групповых соревнования (до 200 баллов каждое);
- Экзамен:
  - Письменная часть (100 баллов),
  - Устная часть (150 баллов).

Замечания:

- На оценку k требуется набрать $100k$ баллов;
- Экзамен можно пропустить только, если набрано не менее 550 баллов до экзамена;
- Задания содержат задачи более, чем на 100 баллов, поэтому можно выбрать, что выполнять;
- В каждом задании баллы лучшей работы удваиваются, если она оценена более, чем в 100 баллов (не более 250 баллов);
- За каждую неделю опоздания балл за задание снижается в 2 раза. Задание не принимается после его разбора или объявления об этом.

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

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

Пример 1. $\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon}$, $\mathbf{w} \sim \mathcal{N}(\mathbf{w}|\mathbf{0},\ \mathbf{A}^{-1})$, $\boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0},\ \beta^{-1}\mathbf{I})$

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

$\log p(\mathbf{y}|\mathbf{X},\ \underbrace{\mathbf{A},\ \beta^{-1}}_{\boldsymbol{\Theta}}) \propto -\frac{1}{2}\log\det(\beta^{-1}\mathbf{I}+\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^{\mathsf{T}}) - \frac{1}{2}\mathbf{y}^{\mathsf{T}}(\beta^{-1}\mathbf{I}+\mathbf{X}\mathbf{A}^{-1}\mathbf{X}^{\mathsf{T}})^{-1}\mathbf{y}$.

EM-алгоритм

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

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

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

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

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

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

EM-алгоритм для максимизации обоснованности

$\mathbf{y} = \mathbf{X}\mathbf{w} + \boldsymbol{\varepsilon},\ \mathbf{w} \sim \mathcal{N}(\mathbf{w}|\mathbf{0},\ \mathbf{A}^{-1}),\ \boldsymbol{\varepsilon} \sim \mathcal{N}(\mathbf{0},\ \beta^{-1}\mathbf{I})$

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

$\log p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A},\ \beta) \propto \frac{m}{2}\log\beta - \frac{\beta}{2}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \frac{1}{2}\log\det\mathbf{A} - \frac{1}{2}\mathbf{w}^{\mathsf{T}}\mathbf{A}\mathbf{w}.$

$F(q,\ \mathbf{A},\ \beta) = -\int q(\mathbf{w})\log q(\mathbf{w})d\mathbf{w} + \int q(\mathbf{w})\log p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A},\ \beta)d\mathbf{w} =$

$\log p(\mathbf{y}|\mathbf{X},\ \mathbf{A},\ \beta) - D_{\mathrm{KL}}(q(\mathbf{w})\|p(\mathbf{w}|\mathbf{y},\ \mathbf{X},\ \mathbf{A},\ \beta)) \to \max\limits_{q,\ \mathbf{A},\ \beta}.$

Е-шаг (считаем $\mathbf{A},\ \beta$ фиксированными)

$F(q,\ \mathbf{A},\ \beta) \to \max\limits_{q} \Longleftrightarrow q(\mathbf{w}) = p(\mathbf{w}|\mathbf{y},\ \mathbf{X},\ \mathbf{A},\ \beta) = \mathrm{N}(\mathbf{w}|\mathbf{w}_0,\ \boldsymbol{\Sigma}_0^{-1})$, где

$\boldsymbol{\Sigma}_0 = \mathbf{A} + \beta\mathbf{X}^{\mathsf{T}}\mathbf{X},\ \mathbf{w}_0 = \beta\boldsymbol{\Sigma}_0^{-1}\mathbf{X}^{\mathsf{T}}\mathbf{y}.$

М-шаг (считаем $q(\mathbf{w})$ фиксированным)

$\mathsf{E}_{q(\mathbf{w})}\log p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A},\ \beta) = \int q(\mathbf{w})\log p(\mathbf{y},\ \mathbf{w}|\mathbf{X},\ \mathbf{A},\ \beta)d\mathbf{w} \to \max\limits_{\mathbf{A},\ \beta}.$

$\tilde{F}(\mathbf{A},\ \beta) = \frac{m}{2}\log\beta - \frac{\beta}{2}\mathsf{E}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 + \frac{1}{2}\sum\limits_{j=1}^{n}\log\alpha_j - \frac{1}{2}\sum\limits_{j=1}^{n}\alpha_j\mathsf{E}w_j^2 \to \max\limits_{\mathbf{A},\ \beta}.$

$\frac{\partial F}{\partial\alpha_j} = \frac{1}{2\alpha_j} - \frac{1}{2}\mathsf{E}w_j^2 = 0 \Longleftrightarrow \alpha_j = \frac{1}{\mathsf{E}w_j^2}.$

Hint: $1 = \alpha_j(\mathsf{E}^2 w_j + \mathsf{D}w_j) \Longrightarrow \alpha_j^{\text{new}} = \frac{1-\alpha_j^{\text{old}}\mathsf{D}w_j}{\mathsf{E}^2 w_j}.$

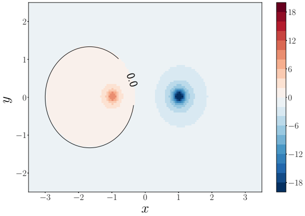
$\frac{\partial F}{\partial\beta} = \frac{m}{2\beta} - \frac{1}{2}\mathsf{E}\|\mathbf{y} - \mathbf{X}\mathbf{w}\|^2 = 0 \Longleftrightarrow \beta = \frac{m}{\mathsf{E}\|\mathbf{y}-\mathbf{X}\mathbf{w}\|^2}.$

Потенциал поля точечного заряда: $\varphi = k\frac{q}{r}$.

Пусть имеется несколько зарядов $q_1, \ldots, q_l$ в точках $\mathbf{z}_1, \ldots, \mathbf{z}_l$.

Тогда $\varphi(\mathbf{x}) = k\sum\limits_{i=1}^{l} \frac{q_l}{\|\mathbf{x}-\mathbf{z}_l\|}$. По набору точек $\mathbf{x}_1, \ldots, \mathbf{x}_m$ и измеренным $y_i = \varphi(\mathbf{x}_i) - \underbrace{\varphi(\infty)}_{=0} + \varepsilon_i,\ \varepsilon_i \sim \mathcal{N}(\varepsilon_i|0,\ \beta^{-1})$

требуется оценить $\varphi(\mathbf{x})$ для $\mathbf{x}$ из тестовой выборки.



$\mathbf{y} = \mathbf{\Phi}\mathbf{w} + \boldsymbol{\varepsilon},\ \boldsymbol{\varepsilon} \sim \mathcal{N}(\boldsymbol{\varepsilon}|\mathbf{0},\ \beta^{-1}\mathbf{I})$, где

$\mathbf{\Phi} = \left\| \frac{1}{\delta + \|\mathbf{x}_i - \mathbf{x}_j\|} \right\|,\ i,\ j = \overline{1,\ m};$

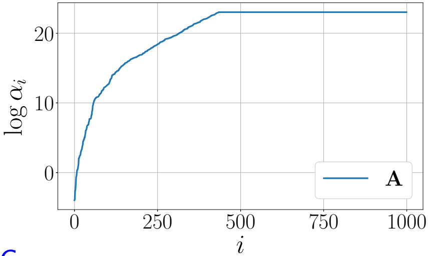
$\mathbf{w} \sim p(\mathbf{w}|\mathbf{A}) = \mathcal{N}(\mathbf{w}|\mathbf{0},\ \mathbf{A}^{-1})$.

Шаг 1: $p(\mathbf{y}_{\text{train}}|\mathbf{\Phi}_{\text{train}},\ \mathbf{A},\ \beta) \to \max\limits_{\mathbf{A},\ \beta}$ позволит отобрать признаки.

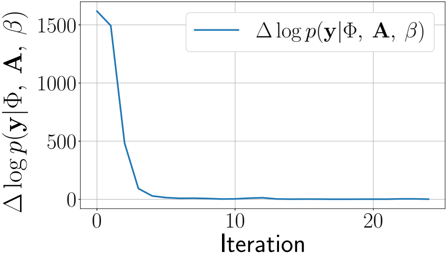
Шаг 2: Прогноз для тестовой выборки:

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

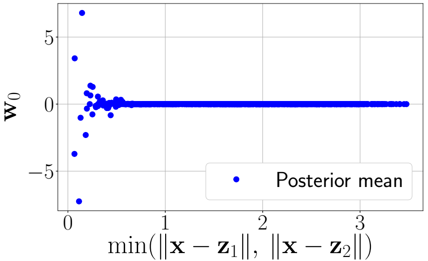
Оптимальный $\boldsymbol{\alpha}$



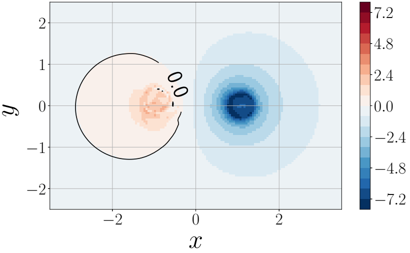
Обоснованность по итерациям



Среднее апостериорного распределения $\mathbf{w}_0$



Восстановленный потенциал



Литература

1. Bishop, Christopher M. "Pattern recognition and machine learning". Springer, New York (2006). Pp. 113-120, 161-171.
2. MacKay, David JC. Bayesian methods for adaptive models. Diss. California Institute of Technology, 1992.
3. Gelman, Andrew, et al. Bayesian data analysis, 3rd edition. Chapman and Hall/CRC, 2013.
4. Chen, Ming-Hui, and Joseph G. Ibrahim. "Conjugate priors for generalized linear models."Statistica Sinica (2003): 461-476.
5. Koller, D., & Friedman, N. (2009). Probabilistic graphical models: principles and techniques. MIT press.
6. Wainwright, M. J., & Jordan, M. I. (2008). Graphical models, exponential families, and variational inference. Now Publishers Inc.