Графические модели: введение

Александр Адуенко

5е марта 2024

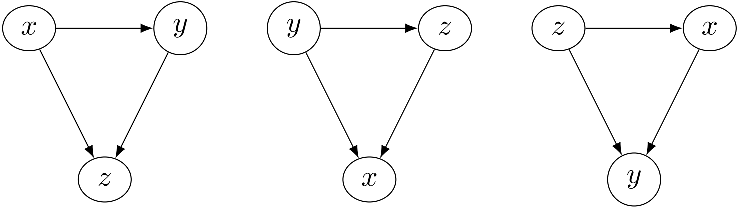
Содержание предыдущих лекций

- EM-алгоритм. Использование EM-алгоритма для отбора признаков в байесовской линейной регрессии.
- Вариационный EM-алгоритм и его использование для вывода в смеси моделей линейной регрессии.

Графические модели: примеры

Идея: Представим совместное распределение переменных в виде графа.
Пример: $p(x,\, y,\, z) = p(x)p(y|x)p(z|x,\, y)$.

Графическая вероятностная модель $p(x,\, y,\, z)$



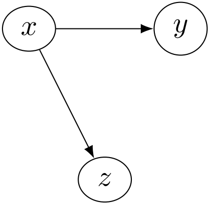
Вопрос 1: Чему соответствуют представления на средней и правой картинке?

$$p(x_1,\, \ldots,\, x_K) = p(x_1)p(x_2|x_1) \cdot \ldots \cdot p(x_K|x_1,\, \ldots,\, x_{K-1}).$$

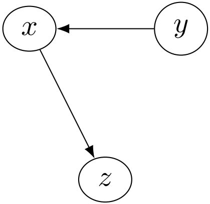
Вопрос 2: Для каких распределений выполнено разложение выше?
Вопрос 3: Какое представление получается для $p(x_1,\, \ldots,\, x_K)$ при таком разложении?

Графические модели: примеры

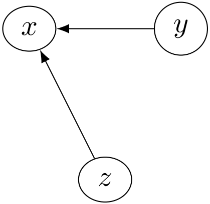
Идея: Представим совместное распределение переменных в виде графа.



Граф 1



Граф 2



Граф 3

Вопрос: Одинаковые ли совместные распределения соответствуют графическим представлениям выше?

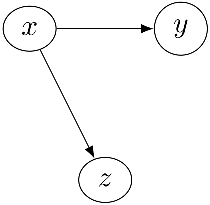
Граф 1: $p(x, y, z) = p(x)p(y|x)p(z|x)$;
Граф 2: $p(x, y, z) = p(y)p(x|y)p(z|x)$;
Граф 3: $p(x, y, z) = p(z)p(y)p(x|z, y)$.

Понятие условной независимости

Независимость: $p(y,\ z) = p(y)p(z)$.
Условная независимость: $p(y,\ z|x) = p(y|x)p(z|x)$.
Вопрос: Какое из определений более требовательное? Следует ли из независимости условная независимость и наоборот?

Свойство: $p(x|y,\ z) \propto p(x|y)p(x|z)$.

Граф 1: $p(x,\ y,\ z) = p(x)p(y|x)p(z|x)$;
Граф 2: $p(x,\ y,\ z) = p(y)p(x|y)p(z|x)$;
Граф 3: $p(x,\ y,\ z) = p(z)p(y)p(x|z,\ y)$.



Граф 1

$p(y,\ z|x) = p(z|x)p(y|x,\ z) = p(z|x)p(y|x) \Longrightarrow$
$(y,\ z)$ условно независимы при x.

$$p(y,\ z) = \int p(x)p(y|x)p(z|x)dx \neq p(y)p(z) \Longrightarrow$$

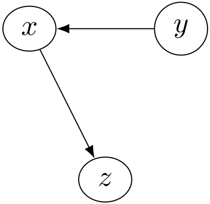
$(y,\ z)$ зависимы.

Пример:

$x \to \mathbf{w}$, $y \to y_1 = \mathbf{w}^{\mathsf{T}}\mathbf{x}_1 + \varepsilon_1$, $z \to y_2 = \mathbf{w}^{\mathsf{T}}\mathbf{x}_2 + \varepsilon_2$.

Понятие условной независимости (продолжение)

Граф 1: $p(x,\ y,\ z) = p(x)p(y|x)p(z|x)$;
Граф 2: $p(x,\ y,\ z) = p(y)p(x|y)p(z|x)$;
Граф 3: $p(x,\ y,\ z) = p(z)p(y)p(x|z,\ y)$.



Граф 2

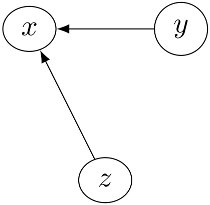
$p(y,\ z|x) = p(z|x)p(y|x,\ z) = p(z|x)p(y|x) \Longrightarrow$
$(y,\ z)$ условно независимы при x.

$$p(y,\ z) = \int p(y)p(x|y)p(z|x)dx \neq p(y)p(z) \Longrightarrow$$

$(y,\ z)$ зависимы.

Пример:

$x \to \mathbf{w};\ y \to \mathbf{A},\ \alpha_j \sim \Gamma(\nu,\ \eta);\ z \to y = \mathbf{w}^{\mathsf{T}}\mathbf{x} + \varepsilon$.



Граф 3

$p(y,\ z|x) \neq p(z|x)p(y|x)$, т.к.
$p(x|y,\ z) \neq p(x|y)p(x|z) \Longrightarrow$
$(y,\ z)$ условно зависимы при x.

$$p(y,\ z) = \int p(y)p(z)p(x|y,\ z)dx = p(y)p(z) \Longrightarrow$$

$(y,\ z)$ независимы.

Вопрос: Приведите пример модели с таким правдоподобием.

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 $p(\mathbf{w}_k|\mathbf{A}_k) = \mathcal{N}(\mathbf{w}_k|\mathbf{0},\ \mathbf{A}_k^{-1}),\ k = 1, \dots, K$;
- Для каждого объекта $\mathbf{x}_i$ выбрана модель f_{k_i}, которой он описывается, причем $p(k_i = k) = \pi_k$;
- Для каждого объекта $\mathbf{x}_i$ целевая переменная y_i определена в соответствии с моделью f_{k_i}: $y_i \sim \mathcal{N}(y_i|\mathbf{w}_{k_i}^{\mathsf{T}}\mathbf{x}_i,\ \sigma_{k_i}^2)$.

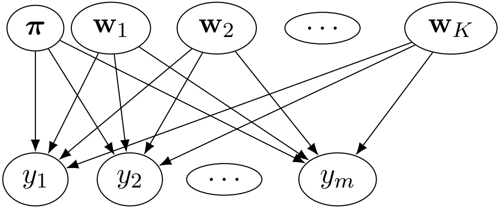
Совместное правдоподобие модели

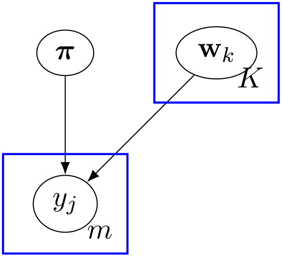
$$p(\mathbf{y},\ \mathbf{w}_1,\ \dots,\ \mathbf{w}_K,\ \boldsymbol{\pi}|\mathbf{X},\ \mathbf{A}_1,\ \dots,\ \mathbf{A}_K,\ \sigma_1^2,\ \dots,\ \sigma_K^2,\ \boldsymbol{\mu}) =$$
$$p(\boldsymbol{\pi}|\boldsymbol{\mu})\prod_{k=1}^{K}\mathcal{N}(\mathbf{w}_k|\mathbf{0},\ \mathbf{A}_k^{-1})\prod_{i=1}^{m}\left(\sum_{l=1}^{K}\pi_l\mathcal{N}(y_i|\mathbf{w}_l^{\mathsf{T}}\mathbf{x}_i,\ \sigma_l^2)\right).$$

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

$$p(\mathbf{y},\ \mathbf{w}_1,\ \dots,\ \mathbf{w}_K,\ \boldsymbol{\pi},\ \mathbf{Z}|\mathbf{X},\ \mathbf{A}_1,\ \dots,\ \mathbf{A}_K,\ \sigma_1^2,\ \dots,\ \sigma_K^2,\ \boldsymbol{\mu}) =$$
$$p(\boldsymbol{\pi}|\boldsymbol{\mu})\prod_{k=1}^{K}\mathcal{N}(\mathbf{w}_k|\mathbf{0},\ \mathbf{A}_k^{-1})\prod_{i=1}^{m}\prod_{l=1}^{K}\left(\pi_l\mathcal{N}(y_i|\mathbf{w}_l^{\mathsf{T}}\mathbf{x}_i,\ \sigma_l^2)\right)^{z_{il}}.$$

Граф 3: $p(x,\ y,\ z) = p(z)p(y)p(x|z,\ y)$.





Вопрос 1: Как в представлении учитывается то, что смесь составлена из моделей линейной регрессии?

Вопрос 2: Как учесть, что $p(\boldsymbol{\pi}) = \mathrm{Dir}(\boldsymbol{\mu})$?

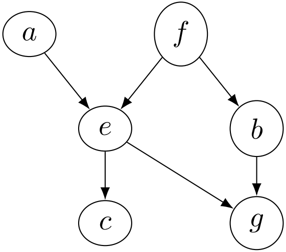
Вопрос 3: Как указать наличие наблюдаемого признакового описания $\mathbf{x}_1,\ \ldots,\ \mathbf{x}_m$ и гиперпараметров модели $\mathbf{A},\ \boldsymbol{\sigma}^2$?

Вопрос 4: Зачем нам графическое представление вероятностных моделей?

Критерий условной независимости D-separation

Рассмотрим две группы переменных A, B и проверим их условную независимость при условии группы переменных C.

D-separation. Группы переменных A и B условно независимы, если все неориентированные пути из A в B блокированы C.



Путь из вершины a в вершину b называется блокированным набором вершин C, если выполнено хотя бы одно из условий

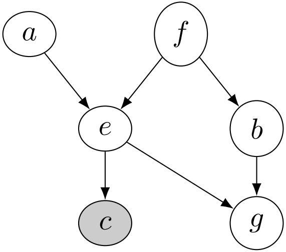
- Стрелки на пути встречаются перед-хвост или хвост-хвост в вершине $c \in C$;
- Стрелки на пути встречаются перед-перед в вершине x, такой, что $x \notin C$ и все ее ориентированные потомки $y \notin C$.

Пути из a в b: $a \to e \to g \leftarrow b$; $a \to e \leftarrow f \to b$.

Вопрос: Зависимы ли переменные a и b?

Рассмотрим две группы переменных A, B и проверим их условную независимость при условии группы переменных C.

D-separation. Группы переменных A и B условно независимы, если все неориентированные пути из A в B блокированы C.



Путь из вершины a в вершину b называется блокированным набором вершин C, если выполнено хотя бы одно из условий

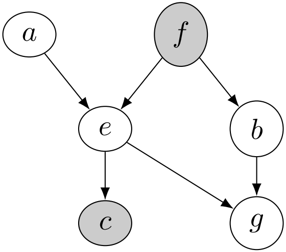
- Стрелки на пути встречаются перед-хвост или хвост-хвост в вершине $c \in C$;
- Стрелки на пути встречаются перед-перед в вершине x, такой, что $x \notin C$ и все ее ориентированные потомки $y \notin C$.

Пути из a в b: $a \to e \to g \leftarrow b$; $a \to e \leftarrow f \to b$.

Вопрос: Независимы ли переменные a и b при условии c?

Рассмотрим две группы переменных A, B и проверим их условную независимость при условии группы переменных C.

D-separation. Группы переменных A и B условно независимы, если все неориентированные пути из A в B блокированы C.



Путь из вершины a в вершину b называется блокированным набором вершин C, если выполнено хотя бы одно из условий

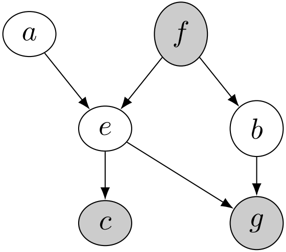
- Стрелки на пути встречаются перед-хвост или хвост-хвост в вершине $c \in C$;
- Стрелки на пути встречаются перед-перед в вершине x, такой, что $x \notin C$ и все ее ориентированные потомки $y \notin C$.

Пути из a в b: $a \to e \to g \leftarrow b$; $a \to e \leftarrow f \to b$.

Вопрос: Независимы ли переменные a и b при условии c, f?

Рассмотрим две группы переменных A, B и проверим их условную независимость при условии группы переменных C.

D-separation. Группы переменных A и B условно независимы, если все неориентированные пути из A в B блокированы C.



Путь из вершины a в вершину b называется блокированным набором вершин C, если выполнено хотя бы одно из условий

- Стрелки на пути встречаются перед-хвост или хвост-хвост в вершине $c \in C$;
- Стрелки на пути встречаются перед-перед в вершине x, такой, что $x \notin C$ и все ее ориентированные потомки $y \notin C$.

Пути из a в b: $a \to e \to g \leftarrow b$; $a \to e \leftarrow f \to b$.

Вопрос: Независимы ли переменные a и b при условии c, f, g?

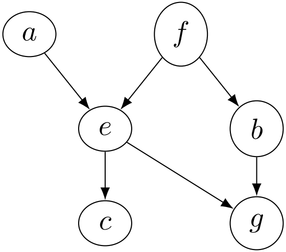
Графические модели: ориентированное и неориентированное представление

Александр Адуенко

17е марта 2024

- EM-алгоритм. Использование EM-алгоритма для отбора признаков в байесовской линейной регрессии.
- Вариационный EM-алгоритм и его использование для вывода в смеси моделей линейной регрессии.
- Ориентированные графические модели и их представление plate notation. Критерий условной независимости D-separation.

Условная независимость как фильтр



Условия условной независимости:

- $a \perp b | \emptyset$;
- $a \perp f | \emptyset$;
- $a \not\perp f | e$;
- $a \not\perp c | \emptyset$;
- $a \perp c | e$;

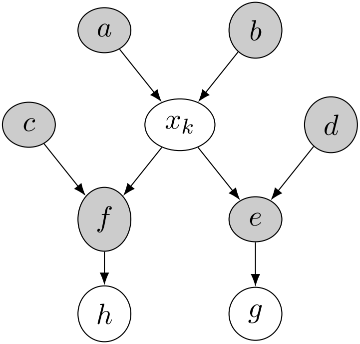
$\vdots$

-

Утверждение: Множество распределений, представимых ориентированным графом заданного вида, совпадает с множеством распределений, удовлетворяющим всем условиям условной независимости, им порожденным.

Марковское «одеяло»

$$p(x_1,\,\dots,\,x_K)=\prod_k p(x_k|Pa_k).$$



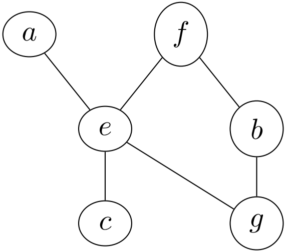
$$p(x_k|\mathbf{x}_{j\neq k})=\frac{\prod_l p(x_l|Pa_l)}{\int\prod_l p(x_l|Pa_l)dx_k}.$$

$$p(x_1,\,\dots,\,x_K)=p(a)p(b)p(x_k|a,b)p(c)p(d)$$
$$p(e|d,\,x_k)p(f|c,\,x_k)p(g|e)p(h|f).$$

$$p(x_k|\mathbf{x}_{j\neq i})=$$
$$\frac{p(x_k|a,b)p(e|d,\,x_k)p(f|c,\,x_k)}{\int p(x_k|a,b)p(e|d,\,x_k)p(f|c,\,x_k)dx_k}.$$

Вопрос: От каких переменных действительно зависит $x_k|\mathbf{x}_{j\neq k}$?

Неориентированные графические модели $\equiv$ марковские случайные поля.



Идея: Определить критерий условной независимости в терминах разделимости, не d-разделимости.

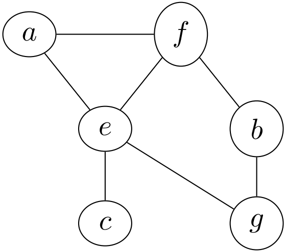
Вопрос: Независимы ли a и b

- Безусловно?
- При условии e?
- При условии f?
- При условии $f,\ g$?

Вопрос: Что есть марковское «одеяло» для неориентированных графических моделей?

Замечание: Если x_i и x_j не соединены ребром, то они независимы при условии всех остальных переменных:
$p(x_i,\ x_j|\mathbf{x}_{\setminus\{i,\ j\}}) = p(x_i|\mathbf{x}_{\setminus\{i,\ j\}})p(x_j|\mathbf{x}_{\setminus\{i,\ j\}})$.

Замечание: Если x_i и x_j не соединены ребром, то они независимы при условии всех остальных переменных:
$p(x_i, x_j|\mathbf{x}_{\setminus\{i, j\}}) = p(x_i|\mathbf{x}_{\setminus\{i, j\}})p(x_j|\mathbf{x}_{\setminus\{i, j\}})$.



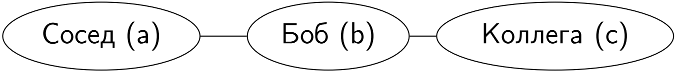
$$p(a, b, c, e, f, g) = \frac{1}{Z}\psi_{afe}(a, f, e)\psi_{ec}(e, c)\psi_{eg}(e, g)\psi_{bg}(b, g)\psi_{bf}(b, f).$$

$$Z = \int \prod_i \psi_{C_i}(\mathbf{x}_{C_i}).$$

Вопрос: Какими свойствами должны обладать $\psi_{C_i}(\mathbf{x}_{C_i})$, чтобы задать корректное распределение?

Теорема (Hammersley-Clifford). Предположим, что все потенциалы строго положительны $\psi_C(\mathbf{x}_C) > 0 \, \forall \, \mathbf{x}_C$. Тогда факторизация по максимальным кликам графа задает то же множество распределений, что и набор условий условной независимости в терминах графовой разделимости.

Пусть моделируется заболеваемость простудой для трех человек: Боба, его соседа и коллеги.
Метка 0 соответствует тому, что человек здоров, а 1 – болезни.



$$p(a,\ b,\ c) = \frac{1}{Z}\psi_{ab}(a,\ b)\psi_{bc}(b,\ c),\ a,\ b,\ c \in \{0,\ 1\}.$$

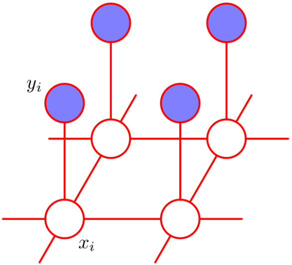
$$\psi_{a,\,b}(a,\ b) = \begin{cases} 20,\ a = 0,\ b = 0, \\ 2,\ a = 1,\ b = 1, \\ 1,\ a = 0,\ b = 1, \\ 1,\ a = 1,\ b = 0. \end{cases},\ \psi_{b,\,c}(b,\ c) = \begin{cases} 5,\ b = 0,\ c = 0, \\ 0.7,\ b = 1,\ c = 1, \\ 5,\ b = 0,\ c = 1, \\ 0.1,\ b = 1,\ c = 0. \end{cases}$$

Идея: $\psi_{C_i}(\mathbf{x}_{C_i}) = \exp(-E_{C_i}(\mathbf{x}_{C_i}))$.

Тогда $\log p(\mathbf{x}) \propto -E(\mathbf{x}) = -\sum_i E_{C_i}(\mathbf{x}_{C_i})$.

Пример: Пусть имеется бинарное изображение $\mathbf{y}$, $y_i \in \{-1,\ 1\}$, которое зашумлено. Требуется восстановить исходное изображение $\mathbf{x}$.

$$E(\mathbf{x},\ \mathbf{y}) = h\sum_i x_i - \beta \sum_{(i,\ j)\in\varepsilon} x_i x_j - \eta \sum_i x_i y_i.$$



Графическая модель $p(\mathbf{x},\ \mathbf{y})$ [Bishop, 2006]

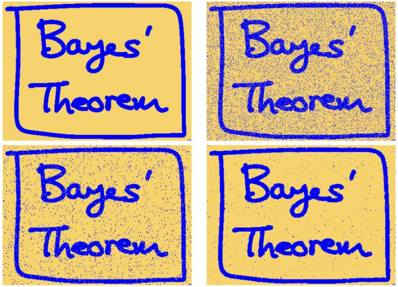
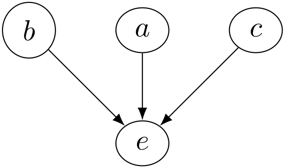


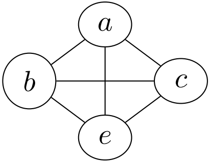
Иллюстрация шумоподавления [Bishop, 2006]

Связь между ориентированными и неориентированными моделями

Вопрос: Можно ли по ориентированной модели построить соответствующую ей неориентированную и наоборот?



$$p(a,\ b,\ c,\ e) = p(a)p(b)p(c)p(e|a,\ b,\ c).$$



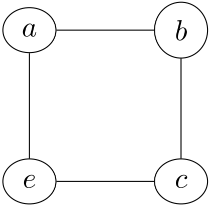
$$p(a,\ b,\ c,\ e) = \psi(a,\ b,\ c,\ e).$$

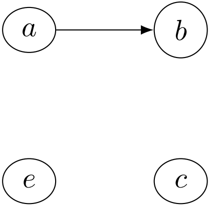
Конверсия из ориентированного в неориентированный граф:

- Соединяем ребрами попарно всех родителей каждой вершины («moralization»);
- Удаляем ориентацию на ребрах и получаем неориентированный граф;
- Для каждой клики задаем потенциал $\psi_{C_i}(\mathbf{x}_{C_i}) = 1$;
- Для каждого условного распределения, умножаем на него потенциал той клики, к которой относятся все переменные из него.

Из неориентированной модели к ориентированной

Вопрос: Верно ли, что ориентированная модель позволяет всегда «более аккуратно» задать зависимости, не потеряв условных независимостей между переменными?



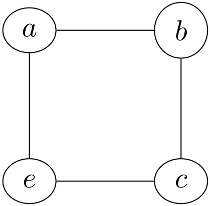


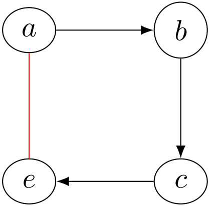
$p(a,\ b,\ c,\ e) = \psi_1(a,\ b)\psi_2(b,\ c)\psi_3(c,\ e)\psi_4(e,\ a).$

Свойства зависимости и независимости переменных:

- $a \not\perp c|\emptyset$;
- $a \perp c|b,\ e$;
- $b \perp e|a,\ c$;
- $\vdots$

Из неориентированной модели к ориентированной 2



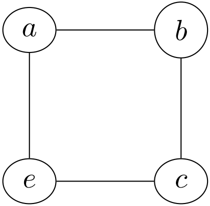


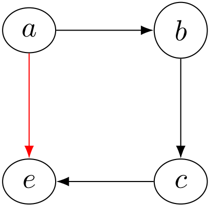
$p(a,\, b,\, c,\, e) = \psi_1(a,\, b)\psi_2(b,\, c)\psi_3(c,\, e)\psi_4(e,\, a)$.
Свойства зависимости и независимости переменных:

- $a \not\perp c|\emptyset$;
- $a \perp c|b,\, e$;
- $b \perp e|a,\, c$;
- $\vdots$

Вопрос: Могли ли мы ориентировать ребро $(b,\, c)$ иначе и выполнить условие $a \perp c|b,\, e$?

Из неориентированной модели к ориентированной 3



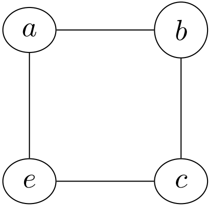


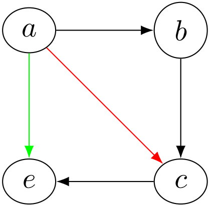
$p(a,\ b,\ c,\ e) = \psi_1(a,\ b)\psi_2(b,\ c)\psi_3(c,\ e)\psi_4(e,\ a)$.

Свойства зависимости и независимости переменных:

- $a \not\perp c|\emptyset$;
- $a \perp c|b,\ e$;
- $b \perp e|a,\ c$;
- $\vdots$

Вопрос: Соответствует ли граф справа правдоподобию? Нет ли дополнительной условной независимости?





Слева: $p(a,\ b,\ c,\ e) = \psi_1(a,\ b)\psi_2(b,\ c)\psi_3(c,\ e)\psi_4(e,\ a)$.
Свойства зависимости и независимости переменных:
$a \perp c|b,\ e,\ b \perp e|a,\ c,\ \dots$

Справа: $p(a,\ b,\ c,\ e) = p(a)p(b|a)p(c|b,\ a)p(e|a,\ c)$.
Замечание 1: Без ребра из a в c, была дополнительная условная независимость $a \perp c \mid b$.
Вопрос 1: Почему нельзя было бы провести вместо $a \to c$ ребра $c \to a,\ e \to b$?
Вопрос 2: Помогло ли бы ребро $b \to e$ убрать $a \perp c \mid b$?
Вопрос: Какую условную независимость мы потеряли?

Графические модели: факторные графы и вывод в графических моделях

Александр Адуенко

19е марта 2024

Содержание предыдущих лекций

- EM-алгоритм. Использование EM-алгоритма для отбора признаков в байесовской линейной регрессии.
- Вариационный EM-алгоритм и его использование для вывода в смеси моделей линейной регрессии.
- Ориентированные графические модели и их представление plate notation. Критерий условной независимости d-separation.
- Неориентированные графические модели и их связь с ориентированными.

Пример вывода в графической модели

$$\begin{tikzcd} x_1 \arrow[r] & x_2 \arrow[r] & \cdots \arrow[r] & x_{N-1} \arrow[r] & x_N \end{tikzcd}$$

$$p(\mathbf{x}) = p(x_1)p(x_2|x_1)\cdot\ldots\cdot p(x_{N-1}|x_{N-2})p(x_N|x_{N-1}).$$

$$\begin{tikzcd} x_1 \arrow[r, dash] & x_2 \arrow[r, dash] & \cdots \arrow[r, dash] & x_{N-1} \arrow[r, dash] & x_N \end{tikzcd}$$

$$p(\mathbf{x}) = \frac{1}{Z}\psi_{12}(x_1,\, x_2)\psi_{23}(x_2,\, x_3)\cdot\ldots\cdot\psi_{N-1,N}(x_{N-1},\, x_N).$$

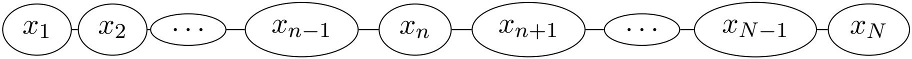
Пусть для простоты $x_i \in \{1,\, \ldots,\, K\}$ и требуется найти $p(x_n)$.

$$p(x_n) = \sum_{x_1}\sum_{x_2}\ldots\sum_{x_{n-1}}\sum_{x_{n+1}}\ldots\sum_{x_N} p(\mathbf{x}).$$

Вопрос: Сколько нужно операций, чтобы выполнить суммирование для подсчета $p(x_n)$, то есть $\mathsf{P}(x_n = k)$, $k = 1,\, \ldots,\, K$?

Замечание: Формула суммирования не учитывает структуру модели.

Пример вывода в графической модели



$$p(\mathbf{x}) = \frac{1}{Z}\psi_1(x_1,\, x_2)\psi_2(x_2,\, x_3)\cdot \ldots \cdot \psi_{N-1}(x_{N-1},\, x_N).$$

$$p(x_n) = \sum_{x_1}\sum_{x_2}\ldots\sum_{x_{n-1}}\sum_{x_{n+1}}\ldots\sum_{x_N} p(\mathbf{x}).$$

$$p(x_n) = \frac{1}{Z}\sum_{x_{n-1}}\psi_{n-1}(x_{n-1},\, x_n)\cdot\ldots\left[\sum_{x_2}\psi_2(x_2,\, x_3)\cdot\left[\sum_{x_1}\psi_1(x_1,\, x_2)\right]\right]\times$$

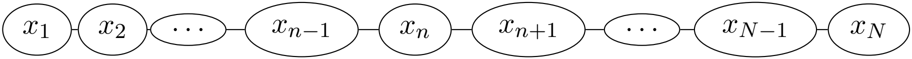
$$\sum_{x_{n+1}}\psi_n(x_n,\, x_{n+1})\cdot\ldots\left[\sum_{x_{N-1}}\psi_{N-2}(x_{N-2},\, x_{N-1})\left[\sum_{x_N}\psi_{N-1}(x_{N-1},\, x_N)\right]\right].$$

$$p(x_n) = \frac{1}{Z}\mu_\alpha(x_n)\mu_\beta(x_n),$$ где

$\mu_\alpha(x_n)$ – сообщение вперед от x_{n-1} к x_n;

$\mu_\beta(x_n)$ – сообщение назад от x_{n+1} к x_n.

Пример вывода в графической модели



$$p(x_n) = \frac{1}{Z}\sum_{x_{n-1}} \psi_{n-1}(x_{n-1},\, x_n)\cdot\ldots\left[\sum_{x_2}\psi_2(x_2,\, x_3)\cdot\left[\sum_{x_1}\psi_1(x_1,\, x_2)\right]\right]\times$$

$$\sum_{x_{n+1}}\psi_n(x_n,\, x_{n+1})\cdot\ldots\left[\sum_{x_{N-1}}\psi_{N-2}(x_{N-2},\, x_{N-1})\left[\sum_{x_N}\psi_{N-1}(x_{N-1},\, x_N)\right]\right].$$

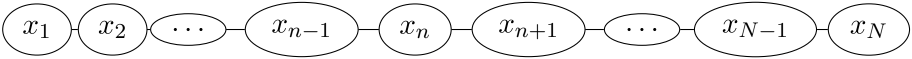
$$p(x_n) = \frac{1}{Z}\mu_\alpha(x_n)\mu_\beta(x_n).$$

$$\mu_\alpha(x_n) = \sum_{x_{n-1}}\psi_{n-1}(x_{n-1},\, x_n)\left[\sum_{x_{n-2}}\ldots\right] = \sum_{x_{n-1}}\psi_{n-1}(x_{n-1},\, x_n)\mu_\alpha(x_{n-1}).$$

$$\mu_\beta(x_n) = \sum_{x_{n+1}}\psi_n(x_n,\, x_{n+1})\left[\sum_{x_{n+2}}\ldots\right] = \sum_{x_{n+1}}\psi_n(x_n,\, x_{n+1})\mu_\beta(x_{n+1}).$$

Вопрос: Как определить базу рекурсии?

Пример вывода в графической модели



$$p(x_n) = \frac{1}{Z}\mu_\alpha(x_n)\mu_\beta(x_n).$$

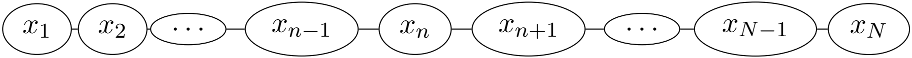
$$\mu_\alpha(x_n) = \sum_{x_{n-1}} \psi_{n-1}(x_{n-1},\, x_n)\mu_\alpha(x_{n-1}).$$

$$\mu_\beta(x_n) = \sum_{x_{n+1}} \psi_n(x_n,\, x_{n+1})\mu_\beta(x_{n+1}).$$

База рекурсии: $\mu_\alpha(x_1) = 1,\ \mu_\beta(x_N) = 1.$

Вопрос 1: Сколько операций требуется для подсчета $\mu_\alpha(x_n),\ \mu_\beta(x_n)$?
Вопрос 2: Как определить Z?
Вопрос 3: Как обобщить результат на случай непрерывных переменных?
Вопрос 4: Как изменится вывод, если часть переменных наблюдаемы, то есть ищем $p(x_n | x_{i_1},\, \ldots,\, x_{i_l})$?

Пример вывода в графической модели



$$p(x_n) = \frac{1}{Z}\mu_\alpha(x_n)\mu_\beta(x_n).$$

$$\mu_\alpha(x_n) = \sum_{x_{n-1}} \psi_{n-1}(x_{n-1},\, x_n)\mu_\alpha(x_{n-1}).$$

$$\mu_\beta(x_n) = \sum_{x_{n+1}} \psi_n(x_n,\, x_{n+1})\mu_\beta(x_{n+1}).$$

База рекурсии: $\mu_\alpha(x_1) = 1,\ \mu_\beta(x_N) = 1.$

Вопрос: Как найти $p(x_l),\ \forall\, l \neq n$?

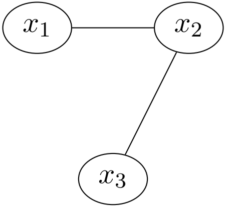
Идея: Сосчитать $\mu_\alpha(x_q),\ q = 1,\, \dots,\, N$ и $\mu_\beta(x_q),\ q = N,\, \dots,\, 1.$

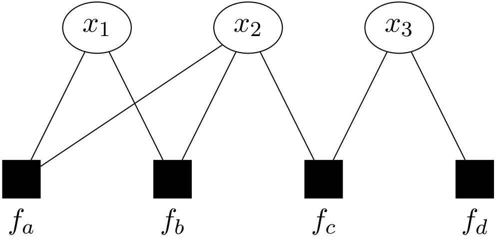
Задание: Показать $p(x_{n-1},\, x_n) = \dfrac{1}{Z}\mu_\alpha(x_{n-1})\psi_{n-1}(x_{n-1},\, x_n)\mu_\beta(x_n).$

Фактор-графы и их построение по графической модели

Идея: Построить общее представление для ориентированных и неориентированных моделей.

$$p(\mathbf{x}) = \prod_s f_s(\mathbf{x}_s).$$





Вопрос: Задает ли граф справа другой набор условных независимостей, чем граф слева?

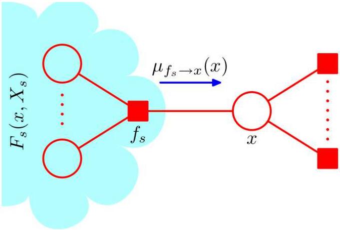
Алгоритм Sum-Product вывода в ациклических ГМ

Утверждение: Если исходная графическая модель есть направленное или ненаправленное дерево, то для нее можно построить ациклический фактор-граф.

Найти: $p(x) = \sum_{\mathbf{x}\backslash x} p(\mathbf{x})$.

$p(\mathbf{x}) = \frac{1}{Z}\prod_s f_s(\mathbf{x}_s) =$

$\frac{1}{Z}\prod_{s\in N(x)} F_s(x,\ X_s) = \frac{1}{Z}\tilde{p}(x)$.



Фактор-граф в окрестности вершины x [Bishop, 2006]

$$\tilde{p}(x) = \sum_{\mathbf{x}\backslash x}\tilde{p}(\mathbf{x}) = \sum_{\mathbf{x}\backslash x}\prod_{s\in N(x)} F_s(x,\ X_s) = \prod_{s\in N(x)}\sum_{\mathbf{x}\backslash x} F_s(x,\ X_s) =$$

$$\prod_{s\in N(x)}\sum_{X_s} F_s(x,\ X_s) = \prod_{s\in N(x)}\mu_{f_s\to x}(x).$$

Алгоритм Sum-Product вывода в ациклических ГМ

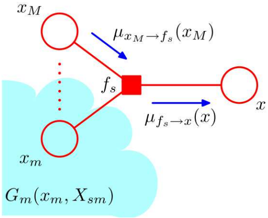
$$F_s(x,\ X_s) = f_s(x,\ x_1,\ \dots,\ x_M)G_1(x_1,\ X_{s1})\cdot\ldots\cdot G_M(x_M,\ X_{sM}).$$

$$\mu_{f_s\to x}(x) = \sum_{X_s} F_s(x,\ X_s) =$$

$$\sum_{x_{1:M}} f_s(x,\ x_{1:M}) \prod_{m\in N(f_s)\setminus x}\left[\sum_{X_{sm}} G_m(x_m,\ X_{sm})\right] =$$

$$\sum_{x_{1:M}} f_s(x,\ x_{1:M}) \prod_{m\in N(f_s)\setminus x} \mu_{x_m\to f_s}(x_m),\ \text{где}$$

$$\mu_{x_m\to f_s}(x_m) = \sum_{X_{sm}} G_m(x_m,\ X_{sm}).$$



Фактор-граф в окрестности вершины f_s [Bishop, 2006]

$$G_m(x_m,\ X_{sm}) = \prod_{l\in N(x_m)\setminus f_s} F_l(x_m,\ X_{ml}).$$

$$\mu_{x_m\to f_s}(x_m) = \sum_{X_{sm}} \prod_{l\in N(x_m)\setminus f_s} F_l(x_m,\ X_{ml}) =$$

$$\prod_{l\in N(x_m)\setminus f_s}\left[\sum_{X_{ml}} F_l(x_m,\ X_{ml})\right] = \prod_{l\in N(x_m)\setminus f_s} \mu_{f_l\to x_m}(x_m).$$

Алгоритм Sum-Product вывода в ациклических ГМ

Получаем следующие формулы пересчета сообщений:

- $\mu_{x_m \to f_s}(x_m) = \prod\limits_{l \in N(x_m) \setminus f_s} \mu_{f_l \to x_m}(x_m)$;
- $\mu_{f_s \to x}(x) = \sum\limits_{x_{1:M}} f_s(x,\, x_{1:M}) \prod\limits_{m \in N(f_s) \setminus x} \mu_{x_m \to f_s}(x_m)$.

Алгоритм:

1. Объявляем вершину x корнем;
2. От листьев фактор-графа движемся к корню, пересылая сообщения по правилам выше;
3. По достижении корня имеем: $p(x) = \dfrac{1}{Z} \prod\limits_{s \in N(x)} \mu_{f_s \to x}(x)$.

База рекурсии (сообщения от листьев): $\mu_{x \to f} = 1,\ \mu_{f \to x} = f(x)$.

Вопрос 1: Как показать, что процедура работает, то есть все вершины получат достаточно сообщений, чтобы отправить своё?

Вопрос 2: Как получить $p(x_l)\ \forall\, x_l \neq x$?

Вопрос 3: Как определить нормировочную постоянную Z?

Вопрос 4: Как получить $p(\mathbf{x}_s)$?

Графические модели: Скрытые марковские модели

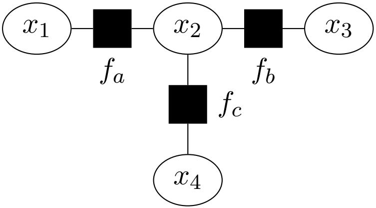
Александр Адуенко

31е марта 2024

Содержание предыдущих лекций

- EM-алгоритм. Использование EM-алгоритма для отбора признаков в байесовской линейной регрессии.
- Вариационный EM-алгоритм и его использование для вывода в смеси моделей линейной регрессии.
- Ориентированные графические модели и их представление plate notation. Критерий условной независимости d-separation.
- Неориентированные графические модели и их связь с ориентированными.
- Факторные графы и алгоритм Sum-Product для вывода в ациклических графических моделях.

Пример работы алгоритма Sum-Product



$p(\mathbf{x}) = f_a(x_1, x_2) f_b(x_2, x_3) f_c(x_2, x_4).$

Прямой проход (x_3 – корень):

$$\mu_{x_1 \to f_a}(x_1) = 1, \ \mu_{x_4 \to f_c}(x_4) = 1,$$

$$\mu_{f_a \to x_2}(x_2) = \sum_{x_1} f_a(x_1, x_2) \mu_{x_1 \to f_a}(x_1) = \sum_{x_1} f_a(x_1, x_2),$$

$$\mu_{f_c \to x_2}(x_2) = \sum_{x_4} f_c(x_2, x_4),$$

$$\mu_{x_2 \to f_b}(x_2) = \mu_{f_a \to x_2}(x_2) \mu_{f_c \to x_2}(x_2),$$

$$\mu_{f_b \to x_3}(x_3) = \sum_{x_2} f_b(x_2, x_3) \mu_{x_2 \to f_b}(x_2).$$

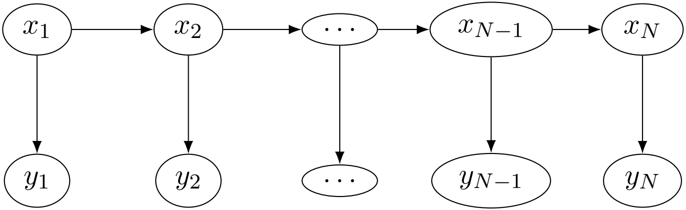
Обратный проход:

$$\mu_{x_3 \to f_b}(x_3) = 1, \ \mu_{f_b \to x_2}(x_2) = \sum_{x_3} f_b(x_2, x_3),$$

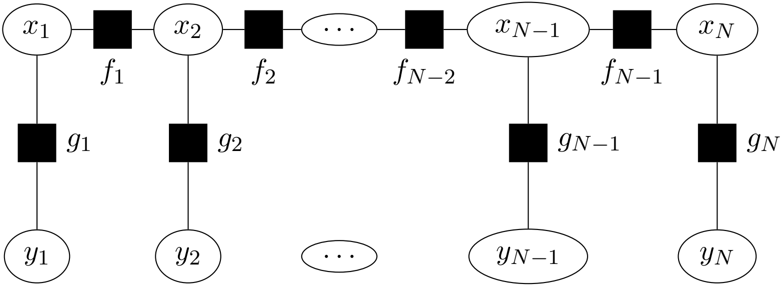
$$\mu_{x_2 \to f_a}(x_2) = \mu_{f_b \to x_2}(x_2) \mu_{f_c \to x_2}(x_2), \ \mu_{x_2 \to f_c}(x_2) = \mu_{f_a \to x_2}(x_2) \mu_{f_b \to x_2}(x_2),$$

$$\mu_{f_a \to x_1}(x_1) = \sum_{x_2} f_a(x_1, x_2) \mu_{x_2 \to f_a}(x_2),$$

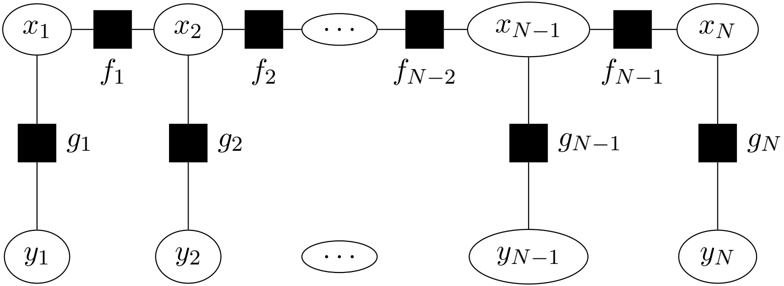
$$\mu_{f_c \to x_4}(x_4) = \sum_{x_2} f_c(x_2, x_4) \mu_{x_2 \to f_c}(x_2).$$



$$p(\mathbf{x},\ \mathbf{y}) = \prod_{i=1}^{N} p(x_i|x_{i-1}) \prod_{i=1}^{N} p(y_i|x_i), x_0 = \emptyset.$$



Скрытые марковские модели 2



$$p(\mathbf{x},\ \mathbf{y}) = \prod_{i=1}^{N} p(x_i|x_{i-1}) \prod_{i=1}^{N} p(y_i|x_i), x_0 = \emptyset.$$

Замечание: С помощью алгоритма Sum-Product можем найти $p(x_i|\mathbf{y})\,\forall\, i$.

Вопрос 1: Как найти $\mathbf{x}^* = \arg\max_{\mathbf{x}} p(\mathbf{x}|\mathbf{y})$?

Идея: $\tilde{x}_i^* = \arg\max_{x_i} p(x_i|\mathbf{y})$.

Вопрос 2: Верно ли, что $\mathbf{x}^* = \tilde{\mathbf{x}}^*$?

Пример скрытой марковской модели

Пусть $x_t \in [\text{Подъем, Зависание, Спуск}]$ есть состояние воздушного шара, а $\sqrt{y_t}$ полная скорость.

$$\begin{tikzcd}
x_1 \arrow[r] \arrow[d] & x_2 \arrow[r] \arrow[d] & \cdots \arrow[r] \arrow[d] & x_{N-1} \arrow[r] \arrow[d] & x_N \arrow[d] \\
y_1 & y_2 & \cdots & y_{N-1} & y_N
\end{tikzcd}$$

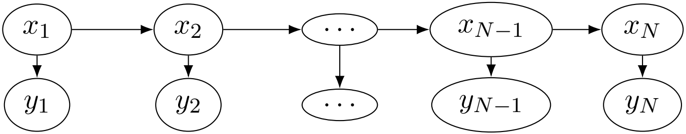
$$p(\mathbf{x},\ \mathbf{y}) = \prod_{i=1}^{N} p(x_i|x_{i-1}) \prod_{i=1}^{N} p(y_i|x_i), x_0 = \emptyset.$$

$$p(x_1) = \boldsymbol{\pi} = [1,\ 0,\ 0]^{\mathsf{T}},\ p(x_t|x_{t-1}) = \mathbf{A} = \begin{pmatrix} & \text{П} & \text{З} & \text{С} \\ \text{П} & 0.98 & 0.02 & 0 \\ \text{З} & 0 & 0.8 & 0.2 \\ \text{С} & 0 & 0 & 1 \end{pmatrix}.$$

$$y_t|x_t = v^2_{\text{ветер}} + v^2_{\text{вертик.}} = \underbrace{\varepsilon_t}_{\sim\mathcal{N}(5,\,3^2)^2} + \underbrace{v^2_{\text{вертик.}}|x_t}_{\text{П: }1,\ \text{З: }0,\ \text{С: }4}.$$

Вопрос: Что можно сказать про $x_t^* = \arg\max_{x_t} p(x_t|\mathbf{y})$?

Алгоритм Витерби



$$p(\mathbf{x},\ \mathbf{y}) = \prod_{i=1}^{N} p(x_i|x_{i-1}) \prod_{i=1}^{N} p(y_i|x_i), x_0 = \emptyset.$$

Задача: $p(\mathbf{x}|\mathbf{y}) \to \max\limits_{\mathbf{x}} \equiv p(\mathbf{x},\ \mathbf{y}) \to \max\limits_{\mathbf{x}}.$

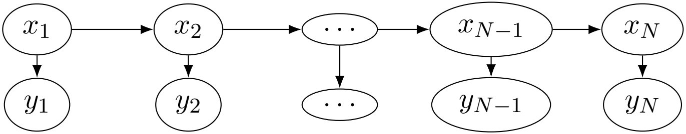
$$V_{1,\,k} = \pi_k p(y_1|x_1 = k),$$
$$V_{t,\,k} = \max_{j \in S} V_{t-1,\,j}\, a_{jk}\, p(y_t|x_t = k).$$

Вопрос 1: Что показывает $V_{t,\,k}$?

Вопрос 2: Как изменятся формулы для $V_{t,\,k}$, если y_t ненаблюдаемо?

Вопрос 3: Что мы получим в $V_{N,\,k}$?

Алгоритм Витерби 2



$$p(\mathbf{x},\ \mathbf{y}) = \prod_{i=1}^{N} p(x_i|x_{i-1}) \prod_{i=1}^{N} p(y_i|x_i), x_0 = \emptyset.$$

Задача: $p(\mathbf{x}|\mathbf{y}) \to \max\limits_{\mathbf{x}} \equiv p(\mathbf{x},\ \mathbf{y}) = p(\mathbf{y}|\mathbf{x})p(\mathbf{x}) \to \max\limits_{\mathbf{x}}$.

$$V_{1,\,k} = \pi_k p(y_1|x_1 = k),$$
$$V_{t,\,k} = \max_{j\in S} V_{t-1,\,j}\, a_{jk}\, p(y_t|x_t = k).$$

$V_{N,\,k}$ – вероятность наиболее вероятной последовательности состояний, оканчивающейся в $x_N = k$, то есть $V_{N,\,k} = \max\limits_{\mathbf{x}\setminus x_N} p(\mathbf{x},\ \mathbf{y}|x_N = k)$.

Замечание: $x_N^* = \arg\max\limits_{k} V_{N,\,k}$.

Вопрос: Как получить x_{N-1}^* из $\mathbf{x}^* = \arg\max\limits_{\mathbf{x}} p(\mathbf{x},\ \mathbf{y})$?

Идея: Запомнить j^* из $V_{N,\,k} = \max\limits_{j\in S} V_{N-1,\,j}\, a_{jk}\, p(y_N|x_N = k)$.

Задача Sum-Product

$$p(\mathbf{x}) = \frac{1}{Z}\prod_s f_s(\mathbf{x}_s)$$

Найти: $p(x) = \sum_{\mathbf{x}\setminus x} p(\mathbf{x})$.

Свойство:

$ab + ac = a(b + c)$.

Задача Max-Sum

$$g(\mathbf{x}) = \log p(\mathbf{x}) = C + \sum_s \log f_s(\mathbf{x}_s)$$

Найти: $g(x) = \max_{\mathbf{x}\setminus x} g(\mathbf{x})$.

Свойство:

$\max(a + b,\ a + c) = a + \max(b,\ c)$.

Формулы пересчета сообщений для Sum-Product:

- $\mu_{x_m \to f_s}(x_m) = \prod_{l \in N(x_m)\setminus f_s} \mu_{f_l \to x_m}(x_m)$;
- $\mu_{f_s \to x}(x) = \sum_{x_{1:M}} f_s(x,\ x_{1:M}) \prod_{m \in N(f_s)\setminus x} \mu_{x_m \to f_s}(x_m)$.

Формулы пересчета сообщений для Max-Sum:

- $\mu_{x_m \to f_s}(x_m) = \sum_{l \in N(x_m)\setminus f_s} \mu_{f_l \to x_m}(x_m)$;
- $\mu_{f_s \to x}(x) = \max_{x_{1:M}} \log f_s(x,\ x_{1:M}) + \sum_{m \in N(f_s)\setminus x} \mu_{x_m \to f_s}(x_m)$.

Алгоритм Max-Sum 2

$$g(\mathbf{x}) = \log p(\mathbf{x}) = C + \sum_s \log f_s(\mathbf{x}_s)$$

Найти: $g(x) = \max_{\mathbf{x}\setminus x} g(\mathbf{x})$.

Формулы пересчета сообщений для Max-Sum:

- $\mu_{x_m \to f_s}(x_m) = \sum_{l \in N(x_m)\setminus f_s} \mu_{f_l \to x_m}(x_m)$;
- $\mu_{f_s \to x}(x) = \max_{x_{1:M}} \log f_s(x,\ x_{1:M}) + \sum_{m \in N(f_s)\setminus x} \mu_{x_m \to f_s}(x_m)$.

Сообщения из листьев: $\mu_{x\to f} = 0,\ \mu_{f\to x} = \log f(x)$.

Вопрос 1: Как получить $p(\mathbf{x}^*) = \max_{\mathbf{x}} p(\mathbf{x})$?

Вопрос 2: Как получить $\mathbf{x}^* = \arg\max \log p(\mathbf{x})$?

Алгоритм Max-Sum 3

$$g(\mathbf{x}) = \log p(\mathbf{x}) = C + \sum_s \log f_s(\mathbf{x}_s)$$

Найти: $g(x) = \max\limits_{\mathbf{x}\backslash x} g(\mathbf{x})$.

Формулы пересчета сообщений для Max-Sum:

- $\mu_{x_m \to f_s}(x_m) = \sum\limits_{l \in N(x_m)\backslash f_s} \mu_{f_l \to x_m}(x_m)$;
- $\mu_{f_s \to x}(x) = \max\limits_{x_{1:M}} \left[\log f_s(x,\ x_{1:M}) + \sum\limits_{m \in N(f_s)\backslash x} \mu_{x_m \to f_s}(x_m)\right]$.

Сообщения из листьев: $\mu_{x \to f} = 0,\ \mu_{f \to x} = \log f(x)$.

$\max\limits_{\mathbf{x}} g(\mathbf{x}) = \max\limits_{x_R} g(x_R),\ x_R^* = \arg\max\limits_{x_R} g(x_R)$, где x_R – корень ф-дерева.

Вопрос 1: $x_i^* = \arg\max\limits_{x_i} g(x_i)$ для всех вершин для получения $\mathbf{x}^*$?

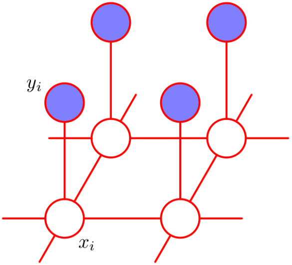
Идея: Хранить конфигурацию $x_{1:M}$, доставляющую максимум в $\mu_{f_s \to x}$.

Вопрос 2: Сколько потребуется памяти для хранения таких конфигураций?

Иллюстрация работы алгоритма Max-Sum

Пример: Пусть имеется бинарное изображение $\mathbf{y}$, $y_i \in \{-1, 1\}$, которое зашумлено. Требуется восстановить исходное изображение $\mathbf{x}$.

$$E(\mathbf{x}, \mathbf{y}) = h\sum_i x_i - \beta \sum_{(i,j)\in\varepsilon} x_i x_j - \eta \sum_i x_i y_i.$$



Графическая модель $p(\mathbf{x}, \mathbf{y})$ [Bishop, 2006]

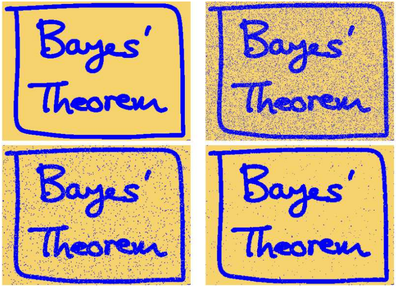


Иллюстрация шумоподавления [Bishop, 2006]

Скрытые марковские модели: Алгоритм Баума-Велча

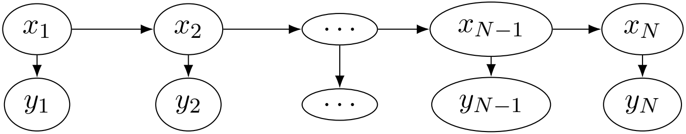
Александр Адуенко

2е апреля 2024

Содержание предыдущих лекций

- EM-алгоритм. Использование EM-алгоритма для отбора признаков в байесовской линейной регрессии.
- Вариационный EM-алгоритм и его использование для вывода в смеси моделей линейной регрессии.
- Ориентированные графические модели и их представление plate notation. Критерий условной независимости d-separation.
- Неориентированные графические модели и их связь с ориентированными.
- Факторные графы и алгоритм Sum-Product для вывода в ациклических графических моделях.
- Скрытые марковские модели и алгоритм Витерби. Алгоритм Max-Sum как обобщение алгоритма Витерби.

Скрытые марковские модели



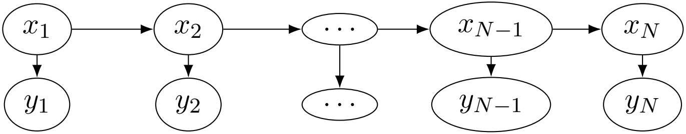
$$p(\mathbf{x},\,\mathbf{y}) = p(x_1)\prod_{i=2}^{N} p(x_i|x_{i-1})\prod_{i=1}^{N} p(y_i|x_i).$$

Пусть $x_i \in [K]$, $\mathbf{A} = \|a_{ij}\| = \|\mathsf{P}(x_l = j|x_{l-1} = i)\|$, $\pi_k = \mathsf{P}(x_1 = k)$.

$$p(\mathbf{x},\,\mathbf{y}|\mathbf{A},\,\boldsymbol{\pi},\,\mathbf{B}) = p(x_1|\boldsymbol{\pi})\prod_{i=2}^{N} p(x_i|x_{i-1},\,\mathbf{A})\prod_{i=1}^{N} p(y_i|x_i,\,\mathbf{B}).$$

Задачи:

- $p(x_i|\mathbf{y},\,\mathbf{A},\,\mathbf{B},\,\boldsymbol{\pi})$ – алгоритм Sum-Product;
- $p(x_i,\,x_{i+1}|\mathbf{y},\,\mathbf{A},\,\mathbf{B},\,\boldsymbol{\pi})$ – алгоритм Sum-Product;
- $p(\mathbf{x}|\mathbf{y},\,\mathbf{A},\,\mathbf{B},\,\boldsymbol{\pi}) \to \max\limits_{\mathbf{x}}$ – алгоритм Витерби / Max-Sum;
- $p(\mathbf{x}|\mathbf{y},\,\mathbf{A},\,\mathbf{B},\,\boldsymbol{\pi})$ – последовательное сэмплирование;
- $p(\mathbf{y}|\mathbf{A},\,\mathbf{B},\,\boldsymbol{\pi}) \to \max\limits_{\mathbf{A},\,\mathbf{B},\,\boldsymbol{\pi}}$.



Задача: Найти $p(\mathbf{x}|\mathbf{y},\ \mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi})$.

$$p(\mathbf{x},\ \mathbf{y}|\mathbf{A},\ \boldsymbol{\pi},\ \mathbf{B}) = p(x_1|\boldsymbol{\pi}) \prod_{i=2}^{N} p(x_i|x_{i-1},\ \mathbf{A}) \prod_{i=1}^{N} p(y_i|x_i,\ \mathbf{B}).$$

$$p(\mathbf{x}|\mathbf{y},\ \mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) \propto \underbrace{p(x_1|\boldsymbol{\pi})p(y_1|x_1,\ \mathbf{B})}_{\tilde{g}(x_1)} \underbrace{p(x_2|x_1,\ \mathbf{A})p(y_2|x_2,\ \mathbf{B})}_{\tilde{g}(x_2|x_1)} \cdot \ldots \cdot \underbrace{p(x_N|x_{N-1},\ \mathbf{A})p(y_N|x_N,\ \mathbf{B})}_{\tilde{g}(x_N|x_{N-1})}.$$

Идея:

- $x_1 \sim g(x_1)$;
- $x_2 \sim g(x_2|x_1)$;
- $\vdots$
- $x_N \sim g(x_N|x_{N-1})$.

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

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

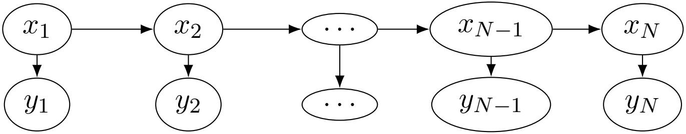
$$\log p(\mathbf{D}|\boldsymbol{\Theta}) - \int q(\mathbf{Z})\log\frac{q(\mathbf{Z})}{p(\mathbf{Z}|\mathbf{D},\boldsymbol{\Theta})}d\mathbf{Z} = \log p(\mathbf{D}|\boldsymbol{\Theta}) - D_{\mathrm{KL}}(q\|p(\mathbf{Z}|\mathbf{D},\ \boldsymbol{\Theta})).$$

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: $\boldsymbol{\Theta}^s = F(q^s,\ \boldsymbol{\Theta}) \to \max\limits_{\boldsymbol{\Theta}}$.

Вывод параметров скрытой марковской модели



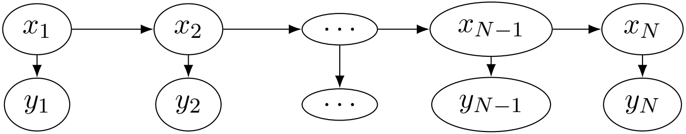
Задача: $p(\mathbf{y}|\mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) \to \max\limits_{\mathbf{A},\,\mathbf{B},\,\boldsymbol{\pi}}$.

$$p(\mathbf{x},\ \mathbf{y}|\mathbf{A},\ \boldsymbol{\pi},\ \mathbf{B}) = p(x_1|\boldsymbol{\pi}) \prod_{i=2}^{N} p(x_i|x_{i-1},\ \mathbf{A}) \prod_{i=1}^{N} p(y_i|x_i,\ \mathbf{B}).$$

Введем $\mathbf{Z} = \|z_{ik}\|,\ z_{ik} \in \{0,\ 1\}$ и пусть $y_i|x_i = k = \mathcal{N}(y_i|m_k,\ \sigma_k^2)$.

$$p(\mathbf{y},\ \mathbf{z}|\mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{1k}} \prod_{i=2}^{N} \prod_{k=1}^{K} \prod_{l=1}^{K} a_{kl}^{z_{i-1,\,k} z_{il}} \prod_{i=1}^{N} \prod_{k=1}^{K} \mathcal{N}(y_i|m_k,\ \sigma_k^2)^{z_{ik}}.$$

Вопрос: Как решить $p(\mathbf{y}|\mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) \to \max\limits_{\mathbf{A},\,\mathbf{B},\,\boldsymbol{\pi}}$, где $\mathbf{B} = (\mathbf{m},\ \boldsymbol{\sigma}^2)$?



Задача: $p(\mathbf{y}|\mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) \to \max\limits_{\mathbf{A},\, \mathbf{B},\, \boldsymbol{\pi}}$, где $\mathbf{B} = (\mathbf{m},\ \boldsymbol{\sigma}^2)$.

$$p(\mathbf{y},\ \mathbf{z}|\mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{1k}} \prod_{i=2}^{N} \prod_{k=1}^{K} \prod_{l=1}^{K} a_{kl}^{z_{i-1,\,k} z_{il}} \prod_{i=1}^{N} \prod_{k=1}^{K} \mathcal{N}(y_i|m_k,\ \sigma_k^2)^{z_{ik}}.$$

$$\log p(\mathbf{y},\ \mathbf{z}|\mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) = \sum_{k=1}^{K} z_{1k} \log \pi_k + \sum_{i=2}^{N} \sum_{k=1}^{K} \sum_{l=1}^{K} z_{i-1,\,k} z_{il} \log a_{kl} +$$

$$\sum_{i=1}^{N} \sum_{k=1}^{K} z_{ik} \left(-\frac{1}{2} \log \sigma_k^2 - \frac{1}{2\sigma_k^2} (y_i - m_k)^2 \right).$$

Е-шаг: $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{y},\ \mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi})$.

М-шаг: $\mathsf{E}_q \log p(\mathbf{y},\ \mathbf{z}|\mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) \to \max\limits_{\mathbf{A},\, \mathbf{B},\, \boldsymbol{\pi}}$.

М-шаг

$$\mathsf{E}_q \log p(\mathbf{y},\ \mathbf{z}|\mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) = \sum_{k=1}^{K} \mathsf{E} z_{1k} \log \pi_k + \sum_{i=2}^{N}\sum_{k=1}^{K}\sum_{l=1}^{K} \mathsf{E} z_{i-1,\,k} z_{il} \log a_{kl} +$$

$$\sum_{i=1}^{N}\sum_{k=1}^{K} \mathsf{E} z_{ik} \left(-\frac{1}{2}\log\sigma_k^2 - \frac{1}{2\sigma_k^2}(y_i - m_k)^2 \right).$$

$$\mathsf{E}_q \log p(\mathbf{y},\ \mathbf{z}|\mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) \to \max_{\mathbf{A},\,\mathbf{B},\,\boldsymbol{\pi}}.$$

$$\pi_k = \mathsf{E} z_{1k},\ a_{kl} \propto \sum_{i=2}^{N} \mathsf{E} z_{i-1,\,k} z_{il};$$

$$m_k = \frac{\sum_{i=1}^{N} \mathsf{E} z_{ik} y_i}{\sum_{i=1}^{N} \mathsf{E} z_{ik}},\ \sigma_k^2 = \frac{\sum_{i=1}^{N} \mathsf{E} z_{ik}(y_i - m_k)^2}{\sum_{i=1}^{N} \mathsf{E} z_{ik}}.$$

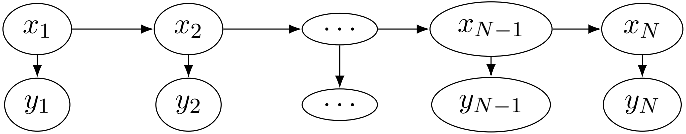
Вопрос: Что требуется знать про $q(\mathbf{Z})$, чтобы осуществить М-шаг?

Общий шаг: $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{y},\ \mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi})$.
Достаточно: $\mathsf{E}z_{ik}$, $\mathsf{E}z_{i-1,\,k}z_{il}$.
Вопрос: Можно ли воспользоваться алгоритмом Sum-Product для получения $\mathsf{E}z_{ik}$, $\mathsf{E}z_{i-1,\,k}z_{il}$?

Введем $\alpha_k(t) = p(y_1,\ \ldots,\ y_t,\ x_t = k|\mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi})$ и
$\beta_k(t) = p(y_{t+1},\ \ldots,\ y_N|x_t = k,\ \mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi})$.
$\mathsf{E}z_{tk} = \mathsf{P}(x_t = k|\mathbf{y},\ \mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) \propto \mathsf{P}(x_t = k,\ \mathbf{y}|\mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) =$
$\underbrace{p(y_1,\ \ldots,\ y_t,\ x_t = k|\mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi})}_{\alpha_k(t)}\underbrace{p(y_{t+1},\ \ldots,\ y_N|x_t = k,\ \mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi})}_{\beta_k(t)}$.
Вопрос 1: Какое свойство СММ было использовано при выводе выше?

$\mathsf{E}z_{t-1,\,k}z_{tl} = \mathsf{P}(x_{t-1} = k,\ x_t = l|\mathbf{y},\ \mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) \propto p(x_{t-1} = k,\ x_t = l,\ \mathbf{y}|\mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) = p(y_1,\ \ldots,\ y_{t-1},\ x_{t-1} = k|\mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi})p(x_t = l|x_{t-1} = k)p(y_t|x_t = l)p(y_{t+1},\ \ldots,\ y_N|x_t = l,\ \mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) \Longrightarrow$
$\mathsf{E}z_{t-1,\,k}z_{tl} \propto \alpha_k(t-1)a_{kl}p(y_t|x_t = l,\ \mathbf{B})\beta_l(t)$.
Вопрос 2: Какие свойства СММ были использованы при выводе выше?

Е-шаг 2: Получение $\alpha_k(t)$ и $\beta_k(t)$



Сосчитаем $\alpha_k(t) = p(y_1, \ldots, y_t, x_t = k|\mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ пошагово:

- $\alpha_k(1) = \pi_k p(y_1|x_1 = k, \mathbf{B})$;
- $\alpha_k(t+1) = \sum\limits_{j=1}^{K} \alpha_j(t) a_{jk} p(y_{t+1}|x_{t+1} = k, \mathbf{B})$.

Сосчитаем $\beta_k(t) = p(y_{t+1}, \ldots, y_N|x_t = k, \mathbf{A}, \mathbf{B}, \boldsymbol{\pi})$ пошагово:

- $\beta_k(N) = 1$;
- $\beta_k(t) = \sum\limits_{j=1}^{K} \beta_j(t+1) a_{kj} p(y_{t+1}|x_{t+1} = j, \mathbf{B})$.

Вопрос 1: Какие численные проблемы стоит ожидать при вычислениях по описанной схеме?

Вопрос 2: Как разрешить численные проблемы при угасании значений $\alpha_k(t)$, $\beta_k(N-t)$, $t \gg 1$?

EM-алгоритм: Итоговая схема

$p(\mathbf{y}|\mathbf{A},\, \mathbf{B},\, \boldsymbol{\pi}) \to \max\limits_{\mathbf{A},\, \mathbf{B},\, \boldsymbol{\pi}}$, где $\mathbf{B} = (\mathbf{m},\, \boldsymbol{\sigma}^2)$.

Е-шаг:

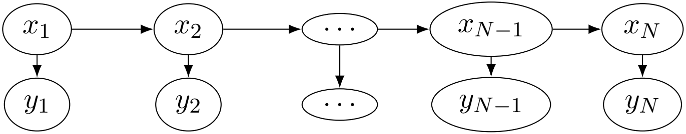
- Вычисляем $\alpha_k(t),\; \beta_k(t),\; t \in [N],\; k \in [K]$ при фиксированных $\mathbf{A},\; \mathbf{B},\; \boldsymbol{\pi}$;
- $\mathsf{E} z_{tk} \propto \alpha_k(t)\beta_k(t)$ и нормируем;
- $\mathsf{E} z_{t-1,\, k} z_{tl} \propto \alpha_k(t-1) a_{kl} p(y_t | x_t = l,\; \mathbf{B}) \beta_l(t)$ и нормируем.

М-шаг:

- $\pi_k = \mathsf{E} z_{1k},\; a_{kl} \propto \sum\limits_{i=2}^{N} \mathsf{E} z_{i-1,\, k} z_{il}$;
- $m_k = \dfrac{\sum_{i=1}^{N} \mathsf{E} z_{ik} y_i}{\sum_{i=1}^{N} \mathsf{E} z_{ik}},\; \sigma_k^2 = \dfrac{\sum_{i=1}^{N} \mathsf{E} z_{ik} (y_i - m_k)^2}{\sum_{i=1}^{N} \mathsf{E} z_{ik}}$.

Вопрос 1: Как учесть ненаблюдаемость части $\mathbf{y}$?

Вопрос 2: Как предсказать $y_{N+\Delta}$?



$$\alpha_k(t+1) = p(y_1,\, \ldots,\, y_{t+1},\, x_{t+1} = k|\mathbf{A},\, \mathbf{B},\, \boldsymbol{\pi}) =$$

$$\sum_{j=1}^{K} p(y_1,\, \ldots,\, y_{t+1},\, x_{t+1} = k,\, x_t = j|\mathbf{A},\, \mathbf{B},\, \boldsymbol{\pi}) =$$

$$\sum_{j=1}^{K} p(y_1,\, \ldots,\, y_t,\, x_t = j|\mathbf{A},\, \mathbf{B},\, \boldsymbol{\pi}) p(y_{t+1},\, x_{t+1} =$$
$$k|\mathbf{A},\, \mathbf{B},\, \boldsymbol{\pi},\, y_1,\, \ldots,\, y_t,\, \mathbf{x}_t = j) =$$

$$\sum_{j=1}^{K} p(y_1,\, \ldots,\, x_t = j|\mathbf{A},\, \mathbf{B},\, \boldsymbol{\pi}) p(y_{t+1},\, x_{t+1} = k|\mathbf{A},\, \mathbf{B},\, \boldsymbol{\pi},\, \mathbf{x}_t = j) =$$

$$\alpha_j(t) p(x_{t+1} = k|\mathbf{A},\, \mathbf{B},\, \boldsymbol{\pi},\, \mathbf{x}_t = j) p(y_{t+1}|x_{t+1} = k,\, \mathbf{A},\, \mathbf{B},\, \boldsymbol{\pi},\, \mathbf{x}_t = j) =$$

$$\sum_{j=1}^{K} \alpha_j(t) a_{jk} p(y_{t+1}|x_{t+1} = k,\, \mathbf{A},\, \mathbf{B},\, \boldsymbol{\pi}).$$

Алгоритмы поиска минимального разреза в графах для вывода в графических моделях

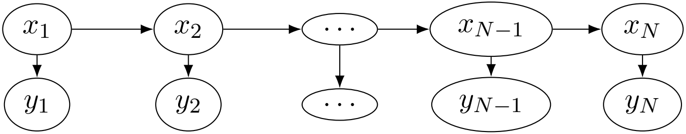
Александр Адуенко

21е апреля 2024

Содержание предыдущих лекций

- EM-алгоритм. Использование EM-алгоритма для отбора признаков в байесовской линейной регрессии.
- Вариационный EM-алгоритм и его использование для вывода в смеси моделей линейной регрессии.
- Ориентированные графические модели и их представление plate notation. Критерий условной независимости d-separation.
- Неориентированные графические модели и их связь с ориентированными.
- Факторные графы и алгоритм Sum-Product для вывода в ациклических графических моделях.
- Скрытые марковские модели (СММ) и алгоритм Витерби. Алгоритм Max-Sum как обобщение алгоритма Витерби.
- Алгоритм Баума-Велча для определения параметров СММ.

Скрытые марковские модели



$$p(\mathbf{x},\ \mathbf{y}) = p(x_1)\prod_{i=2}^{N} p(x_i|x_{i-1})\prod_{i=1}^{N} p(y_i|x_i).$$

Пусть $x_i \in [K]$, $\mathbf{A} = \|a_{ij}\| = \|\mathsf{P}(x_l = j|x_{l-1} = i)\|$, $\pi_k = \mathsf{P}(x_1 = k)$.

$$p(\mathbf{x},\ \mathbf{y}|\mathbf{A},\ \boldsymbol{\pi},\ \mathbf{B}) = p(x_1|\boldsymbol{\pi})\prod_{i=2}^{N} p(x_i|x_{i-1},\ \mathbf{A})\prod_{i=1}^{N} p(y_i|x_i,\ \mathbf{B}).$$

Задачи:

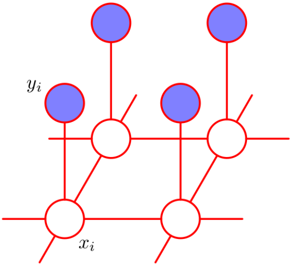
- $p(x_i|\mathbf{y},\ \mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi})$ – алгоритм Sum-Product;
- $p(x_i,\ x_{i+1}|\mathbf{y},\ \mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi})$ – алгоритм Sum-Product;
- $p(\mathbf{x}|\mathbf{y},\ \mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) \to \max\limits_{\mathbf{x}}$ – алгоритм Витерби / Max-Sum;
- $p(\mathbf{x}|\mathbf{y},\ \mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi})$ – последовательное сэмплирование;
- $p(\mathbf{y}|\mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) \to \max\limits_{\mathbf{A},\,\mathbf{B},\,\boldsymbol{\pi}}$ – алгоритм Баума-Велча.

Постановка задачи

Пусть имеется наблюдаемое изображение $\mathbf{y}$.

Требуется восстановить скрытые состояния $\mathbf{x}$ для каждого пикселя.

$p(\mathbf{x},\ \mathbf{y}) = \psi_0 \prod\limits_{(i,\,j)\in\varepsilon} \psi_{ij}(x_i,\ x_j) \prod\limits_{i} \psi_i(x_i,\ y_i).$

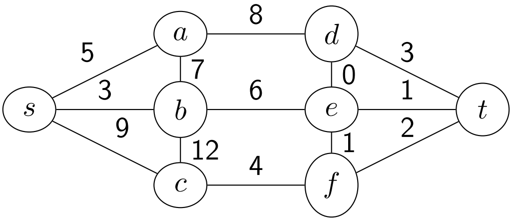


$$p(\mathbf{x}|\mathbf{y}) \to \max_{\mathbf{x}} \equiv p(\mathbf{x},\ y) \to \max_{\mathbf{x}} \equiv$$
$$E(\mathbf{x},\ \mathbf{y}) = -\log\psi_0 - \sum_{(i,\,j)\in\varepsilon} \log\psi_{ij}(x_i,\ x_j) +$$
$$\sum_i \log\psi_i(x_i,\ y_i) \to \min_{\mathbf{x}} \equiv \tilde{E}(\mathbf{x}) =$$
$$\theta_0 + \sum_{(i,\,j)\in\varepsilon} \theta_{ij}(x_i,\ x_j) + \sum_i \theta_i(x_i) \to \min_{\mathbf{x}}.$$

Графическая модель $p(\mathbf{x},\ \mathbf{y})$ [Bishop, 2006]

Замечание: Задача $\tilde{E}(\mathbf{x}) \to \min\limits_{\mathbf{x}}$ является NP-трудной для произвольных $\theta_{ij}(x_i,\ x_j)$, $x_i \in \{0,\ 1\}$.

Вопрос: При каких условиях на $\theta_i(x_i)$, $\theta_{ij}(x_i,\ x_j)$ задача разрешима полиномиально?



$c(u,\ v)$ – пропускная способность;

$f(u,\ v) \le c(u,\ v)$ – поток

$$\sum_{v:\ (u,\ v)\in\varepsilon} f(u,\ v) = \sum_{v:\ (v,\ u)\in\varepsilon} f(v,\ u)\ \forall\, u \notin \{s,\ t\}.$$

$$M(\mathbf{f}) = \sum_{v:\ (s,\ v)\in\varepsilon} f(s,\ v) \to \max_{\mathbf{f}}.$$

Разрез графа – разбиение мн-ва вершин $V = S \sqcup T$.

$C(S,\ T) = \sum\limits_{(u,\ v)\in\varepsilon:\ u\in S,\ v\in T} c(u,\ v)$ – величина разреза.

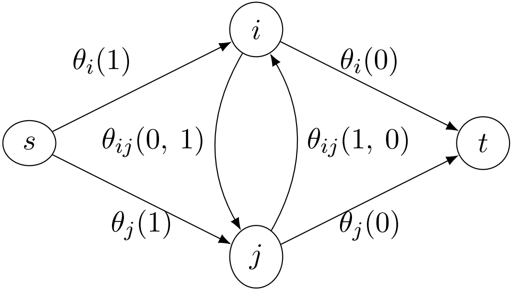
Теорема (Форд-Фалкерсон). Максимальный поток равен минимальному разрезу $\max\limits_{\mathbf{f}} M(\mathbf{f}) = \min\limits_{S,\ T} C(S,\ T)$.

Замечание: Максимальный поток / минимальный разрез эффективно вычислимы.

$$\tilde{E}(\mathbf{x}) = \theta_0 + \sum_{(i,\,j)\in\varepsilon} \theta_{ij}(x_i,\, x_j) + \sum_i \theta_i(x_i) \to \min_{\mathbf{x}}.$$

Пусть потенциалы удовлетворяют условиям:

- $\theta_i(0) \geq 0,\ \theta_i(1) \geq 0;$
- $\forall\,(i,\ j) \in \varepsilon:\ \theta_{ij}(0,\ 0) = \theta_{ij}(1,\ 1) = 0,\ \theta_{ij}(0,\ 1) \geq 0,\ \theta_{ij}(1,\ 0) \geq 0.$



Вопрос: Пусть $x_i = 0$, если $i \in S$ и $x_i = 1$, если $i \in T$.
Чему соответствует минимальный разрез такого графа?

$$\tilde{E}(\mathbf{x}) = \theta_0 + \sum_{(i,\, j)\in\varepsilon} \theta_{ij}(x_i,\, x_j) + \sum_i \theta_i(x_i) \to \min_{\mathbf{x}}.$$

Пусть потенциалы удовлетворяют условиям:

- $\theta_i(0) \geq 0,\ \theta_i(1) \geq 0$;
- $\forall\ (i,\ j) \in \varepsilon :\ \theta_{ij}(0,\ 0) = \theta_{ij}(1,\ 1) = 0,\ \theta_{ij}(0,\ 1) \geq 0,\ \theta_{ij}(1,\ 0) \geq 0.$

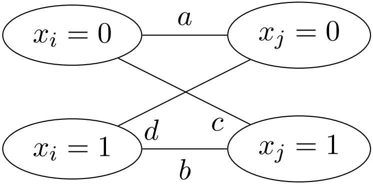
Операции, которые не меняют $\tilde{E}(\mathbf{x})$:

1. $\theta_i(0) \mathrel{-}= \delta,\ \theta_i(1) \mathrel{-}= \delta,\ \theta_0 \mathrel{+}= \delta$;
2. $\theta_{ij}(p,\ 0) \mathrel{-}= \delta,\ \theta_{ij}(p,\ 1) \mathrel{-}= \delta,\ \theta_i(p) \mathrel{+}= \delta.$

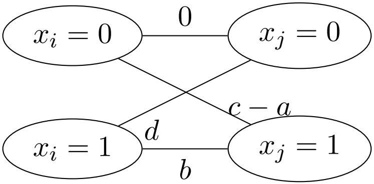
Замечание: $\theta_i(0) \geq 0,\ \theta_i(1) \geq 0$ можно обеспечить с помощью операции 1 с $\delta = \min(\theta_i(0),\ \theta_i(1))$.

Вопрос: Как добиться условия на парные потенциалы?

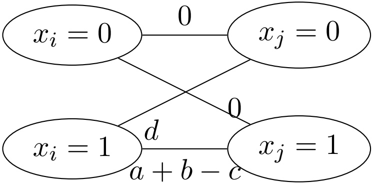
Репараметризация для парных потенциалов



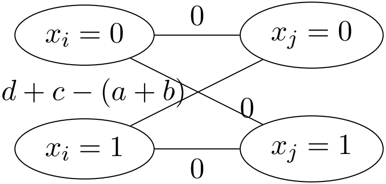
$\theta_{ij}(0,\ 0) = a,\ \theta_{ij}(1,\ 1) = b$
$\theta_{ij}(0,\ 1) = c,\ \theta_{ij}(1,\ 0) = d.$



$\theta_i(0)$ += a, $\theta_{ij}(0,\ 0)$ −= a,
$\theta_{ij}(0,\ 1)$ −= a.

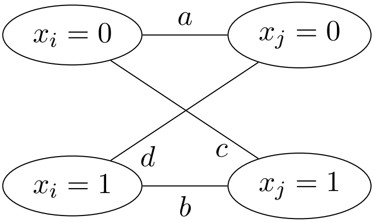


$\theta_j(1)$ += $c - a$, $\theta_{ij}(1,\ 1)$ −= $c - a$,
$\theta_{ij}(0,\ 1)$ −= $c - a$.

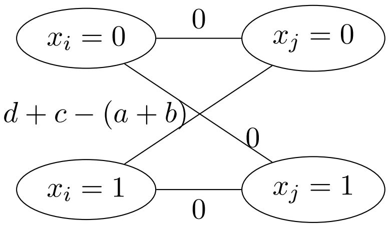


$\theta_i(1)$ += $a + b - c$, $\theta_{ij}(1,\ 0)$ −=
$a + b - c$,
$\theta_{ij}(1,\ 1)$ −= $a + b - c$.

Репараметризация для парных потенциалов 2



$\theta_{ij}(0,\ 0) = a,\ \theta_{ij}(1,\ 1) = b$
$\theta_{ij}(0,\ 1) = c,\ \theta_{ij}(1,\ 0) = d.$



$\theta_{ij}(0,\ 0) = 0,\ \theta_{ij}(1,\ 1) = 0$
$\theta_{ij}(0, 1) = 0, \theta_{ij}(1, 0) = d+c-(a+b).$

Условие субмодулярности (УС):
$\theta_{ij}(0,\ 1) + \theta_{ij}(1,\ 0) \geq \theta_{ij}(0,\ 0) + \theta_{ij}(1,\ 1).$

Утверждение: УС необходимо и достаточно для применимости алгоритмов разрезов графов для решения задачи $\tilde{E}(\mathbf{x}) \to \min_{\mathbf{x}}$.

Вопрос: Как выбрать потенциалы $\theta_i(x_i),\ \theta_{ij}(x_i,\ x_j)$?

Пусть $x_i = 1$ есть метка объекта, а $x_i = 0$ метка фона.
$\theta_i(0) = +\infty \equiv C(x_i, t) = +\infty$ гарантирует $x_i = 1$.
$\theta_i(1) = +\infty \equiv C(s, x_i) = +\infty$ гарантирует $x_i = 0$.

Выбор парного потенциала:

- Модель Поттса: $\theta_{ij}(x_i, x_j) = [x_i \neq x_j]$;
- $\theta_{ij}(x_i, x_j) = [x_i \neq x_j] \cdot \exp\left(-\frac{(y_i - y_j)^2}{2\sigma^2}\right)$.

Вопрос 1: Какую поправку выражает второй потенциал по отношению к модели Поттса?
Вопрос 2: Как учесть многоканальность (то есть $y_i \in \mathbb{R}^3_+$)?
Вопрос 3: Как учесть наличие / отсутствие линий / углов в двух пикселях?

Иллюстрация работы алгоритма GraphCut

Исходные изображения

Результат сшивки

Алгоритм α – расширение

$$\tilde{E}(\mathbf{x}) = \theta_0 + \sum_{(i,\,j)\in\varepsilon} \theta_{ij}(x_i,\, x_j) + \sum_i \theta_i(x_i) \to \min_{\mathbf{x}},\; x_i \in [K],\; K \geq 3.$$

Замечание: Задача NP-трудна даже для $K = 3$ и парных потенциалов Поттса.

Идея (α – расширение): Пошагово решать задачи с бинарными переменными.

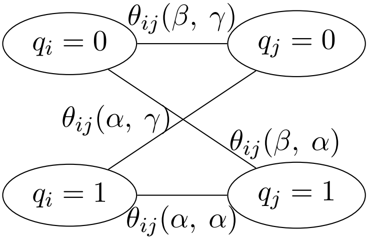
1. Выбираем начальное приближение $\mathbf{x}$, $x_i \in [K]$;
2. В цикле для каждой метки $\alpha \in [K]$ заменяем часть меток на данную, минимизируя энергию;
3. Останавливаемся, когда нет улучшений ни для одной метки.

Шаг 2 соответствует введению переменных $q_j \in \{0,\, 1\}$, что

- $q_j = 0$, если $x_j^{\text{old}} = x_j^{\text{new}}$;
- $q_j = 1$, если $x_j^{\text{old}} \neq \alpha$, $x_j^{\text{new}} = \alpha$.

Алгоритм α – расширение 2

Пусть $x_i^{\text{old}} = \beta,\ x_j^{\text{old}} = \gamma$.



Условие субмодулярности для бинарных потенциалов требует $\theta_{ij}(\beta,\ \alpha) + \theta_{ij}(\alpha,\ \gamma) \geq \theta_{ij}(\beta,\ \gamma) + \theta_{ij}(\alpha,\ \alpha)\ \forall\, \beta,\ \gamma$.

Вопрос 1: Какое условие получаем при $\theta_{ij}(\alpha,\ \alpha) = 0$?

Замечание: Для $x_i^{\text{old}} = \alpha$, $c(s,\ x_i) = +\infty$, $c(x_i,\ t) = \theta_i(\alpha)$.
Для $x_i^{\text{old}} \neq \alpha$, $c(s,\ x_i) = \theta_i(\alpha)$, $c(x_i,\ t) = \theta_i(x_i^{\text{old}})$.
Вопрос 2: Какое условие обеспечивает при $c(s,\ x_i) = +\infty$?

Исходные изображения

Результат сшивки

Иллюстрация работы алгоритма α – расширение 2

Исходные изображения

Результат сшивки

Алгоритм Tree-ReWeighted Message Passing для вывода в циклических графических моделях

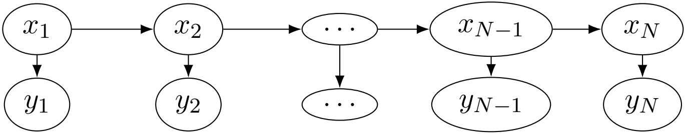
Александр Адуенко

23е апреля 2024

Содержание предыдущих лекций

- ЕМ-алгоритм. Использование ЕМ-алгоритма для отбора признаков в байесовской линейной регрессии.
- Вариационный ЕМ-алгоритм и его использование для вывода в смеси моделей линейной регрессии.
- Ориентированные графические модели и их представление plate notation. Критерий условной независимости d-separation.
- Неориентированные графические модели и их связь с ориентированными.
- Факторные графы и алгоритм Sum-Product для вывода в ациклических графических моделях.
- Скрытые марковские модели (СММ) и алгоритм Витерби. Алгоритм Max-Sum как обобщение алгоритма Витерби.
- Алгоритм Баума-Велча для определения параметров СММ.
- Алгоритмы на основе разрезов графов. Алгоритм α – расширение.

Вывод в графических моделях



$$p(\mathbf{x},\,\mathbf{y}) = p(x_1)\prod_{i=2}^{N} p(x_i|x_{i-1})\prod_{i=1}^{N} p(y_i|x_i).$$

Пусть $x_i \in [K]$, $\mathbf{A} = \|a_{ij}\| = \|\mathsf{P}(x_l = j|x_{l-1} = i)\|$, $\pi_k = \mathsf{P}(x_1 = k)$.

$$p(\mathbf{x},\,\mathbf{y}|\mathbf{A},\,\boldsymbol{\pi},\,\mathbf{B}) = p(x_1|\boldsymbol{\pi})\prod_{i=2}^{N} p(x_i|x_{i-1},\,\mathbf{A})\prod_{i=1}^{N} p(y_i|x_i,\,\mathbf{B}).$$

Задачи:

- $p(x_i|\mathbf{y},\,\Theta)$ – алгоритм Sum-Product;
- $p(\mathbf{x}_C|\mathbf{y},\,\Theta)$ – алгоритм Sum-Product;
- $p(\mathbf{x}|\mathbf{y},\,\Theta) \to \max_{\mathbf{x}}$ – алгоритм Витерби / Max-Sum / Graph-Cut / α – расширение;
- $p(\mathbf{x}|\mathbf{y},\,\Theta)$ – сэмплирование;
- $p(\mathbf{y}|\Theta) \to \max_{\Theta}$ – алгоритм Баума-Велча.

Алгоритм α – расширение

$$\tilde{E}(\mathbf{x}) = \theta_0 + \sum_{(i,\,j)\in\varepsilon} \theta_{ij}(x_i,\, x_j) + \sum_i \theta_i(x_i) \to \min_{\mathbf{x}},\; x_i \in [K],\; K \geq 3.$$

Замечание: Задача NP-трудна даже для $K = 3$ и парных потенциалов Поттса.

Идея (α – расширение): Пошагово решать задачи с бинарными переменными.

1. Выбираем начальное приближение $\mathbf{x}$, $x_i \in [K]$;
2. В цикле для каждой метки $\alpha \in [K]$ заменяем часть меток на данную, минимизируя энергию;
3. Останавливаемся, когда нет улучшений ни для одной метки.

Шаг 2 соответствует введению переменных $q_j \in \{0,\, 1\}$, что

- $q_j = 0$, если $x_j^{\text{old}} = x_j^{\text{new}}$;
- $q_j = 1$, если $x_j^{\text{old}} \neq \alpha$, $x_j^{\text{new}} = \alpha$.

Иллюстрация работы алгоритма α – расширение

Исходные изображения

Результат сшивки

Иллюстрация работы алгоритма α – расширение 2

Исходные изображения

Результат сшивки

Постановка задачи

$$\tilde{E}(\mathbf{x}) = \theta_0 + \sum_{(i,\,j)\in\varepsilon} \theta_{ij}(x_i,\, x_j) + \sum_i \theta_i(x_i) \to \min_{\mathbf{x}},\ x_i \in [K],\ K \geq 3.$$

Замечание: Задача NP-трудна в общем случае даже для $K = 3$ и парных потенциалов Поттса.

- Ациклическая ГМ $\Longrightarrow$ Точное решение через MaxSum;
- $K = 2$ и субмодулярные потенциалы $\Longrightarrow$ Точное решение через GraphCut;
- $K \geq 3$ и неравенство треугольника $\Longrightarrow$ Приближенное решение через α – расширение / $\alpha - \beta$ – замену.

Вопрос 1: Что делать, если $K \geq 3$ и неравенство треугольника не выполнено?

Вопрос 2: Как обработать потенциалы более высоких порядков, например, $\theta(x_1,\, x_2,\, x_3)$?

$$\tilde{E}(\mathbf{x}) = \theta_0 + \sum_{(i,\,j)\in\varepsilon} \theta_{ij}(x_i,\, x_j) + \sum_i \theta_i(x_i) \to \min_{\mathbf{x}},\ x_i \in [K],\ K \geq 3.$$

Вопрос: Как обработать $\theta(x_1,\, x_2,\, x_3)$?
Идея: Введем $X = x_1 + K(x_2 - 1) + K^2(x_3 - 1) \in [K^3]$.
Тогда $x_1 = \phi_1(X) = X\%K,\ x_2 = \phi_2(X) = 1 + (X\%K^2)/K$,
$x_3 = \phi_3(X) = 1 + X/K^2$,
$\theta(x_1,\, x_2,\, x_3) = \tilde{\theta}(X)I[x_1 = \phi_1(X)]I[x_2 = \phi_2(X)]I[x_3 = \phi_3(X)]$.

Алгоритм TRW: LP-релаксация

$$\tilde{E}(\mathbf{x}) = \theta_0 + \sum_{(i,\, j)\in\varepsilon} \theta_{ij}(x_i,\ x_j) + \sum_i \theta_i(x_i) \to \min_{\mathbf{x}},\ x_i \in [K],\ K \geq 3.$$

Введем $z_{ip} = 1 \iff x_i = p,\ z_{ij,\,pq} = 1 \iff x_i = p,\ x_j = q.$

Обозначим $\theta_{ip} = \theta_i(p),\ \theta_{ij}(p,\ q) = \theta_{ij,\,pq}.$

$$E(\mathbf{z},\ \Theta) = \sum_i \sum_{p=1}^K \theta_{ip} z_{ip} + \sum_{(i,\, j)\in\varepsilon} \sum_{p,\, q=1}^K \theta_{ij,\,pq} z_{ij,\,pq} \to \min_{\mathbf{z}}$$

s.t. $\sum_p z_{ip} = 1\, \forall\, i,\ \sum_q z_{ij,\,pq} = z_{ip},\ \sum_p z_{ij,\,pq} = z_{jq},\ z_{ip},\ z_{ij,\,pq} \in \{0,\ 1\}.$

LP-релаксация ($\mathbf{z} \in \mathcal{B} \to \mathbf{z} \in \mathcal{R}$):

$$E(\mathbf{z},\ \Theta) = \sum_i \sum_{p=1}^K \theta_{ip} z_{ip} + \sum_{(i,\, j)\in\varepsilon} \sum_{p,\, q=1}^K \theta_{ij,\,pq} z_{ij,\,pq} \to \min_{\mathbf{z}}$$

s.t. $\sum_p z_{ip} = 1\, \forall\, i,\ \sum_q z_{ij,\,pq} = z_{ip},\ \sum_p z_{ij,\,pq} = z_{jq},\ z_{ip},\ z_{ij,\,pq} \in [0,\ 1].$

$$\min_{\mathbf{z}\in\mathcal{B}} E(\mathbf{z},\ \Theta) \geq \min_{\mathbf{z}\in\mathcal{R}} E(\mathbf{z},\ \Theta).$$

Алгоритм TRW: Двойственная задача

$$E(\mathbf{z},\,\Theta) = \sum_{i\in V}\sum_{p=1}^{K}\theta_{ip}z_{ip} + \sum_{(i,\,j)\in\varepsilon}\sum_{p,\,q=1}^{K}\theta_{ij,\,pq}z_{ij,\,pq} \to \min_{\mathbf{z}}$$

s.t. $\sum_{p} z_{ip} = 1\;\forall\, i,\; \sum_{q} z_{ij,\,pq} = z_{ip},\; \sum_{p} z_{ij,\,pq} = z_{jq},\; z_{ip},\; z_{ij,\,pq} \in [0,\,1].$

Вопрос: В какой точке достигается минимум в задаче линейного программирования?

Идея: Покроем исходный граф $G = (V,\,\varepsilon)$ деревьями $G = \cup_{t=1}^{T} D_t$. Пусть $n_i \geq 1,\, n_{ij} \geq 1$ – количество деревьев, в которые входят вершина i и ребро $(i,\,j)$ соответственно.

Введем $\theta^t_{ip} = \dfrac{\theta_{ip}}{n_i} I[i \in D_t],\; \theta^t_{ij,\,pq} = \dfrac{\theta_{ij,\,pq}}{n_{ij}} I[(i,\,j) \in D_t].$

Тогда $E(\mathbf{z},\,\Theta) = \sum_{t=1}^{T} E^t(\mathbf{z},\,\Theta_t),\; \Theta = \sum_{t=1}^{T}\Theta_t.$

Задача: $E(\mathbf{z},\,\Theta) = \sum_{t=1}^{T} E^t(\mathbf{z},\,\Theta_t) \to \min_{\mathbf{z}\in\mathcal{R}}.$

Алгоритм TRW: Двойственная задача 2

$$E(\mathbf{z},\,\Theta) = \sum_{t=1}^{T} E^t(\mathbf{z},\,\Theta_t) = \sum_{i\in V}\sum_{p=1}^{K}\theta_{ip}z_{ip} + \sum_{(i,\,j)\in\varepsilon}\sum_{p,\,q=1}^{K}\theta_{ij,\,pq}z_{ij,\,pq} \to \min_{\mathbf{z}\in\mathcal{R}}.$$

Идея: Введем $\mathbf{z}^t$ и заменим исходную задачу на эквивалентную

$$\sum_{t=1}^{T} E^t(\mathbf{z}^t,\,\Theta_t) \to \min_{\mathbf{z}_1,\,\dots,\,z_T} \text{ s.t. } \mathbf{z}_t \in \mathcal{R}\,\forall\, t,\ \mathbf{z}_t = \mathbf{z}_1\,\forall\, t \geq 2.$$

$$L(\mathbf{Z},\,\Theta,\,\Lambda) = \sum_{t=1}^{T} E^t(\mathbf{z}^t,\,\Theta_t) + \sum_{i}\sum_{p=1}^{K}\sum_{t=2}^{T}\lambda_{ip}^t(z_{ip}^t - z_{ip}^1) +$$

$$\sum_{(i,\,j)\in\varepsilon}\sum_{p,\,q=1}^{K}\sum_{t=2}^{T}\lambda_{ij,\,pq}^t(z_{ij,\,pq}^t - z_{ij,\,pq}^1) =$$

$$\sum_{t=1}^{T}\left[\sum_{i\in V}\sum_{p=1}^{K}(\theta_{ip}^t + \lambda_{ip}^t)z_{ip}^t + \sum_{(i,\,j)\in\varepsilon}\sum_{p,\,q=1}^{K}(\theta_{ij,\,pq}^t + \lambda_{ij,\,pq}^t)z_{ij,\,pq}^t\right],$$

где $\Lambda \in \mathcal{L} = \left\{\Lambda :\ \lambda_{ip}^1 = -\sum_{t=2}^{T}\lambda_{ip}^t,\ \lambda_{ij,\,pq}^1 = -\sum_{t=2}^{T}\lambda_{ij,\,pq}^t\right\}.$

Алгоритм TRW: Двойственная задача 3

$$E(\mathbf{z},\ \Theta) = \sum_{t=1}^{T} E^t(\mathbf{z},\ \Theta_t) = \sum_{i \in V} \sum_{p=1}^{K} \theta_{ip} z_{ip} + \sum_{(i,\ j) \in \varepsilon} \sum_{p,\ q=1}^{K} \theta_{ij,\ pq} z_{ij,\ pq} \to \min_{\mathbf{z} \in \mathcal{R}}.$$

$$L(\mathbf{Z}, \Theta, \Lambda) = \sum_{t=1}^{T} \left[\sum_{i \in V} \sum_{p=1}^{K} (\theta^t_{ip} + \lambda^t_{ip}) z^t_{ip} + \sum_{(i,\ j) \in \varepsilon} \sum_{p,\ q=1}^{K} (\theta^t_{ij,\ pq} + \lambda^t_{ij,\ pq}) z^t_{ij,\ pq} \right]$$

$$L(\mathbf{Z},\ \Theta,\ \Lambda) = \sum_{t=1}^{T} E^t(\mathbf{z}^t,\ \Theta^t + \Lambda^t),\ \mathbf{z}^t \in \mathcal{R},\ \Lambda^t \in \mathcal{L}.$$

$$\min_{\mathbf{z} \in \mathcal{R}} E(\mathbf{z},\ \Theta) \geq \max_{\Lambda \in \mathcal{L}} \min_{\mathbf{z}_1,\ \dots,\ \mathbf{z}_T \in \mathcal{R}} L(\mathbf{Z},\ \Theta,\ \Lambda) = \max_{\Lambda \in \mathcal{L}} \sum_{t=1}^{T} \min_{\mathbf{z}^t \in \mathcal{R}} E^t(\mathbf{z}^t,\ \Theta^t + \Lambda^t).$$

$$\min_{\mathbf{z} \in \mathcal{R}} E(\mathbf{z},\ \Theta) \geq \max_{\Lambda \in \mathcal{L}} \sum_{t=1}^{T} \min_{\mathbf{z}^t \in \mathcal{R}} E^t(\mathbf{z}^t,\ \Theta^t + \Lambda^t) = \max_{\Lambda \in \mathcal{L}} \sum_{t=1}^{T} \min_{\mathbf{z}^t \in \mathcal{B}} E^t(\mathbf{z}^t,\ \Theta^t + \Lambda^t).$$

Вопрос 1: Что было использовано при получении последнего равенства?

Вопрос 2: Как решить $\min_{\mathbf{z}^t \in \mathcal{B}} E^t(\mathbf{z}^t,\ \Theta^t + \Lambda^t)$ для фиксированного Λ^t?

Алгоритм TRW: Двойственная задача 4

$$E(\mathbf{z},\,\Theta)=\sum_{t=1}^{T}E^t(\mathbf{z},\,\Theta_t)=\sum_{i\in V}\sum_{p=1}^{K}\theta_{ip}z_{ip}+\sum_{(i,\,j)\in\varepsilon}\sum_{p,\,q=1}^{K}\theta_{ij,\,pq}z_{ij,\,pq}\to\min_{\mathbf{z}\in\mathcal{R}}.$$

Эквиvалетная задача:

$$\tilde{E}(\mathbf{Z},\,\Theta)=\sum_{t=1}^{T}E^t(\mathbf{z}^t,\,\Theta_t)\to\min_{\mathbf{Z}\in Q},\ Q=\{\mathbf{Z}:\ \mathbf{z}_t\in\mathcal{R}\,\forall\,t,\ \mathbf{z}_t=\mathbf{z}_1\,\forall\,t\geq 2\}.$$

$$\min_{\mathbf{z}\in\mathcal{R}}E(\mathbf{z},\,\Theta)=\min_{\mathbf{Z}\in Q}\tilde{E}(\mathbf{Z},\,\Theta)=\max_{\Lambda\in\mathcal{L}}\sum_{t=1}^{T}\min_{\mathbf{z}^t\in\mathcal{B}}E^t(\mathbf{z}^t,\,\Theta^t+\Lambda^t).$$

Вопрос 1: Что было использовано для замены неравенства на равенство?

$g(\Lambda)=\sum_{t=1}^{T}\min_{\mathbf{z}^t\in\mathcal{B}}E^t(\mathbf{z}^t,\,\Theta^t+\Lambda^t)$ – вогнутая ли по Λ на выпуклом $\Lambda\in\mathcal{L}$?

Hint: $\min_{\mathbf{z}^t\in\mathcal{B}}E^t(\mathbf{z}^t,\,\Theta^t+\Lambda^t)$ – минимум конечного числа линейных функций по Λ^t, соответствующих разным значениям $\mathbf{z}^t$.

Вопрос 2: Сколько локальных максимумов имеет $g(\Lambda)$?

Вопрос 3: Дифференцируема ли $g(\Lambda)$?

Алгоритм TRW: Двойственная задача 5

$$E^t(\mathbf{z}^t,\ \Theta^t+\Lambda^t)=\sum_{i\in V}\sum_{p=1}^{K}(\theta^t_{ip}+\lambda^t_{ip})z^t_{ip}+\sum_{(i,\,j)\in\varepsilon}\sum_{p,\,q=1}^{K}(\theta^t_{ij,\,pq}+\lambda^t_{ij,\,pq})z^t_{ij,\,pq}.$$

$$g(\Lambda)=\sum_{t=1}^{T}\min_{\mathbf{z}^t\in\mathcal{B}}E^t(\mathbf{z}^t,\ \Theta^t+\Lambda^t)\to\max_{\Lambda\in\mathcal{L}}.$$

Идея: Использовать метод условного субградиентного подъема по Λ и обновлять $\mathbf{Z}(\Lambda)$ до сходимости.

$$\lambda^{t,\,n}_{ip}=\lambda^{t,\,n-1}_{ip}+\alpha_n\left(z^{t,\,n-1}_{ip}-\frac{\sum_{s:\,i\in D_s}z^{s,\,n-1}_{ip}}{n_i}\right);$$

$$\lambda^{t,\,n}_{ij,\,pq}=\lambda^{t,\,n-1}_{ij,\,pq}+\alpha_n\left(z^{t,\,n-1}_{ij,\,pq}-\frac{\sum_{s:\,(i,\,j)\in D_s}z^{s,\,n-1}_{ij,\,pq}}{n_{ij}}\right);$$

$$\mathbf{z}^{t,\,n}=\arg\min_{\mathbf{z}\in\mathcal{B}}E^t(\mathbf{z},\ \Theta^t+\Lambda^{t,\,n}).$$

Вопрос 1: Можно ли по $\mathbf{z}^{t,\,n}$ использовать градиентный шаг?

Вопрос 2: Как в формулах для Λ^n учтено $\Lambda\in\mathcal{L}$?

Алгоритм TRW: Свойства

$$\min_{\mathbf{z}\in\mathcal{B}} E(\mathbf{z},\ \theta) \geq \min_{\mathbf{z}\in\mathcal{R}} E(\mathbf{z},\ \theta) = \max_{\Lambda\in\mathcal{L}} \sum_{t=1}^{T} \min_{\mathbf{z}^t\in\mathcal{B}} E^t(\mathbf{z}^t,\ \Theta^t + \Lambda^t).$$

Вопрос 1: Зависит ли найденная нижняя оценка значения энергии $\max_{\Lambda\in\mathcal{L}} \sum_{t=1}^{T} \min_{\mathbf{z}^t\in\mathcal{B}} E^t(\mathbf{z}^t,\ \Theta^t + \Lambda^t)$ от покрытия графа деревьями?

Вопрос 2: Всегда ли можно покрыть граф $G = (V,\ \varepsilon)$ деревьями? Как выбрать покрытие деревьями?

Вопрос 3: Как получить приближенное решение $\mathbf{z}$ для задачи $\min_{\mathbf{z}\in\mathcal{B}} E(\mathbf{z},\ \theta)$ после нахождения $\mathbf{z}^t,\ \Lambda^t,\ t \in [T]$?

Идея: Рассмотреть ту часть $\mathbf{z}^t$, где оптимальные значения сходятся. Как согласовать остальные?

Литература

1. Wainwright, M. J., Jaakkola, T. S., & Willsky, A. S. (2005). MAP estimation via agreement on trees: message-passing and linear programming. IEEE transactions on information theory, 51(11), 3697-3717.
2. Kolmogorov, V. (2005, January). Convergent tree-reweighted message passing for energy minimization. In International Workshop on Artificial Intelligence and Statistics (pp. 182-189). PMLR.
3. Kolmogorov, V., & Wainwright, M. (2012). On the optimality of tree-reweighted max-product message-passing. arXiv preprint arXiv:1207.1395.
4. Kolmogorov, V. (2014). A new look at reweighted message passing. IEEE transactions on pattern analysis and machine intelligence, 37(5), 919-930.
5. Koller, Daphne, and Nir Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.

Оценивание гиперпараметров графических моделей

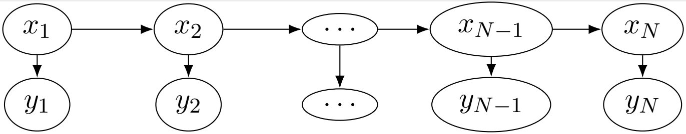
Александр Адуенко

6е мая 2024

Содержание предыдущих лекций

- EM-алгоритм и его сходимость. Использование EM-алгоритма для отбора признаков в байесовской линейной регрессии.
- Вариационный EM-алгоритм и его использование для вывода в смеси моделей линейной регрессии.
- Ориентированные графические модели и их представление plate notation. Критерий условной независимости d-separation.
- Неориентированные графические модели и их связь с ориентированными.
- Факторные графы и алгоритм Sum-Product для вывода в ациклических графических моделях.
- Скрытые марковские модели (СММ) и алгоритм Витерби. Алгоритм Max-Sum как обобщение алгоритма Витерби.
- Алгоритм Баума-Велча для определения параметров СММ.
- Алгоритмы на основе разрезов графов. Алгоритм α – расширение.
- Алгоритм TRW для приближенного вывода в циклических графических моделях с общей энергией.

Вывод в графических моделях



$$p(\mathbf{x},\ \mathbf{y}) = p(x_1)\prod_{i=2}^{N} p(x_i|x_{i-1})\prod_{i=1}^{N} p(y_i|x_i).$$

Пусть $x_i \in [K],\ \mathbf{A} = \|a_{ij}\| = \|\mathsf{P}(x_l = j|x_{l-1} = i)\|,\ \pi_k = \mathsf{P}(x_1 = k)$.

$$p(\mathbf{x},\ \mathbf{y}|\mathbf{A},\ \boldsymbol{\pi},\ \mathbf{B}) = p(x_1|\boldsymbol{\pi})\prod_{i=2}^{N} p(x_i|x_{i-1},\ \mathbf{A})\prod_{i=1}^{N} p(y_i|x_i,\ \mathbf{B}).$$

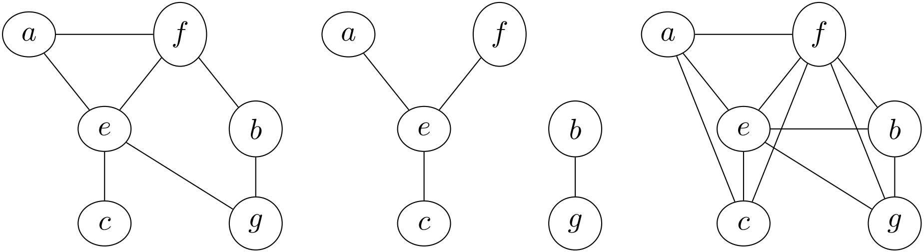
Задачи:

- $p(x_i|\mathbf{y},\ \Theta)$ – алгоритм Sum-Product;
- $p(\mathbf{x}_C|\mathbf{y},\ \Theta)$ – алгоритм Sum-Product;
- $p(\mathbf{x}|\mathbf{y},\ \Theta) \to \max\limits_{\mathbf{x}}$ – алгоритм Витерби / Max-Sum / Graph-Cut / α – расширение / TRW;
- $p(\mathbf{x}|\mathbf{y},\ \Theta)$ – сэмплирование;
- $p(\mathbf{y}|\Theta) \to \max\limits_{\Theta}$ – алгоритм Баума-Велча.

Обучение параметров графических моделей

$$p(\mathbf{y}|\Theta) \to \max_{\Theta}, \; p(\mathbf{x},\, \mathbf{y}|\Theta) = \frac{1}{Z}\prod_i \psi_i(\mathbf{x}_i,\, \mathbf{y}_i|\Theta_i).$$

Вопрос: Как обучить структуру графической модели?



$$p(a,\, b,\, c,\, e,\, f,\, g) = \frac{1}{Z_1}\psi_{afe}(a,\, f,\, e)\psi_{ec}(e,\, c)\psi_{eg}(e,\, g)\psi_{bg}(b,\, g)\psi_{bf}(b,\, f);$$

$$p(a,\, b,\, c,\, e,\, f,\, g) = \frac{1}{Z_2}\psi_{ae}(a,\, e)\psi_{fe}(f,\, e)\psi_{ce}(c,\, e)\psi_{bg}(b,\, g);$$

$$p(a,\, b,\, c,\, e,\, f,\, g) = \frac{1}{Z_3}\psi_{afec}(a,\, f,\, e,\, c)\psi_{efbg}(e,\, f,\, b,\, g);$$

Пример: Обучение структуры ГМ

Пусть $\mathbf{y} = [\mathbf{y}_1, \ldots, \mathbf{y}_K]^\mathsf{T}$, $\mathbf{y}_i \in \mathbb{R}^D$.

$$p(\mathbf{y}) = \frac{1}{Z}\exp(-E(\mathbf{y})),\; E(\mathbf{y}) = -\frac{1}{2}\mathbf{y}^\mathsf{T}\boldsymbol{\Omega}\mathbf{y} = -\frac{1}{2}\sum_{k,\, l=1}^{K} \mathbf{y}_k^\mathsf{T}\boldsymbol{\Omega}_{kl}\mathbf{y}_l.$$

$\boldsymbol{\Omega}_{kl} = \mathbf{O} \iff \mathbf{y}_k,\, \mathbf{y}_l$ – условно независимы при условии остальных переменных.

Идея: Ввести априорное распределение на $\boldsymbol{\Omega}$,
$p(\boldsymbol{\Omega}) \propto I[\boldsymbol{\Omega} > 0]\exp(-\lambda\|\boldsymbol{\Omega}\|_1)$.

$$\log p(\mathbf{y},\, \boldsymbol{\Omega}) \propto \log I[\boldsymbol{\Omega} > 0] - \lambda\|\boldsymbol{\Omega}\|_1 + \frac{m}{2}\log\det\boldsymbol{\Omega} - \frac{1}{2}\mathrm{tr}\left(\boldsymbol{\Omega}\sum_{j=1}^{m}\mathbf{y}^j{\mathbf{y}^j}^\mathsf{T}\right).$$

$$\log p(\boldsymbol{\Omega}|\mathbf{y},\, \lambda) \propto \log p(\mathbf{y},\, \boldsymbol{\Omega}) \to \max_{\boldsymbol{\Omega}}.$$

Вопрос 1: Как изменить $p(\boldsymbol{\Omega})$, чтобы убрать разреживание структуры внутри компонент одной переменной $\mathbf{y}_k$?

Вопрос 2: Как обобщить обучение структуры на случай с ненаблюдаемыми переменными?

$$p(\mathbf{y}|\Theta) \to \max_{\Theta}, \; p(\mathbf{x}, \, \mathbf{y}|\Theta) = \prod_{i=1}^{d} p(\mathbf{x}_i \,/\, \mathbf{y}_i|Pa_i, \, \Theta_i).$$

Пусть все переменные наблюдаемые, то есть $\mathbf{x} = \emptyset$.

Вопрос 1: Что изменилось по отношению к общему случаю?

$$\log p(\mathbf{y}|\Theta) = \sum_{i} \log p(\mathbf{y}_i|Pa_i, \, \Theta_i) \to \max_{\Theta}.$$

Вопрос 2: Что можно сказать про задачу, если Θ_i – непересекающиеся во всех факторах?

Вопрос 3: Пусть $\mathbf{y}_i \in [K]$, $Pa_i \in [L]$. Тогда $\Theta_i^{kl} = \mathsf{P}(\mathbf{y}_i = k|Pa_i = l)$. Что получим для Θ_i^{kl}?

Вопрос 4: Что делать, если $\mathbf{x} \neq \emptyset$?

Оценка гиперпараметров ориентированной ГМ 2

$$p(\mathbf{y}|\Theta) \to \max_{\Theta}, \; p(\mathbf{x}, \, \mathbf{y}|\Theta) = \prod_{i=1}^{d} p(\mathbf{x}_i \, / \, \mathbf{y}_i | Pa_i, \, \Theta_i).$$

$$p(\mathbf{y}|\Theta) = \int \prod_{i=1}^{d} p(\mathbf{x}_i \, / \, \mathbf{y}_i | Pa_i, \, \Theta_i) d\mathbf{x} \to \max_{\Theta} .$$

Идея: Используем EM-алгоритм для поиска гиперпараметров Θ.

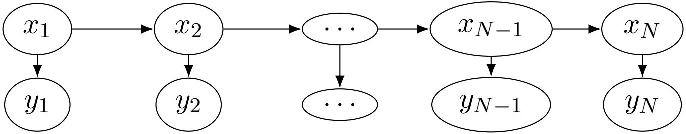
Введем $F(q, \, \Theta) = -\int q(\mathbf{x}) \log q(\mathbf{x}) d\mathbf{x} + \int q(\mathbf{x}) \log p(\mathbf{y}, \, \mathbf{x}|\Theta) d\mathbf{x} =$
$\log p(\mathbf{y}|\Theta) - D_{\mathrm{KL}}(q \| p(\mathbf{x}|\mathbf{y}, \, \Theta)) \to \max_{q, \, \Theta}.$

E-шаг. $q(\mathbf{x}) = \arg\min_{q \in Q} D_{KL}(q || p(\mathbf{x}|\mathbf{y}, \, \Theta)).$

M-шаг. $\sum_{j=1}^{m} \sum_{i=1}^{d} \mathsf{E}_{q(\mathbf{x})} \log p(\mathbf{x}_i^j \, / \, \mathbf{y}_i^j | Pa_i^j, \, \Theta_i) \to \max_{\Theta} .$

Вопрос: Что достаточно знать о $q(\mathbf{x})$ для проведения M-шага?

Пример: Оценка параметров СММ



Задача: $p(\mathbf{y}|\mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) \to \max\limits_{\mathbf{A},\,\mathbf{B},\,\boldsymbol{\pi}}$, где $\mathbf{B} = (\mathbf{m},\ \boldsymbol{\sigma}^2)$.

$$p(\mathbf{y},\ \mathbf{z}|\mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) = \prod_{k=1}^{K} \pi_k^{z_{1k}} \prod_{i=2}^{N}\prod_{k=1}^{K}\prod_{l=1}^{K} a_{kl}^{z_{i-1,\,k} z_{il}} \prod_{i=1}^{N}\prod_{k=1}^{K} \mathcal{N}(y_i|m_k,\ \sigma_k^2)^{z_{ik}}.$$

$$\log p(\mathbf{y},\ \mathbf{z}|\mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) = \sum_{k=1}^{K} z_{1k}\log\pi_k + \sum_{i=2}^{N}\sum_{k=1}^{K}\sum_{l=1}^{K} z_{i-1,\,k} z_{il}\log a_{kl} +$$

$$\sum_{i=1}^{N}\sum_{k=1}^{K} z_{ik}\left(-\frac{1}{2}\log\sigma_k^2 - \frac{1}{2\sigma_k^2}(y_i - m_k)^2\right).$$

Е-шаг: $q(\mathbf{Z}) = p(\mathbf{Z}|\mathbf{y},\ \mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi})$.

М-шаг: $\mathsf{E}_q \log p(\mathbf{y},\ \mathbf{z}|\mathbf{A},\ \mathbf{B},\ \boldsymbol{\pi}) \to \max\limits_{\mathbf{A},\,\mathbf{B},\,\boldsymbol{\pi}}$.

Пример: Оценка параметров СММ 2 (М-шаг)

$$\mathsf{E}_q \log p(\mathbf{y},\, \mathbf{z}|\mathbf{A},\, \mathbf{B},\, \boldsymbol{\pi}) = \sum_{k=1}^{K} \mathsf{E} z_{1k} \log \pi_k + \sum_{i=2}^{N}\sum_{k=1}^{K}\sum_{l=1}^{K} \mathsf{E} z_{i-1,\,k} z_{il} \log a_{kl} +$$

$$\sum_{i=1}^{N}\sum_{k=1}^{K} \mathsf{E} z_{ik} \left(-\frac{1}{2}\log\sigma_k^2 - \frac{1}{2\sigma_k^2}(y_i - m_k)^2 \right).$$

$$\mathsf{E}_q \log p(\mathbf{y},\, \mathbf{z}|\mathbf{A},\, \mathbf{B},\, \boldsymbol{\pi}) \to \max_{\mathbf{A},\, \mathbf{B},\, \boldsymbol{\pi}}.$$

$$\pi_k = \mathsf{E} z_{1k},\; a_{kl} \propto \sum_{i=2}^{N} \mathsf{E} z_{i-1,\,k} z_{il};$$

$$m_k = \frac{\sum_{i=1}^{N} \mathsf{E} z_{ik} y_i}{\sum_{i=1}^{N} \mathsf{E} z_{ik}},\; \sigma_k^2 = \frac{\sum_{i=1}^{N} \mathsf{E} z_{ik}(y_i - m_k)^2}{\sum_{i=1}^{N} \mathsf{E} z_{ik}}.$$

Вопрос: Что требуется знать про $q(\mathbf{Z})$, чтобы осуществить М-шаг?

Оценка параметров неориентированной ГМ

$$p(\mathbf{y}|\Theta) \to \max_{\Theta},\ p(\mathbf{x},\ \mathbf{y}|\Theta) = \frac{1}{Z}\prod_{i}\psi_i(\mathbf{x}_i,\ \mathbf{y}_i|\Theta_i).$$

Вопрос: Пусть все переменные наблюдаемые $\mathbf{x} = \emptyset$;
пусть дополнительно все параметры Θ_i в разных факторах независимы.

Верно ли $\Theta_i^* = \arg\max_{\Theta_i} \sum_{j=1}^{m} \log \psi_i(\mathbf{y}_i^j|\Theta_i)$?

Пусть все переменные наблюдаемые $\mathbf{x} = \emptyset$.

$$\log p(\mathbf{y}_1,\ \ldots,\ \mathbf{y}_m|\Theta) = \sum_{j=1}^{m}\sum_{i} \log \psi_i(\mathbf{y}_i^j|\Theta_i) - m\log Z(\Theta) \to \max_{\Theta}.$$

$$\nabla_{\Theta} \log p(\mathbf{y}_1,\ \ldots,\ \mathbf{y}_m|\Theta) = \sum_{j=1}^{m}\sum_{i} \nabla_{\Theta} \log \psi_i(\mathbf{y}_i^j|\Theta_i) - m\nabla_{\Theta} \log Z(\Theta).$$

Идея: Оценить $\nabla_{\Theta} \log Z(\Theta)$ и построить градиентный алгоритм максимизации $\log p(\mathbf{y}_1,\ \ldots,\ y_m|\Theta)$ по Θ, например:
$\Theta^{n+1} = \Theta^n + \lambda\nabla_{\Theta} \log p(\mathbf{y}_1,\ \ldots,\ \mathbf{y}_m|\Theta^n)$.

Оценка $Z(\Theta)$: Importance Sampling

$$p(\mathbf{y}|\Theta) = \frac{1}{Z(\Theta)}\tilde{p}(\mathbf{y}|\Theta),\; Z(\Theta) = \int \tilde{p}(\mathbf{y}|\Theta)d\mathbf{y}.$$

Пусть $p_0(\mathbf{y})$ – некоторое предложное распределение.

$$Z = \int \frac{p_0(\mathbf{y})}{p_0(\mathbf{y})}\tilde{p}(\mathbf{y})d\mathbf{y} = \int p_0(\mathbf{y})\frac{\tilde{p}(\mathbf{y})}{\frac{1}{Z_0}\tilde{p}_0(\mathbf{y})}d\mathbf{y} = Z_0 \int p_0(\mathbf{y})\frac{\tilde{p}(\mathbf{y})}{\tilde{p}_0(\mathbf{y})}d\mathbf{y}.$$

Выборочная оценка: $\hat{Z} = \frac{Z_0}{K}\sum_{k=1}^{K}\frac{\tilde{p}(\mathbf{y}_k)}{\tilde{p}_0(\mathbf{y}_k)},\; \mathbf{y}_k \sim p_0.$

Вопрос 1: Чем отличаются выборочные оценки $\hat{Z}$, построенные для разных $(p_0(\mathbf{y}),\; Z_0)$?

$$\mathrm{D}\hat{Z} = \frac{Z_0}{K^2}\sum_{k=1}^{K}\left(\frac{\tilde{p}(\mathbf{y}_k)}{\tilde{p}_0(\mathbf{y}_k)} - \hat{Z}\right)^2.$$

Вопрос 2: Как зависит дисперсия оценки $\mathrm{D}\hat{Z}$ от количества сэмплов K?

Замечание: Схема эффективна, если $p_0(\mathbf{y}) \approx p(\mathbf{y})$.

Оценка $Z(\Theta)$: Bridge Sampling

$$p(\mathbf{y}|\Theta) = \frac{1}{Z(\Theta)}\tilde{p}(\mathbf{y}|\Theta), \; Z(\Theta) = \int \tilde{p}(\mathbf{y}|\Theta) d\mathbf{y}.$$

Пусть $p_0(\mathbf{y})$ – некоторое предложное распределение, а $p_*(\mathbf{y})$ – интерполируещее распределение между p_0 и p.

$$\hat{Z}_* = \frac{Z_0}{K}\sum_{k=1}^{K}\frac{\tilde{p}_*(\mathbf{y}_k^0)}{\tilde{p}_0(\mathbf{y}_k^0)}, \; \mathbf{y}_k^0 \sim p_0; \; \hat{Z}_* = \frac{Z}{K}\sum_{k=1}^{K}\frac{\tilde{p}_*(\mathbf{y}_k)}{\tilde{p}(\mathbf{y}_k)}, \; \mathbf{y}_k \sim p.$$

$$\frac{Z}{Z_0} \approx \sum_{k=1}^{K}\frac{\tilde{p}_*(\mathbf{y}_k^0)}{\tilde{p}_0(\mathbf{y}_k^0)} / \sum_{k=1}^{K}\frac{\tilde{p}_*(\mathbf{y}_k)}{\tilde{p}(\mathbf{y}_k)}.$$

Вопрос 1: Пусть p_0 и p_* заданы. Что дополнительно требуется в Bridge Sampling против Importance Sampling с p_0?

Вопрос 2: Как выбрать p_*?

$$p_*^{\text{opt}} \propto \frac{\tilde{p}_0(\mathbf{y})\tilde{p}(\mathbf{y})}{\frac{Z}{Z_0}\tilde{p}_0(\mathbf{y}) + \tilde{p}(\mathbf{y})}$$ – зависит от Z!

Идея: Итеративно обновлять $\frac{Z}{Z_0}$ и p_*.

$$p(\mathbf{y}|\Theta) \to \max_{\Theta}, \; p(\mathbf{x}, \, \mathbf{y}|\Theta) = \frac{1}{Z} \prod_i \psi_i(\mathbf{x}_i, \, \mathbf{y}_i|\Theta_i).$$

Пусть теперь есть ненаблюдаемые переменные, то есть $\mathbf{x} \neq \emptyset$.

$$p(\mathbf{y}|\Theta) = \frac{1}{Z(\Theta)} \int \prod_{i=1}^{d} \psi_i(\mathbf{x}_i, \, \mathbf{y}_i|\Theta_i) d\mathbf{x} \to \max_{\Theta}.$$

Идея: Используем EM-алгоритм для поиска гиперпараметров Θ.

Введем $F(q, \, \Theta) = -\int q(\mathbf{x}) \log q(\mathbf{x}) d\mathbf{x} + \int q(\mathbf{x}) \log p(\mathbf{y}, \, \mathbf{x}|\Theta) d\mathbf{x} =$ $\log p(\mathbf{y}|\Theta) - D_{\mathrm{KL}}(q \| p(\mathbf{x}|\mathbf{y}, \, \Theta)) \to \max_{q, \, \Theta}$.

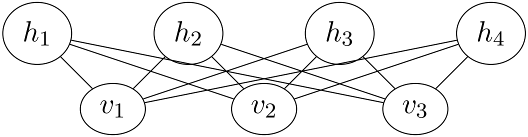
Е-шаг. $q(\mathbf{x}) = \arg\min_{q \in Q} D_{KL}(q||p(\mathbf{x}|\mathbf{y}, \, \Theta))$.

М-шаг. $-m \log Z(\Theta) + \sum_{j=1}^{m} \sum_{i=1}^{d} \mathsf{E}_{q(\mathbf{x})} \log \psi_i(\mathbf{x}_i^j, \, \mathbf{y}_i^j|\Theta_i) \to \max_{\Theta}$.

Идея: На Е-шаге, сэмплировать $\mathbf{x} \sim p(\mathbf{x}|\mathbf{y}, \, \Theta^n)$.

На М-шаге - градиентный шаг в направлении увеличения $F(q, \, \Theta)$.

Пример: Restricted Boltzmann Machine



$E(\mathbf{v},\ \mathbf{h}) = -\mathbf{b}^{\mathsf{T}}\mathbf{v} - \mathbf{c}^{\mathsf{T}}\mathbf{h} - \mathbf{v}^{\mathsf{T}}\mathbf{W}\mathbf{h},\ v_i,\ h_j \in \{0,\ 1\}.$

$$p(\mathbf{v}|\Theta) = p(\mathbf{v}|\mathbf{b},\ \mathbf{c},\ \mathbf{W}) = \frac{1}{Z(\mathbf{b},\ \mathbf{c},\ \mathbf{W})}\int \exp(-E(\mathbf{v},\ \mathbf{h}))d\mathbf{h} \to \max_{\mathbf{b},\,\mathbf{c},\,\mathbf{W}}.$$

Е-шаг: $\mathbf{h}_1,\ \dots,\ \mathbf{h}_K \sim p(\mathbf{h}|\mathbf{v},\ \mathbf{b}^n,\ \mathbf{c}^n,\ \mathbf{W}^n)$;

$$p(\mathbf{h}|\mathbf{v}) = \prod_j p(h_j|\mathbf{v}),\ P(h_j = 1|\mathbf{v}) = \sigma(c_j + \mathbf{v}^{\mathsf{T}}\mathbf{w}_j),\ \mathbf{W} = [\mathbf{w}_1,\ \dots,\ \mathbf{w}_H].$$

М-шаг: $g(\mathbf{b},\ c,\ \mathbf{W}) =$

$$-K\log Z(\mathbf{b},\ \mathbf{c},\ \mathbf{W}) + \mathbf{b}^{\mathsf{T}}\sum_{l=1}^{K}\mathbf{v}_l + \mathbf{c}^{\mathsf{T}}\sum_{l=1}^{K}\mathsf{E}\mathbf{h}_l + \mathbf{v}^{\mathsf{T}}\mathbf{W}\sum_{l=1}^{K}\mathsf{E}\mathbf{h}_l \to \max_{\mathbf{b},\,\mathbf{c},\,\mathbf{W}}.$$

$$\frac{\partial g(\mathbf{b},\ c,\ \mathbf{W})}{\partial w_{ij}} = -K\frac{\partial \log Z(\mathbf{b}^n,\ \mathbf{c}^n,\ \mathbf{W}^n)}{\partial w_{ij}} + v_i\mathsf{E}h_j.$$

Свойство: $\nabla_\Theta \log Z(\Theta) = \mathsf{E}_{(\mathbf{v},\,\mathbf{h})\sim p(\mathbf{v},\,\mathbf{h})}\nabla_\Theta \log\tilde{p}(\mathbf{v},\ \mathbf{h}|\Theta).$

Литература

1. Goodfellow, Ian, Yoshua Bengio, and Aaron Courville. Deep learning. MIT press, 2016: 598-621.
2. Ghahramani Z. Graphical models: parameter learning. URL: https://mlg.eng.cam.ac.uk/zoubin/papers/graphical-models02.pdf
3. Koller, Daphne, and Nir Friedman. Probabilistic graphical models: principles and techniques. MIT press, 2009.
4. Mestres, Adria Caballe, Natalia Bochkina, and Claus Mayer. "Selection of the regularization parameter in graphical models using network characteristics."Journal of Computational and Graphical Statistics 27.2 (2018): 323-333.
5. Gronau, Quentin F., et al. "A tutorial on bridge sampling."Journal of mathematical psychology 81 (2017): 80-97.
6. Gelman, Andrew, and Xiao-Li Meng. "Simulating normalizing constants: From importance sampling to bridge sampling to path sampling."Statistical science (1998): 163-185.

Библиотеки

- STAN: Main; PySTAN; Basic examples;
- Edward: Tutorials; Simple Bayesian NN example;
- Forneylab: Package; Paper; Code Style;
- PyMC: Main;
- CausalNex: Main;
- Factorie: Main;
- GTSAM: Main;
- HMMLearn: Main;
- PGMax: Main;
- BayesPy: Main.

Байесовский выбор моделей: Оценки скорости сходимости EM-алгоритма.

Константин Яковлев

30е апреля 2024

Пусть $\mathbf{D}$ – наблюдаемые переменные, $\mathbf{Z}$ – скрытые переменные.
$p(\mathbf{D}, \mathbf{Z}|\mathbf{\Theta}) = p(\mathbf{D}|\mathbf{Z}, \mathbf{\Theta})p(\mathbf{Z}|\mathbf{\Theta})$, $\mathbf{\Theta} \in \mathbf{\Omega}$ – выпуклое множество.

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

EM-алгоритм

$$\log p(\mathbf{D}|\mathbf{\Theta}) = \underbrace{\mathsf{E}_{q(\mathbf{Z})} \log \frac{p(\mathbf{D}, \mathbf{Z}|\mathbf{\Theta})}{q(\mathbf{Z})}}_{F(q,\mathbf{\Theta})} + \underbrace{D_{\mathrm{KL}}(q(\mathbf{Z})\|p(\mathbf{Z}|\mathbf{D}, \mathbf{\Theta}))}_{\geq 0}.$$

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

$$Q_n(\mathbf{\Theta}'|\mathbf{\Theta}) := \frac{1}{n}\sum_{i=1}^{n} \mathsf{E}_{p(z_i|D_i, \mathbf{\Theta})} \log p(D_i, z_i|\mathbf{\Theta}'), \quad Q(\mathbf{\Theta}'|\mathbf{\Theta}) := \mathsf{E}Q_n(\mathbf{\Theta}'|\mathbf{\Theta}),$$

$$M_n(\mathbf{\Theta}) := \arg\max_{\mathbf{\Theta}'\in\Omega} Q_n(\mathbf{\Theta}'|\mathbf{\Theta}), \quad M(\mathbf{\Theta}) := \arg\max_{\mathbf{\Theta}'\in\Omega} Q(\mathbf{\Theta}'|\mathbf{\Theta}).$$

Вопрос 2: чему соответствуют $M(.)$ и $M_n(.)$?

Пример: разделение смеси гауссиан

Пусть задано $p(\mathbf{x}|\boldsymbol{\Theta}) = \frac{1}{2}\mathcal{N}(\mathbf{x}|\boldsymbol{\Theta}^*,\, \sigma^2\mathbf{I}) + \frac{1}{2}\mathcal{N}(\mathbf{x}|-\boldsymbol{\Theta}^*,\, \sigma^2\mathbf{I}), \quad \mathbf{x} \in \mathbb{R}^d.$

Скрытая переменная $z \in \{0, 1\},\; p(z) = \mathrm{Ber}(0.5)$.

$p(\mathbf{x}|z = 0,\, \boldsymbol{\Theta}) = \mathcal{N}(\mathbf{x}|-\boldsymbol{\Theta},\, \sigma^2\mathbf{I}),\; p(\mathbf{x}|z = 1,\, \boldsymbol{\Theta}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\Theta},\, \sigma^2\mathbf{I})$

Выполняем E-шаг (считаем $\boldsymbol{\Theta}$ фиксированным)

$q(z_i = 1) = p(z_i = 1|\mathbf{x}_i,\, \boldsymbol{\Theta}),$

$p(z_i = 1|\mathbf{x}_i,\, \boldsymbol{\Theta}) = e^{-\frac{\|\mathbf{x}_i - \boldsymbol{\Theta}\|^2}{2\sigma^2}}[e^{-\frac{\|\mathbf{x}_i - \boldsymbol{\Theta}\|^2}{2\sigma^2}} + e^{-\frac{\|\mathbf{x}_i + \boldsymbol{\Theta}\|^2}{2\sigma^2}}]^{-1} := w_{\boldsymbol{\Theta}}(\mathbf{x}_i).$

Выполняем M-шаг (считаем $q(z_i)$ фиксированным)

$$Q_n(\boldsymbol{\Theta}'|\boldsymbol{\Theta}) \propto -\frac{1}{2\sigma^2}\sum_{i=1}^{n}[w_{\boldsymbol{\Theta}}(\mathbf{x}_i)\|\mathbf{x}_i - \boldsymbol{\Theta}'\|^2 + (1 - w_{\boldsymbol{\Theta}}(\mathbf{x}_i))\|\mathbf{x}_i + \boldsymbol{\Theta}'\|^2],$$

$$M_n(\boldsymbol{\Theta}) = \frac{2}{n}\sum_{i=1}^{n} w_{\boldsymbol{\Theta}}(\mathbf{x}_i)\mathbf{x}_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i, \quad M(\boldsymbol{\Theta}) = 2\mathsf{E}[w_{\boldsymbol{\Theta}}(\mathbf{x})\mathbf{x}].$$

Анализ сходимости EM-алгоритма I

Предложение 1 (self-consistency)

Пусть $\boldsymbol{\Theta}^* \in \arg\max_{\boldsymbol{\Theta}\in\boldsymbol{\Omega}} \mathsf{E}\log p(D_i|\boldsymbol{\Theta})$. Тогда $\boldsymbol{\Theta}^* \in \arg\max_{\boldsymbol{\Theta}\in\Omega} Q(\boldsymbol{\Theta}|\boldsymbol{\Theta}^*)$.

Предположение 1

Пусть $q(.) := Q(.|\boldsymbol{\Theta}^*)$ является λ-сильно вогнутой в шаре $\mathbb{B}_r(\boldsymbol{\Theta}^*)$:

$$q(\boldsymbol{\Theta}_1) - q(\boldsymbol{\Theta}_2) - \langle \nabla q(\boldsymbol{\Theta}_2), \boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2 \rangle \leq -\frac{\lambda}{2}\|\boldsymbol{\Theta}_1 - \boldsymbol{\Theta}_2\|^2$$

Вопрос 3: выполнено ли условие выше для смеси гауссиан?

Предположение 2 (First-order Stability (FOS))

Пусть для любого $\boldsymbol{\Theta} \in \mathbb{B}_r(\boldsymbol{\Theta}^*)$ и некоторого $\gamma \geq 0$

$$\|\nabla Q(M(\boldsymbol{\Theta})|\boldsymbol{\Theta}^*) - \nabla Q(M(\boldsymbol{\Theta})|\boldsymbol{\Theta})\|_2 \leq \gamma \|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_2$$

Замечание: для смеси гауссиан выполнено

$$\frac{2}{\sigma^2}\mathsf{E}[\|(w_{\boldsymbol{\Theta}}(\mathbf{x}) - w_{\boldsymbol{\Theta}^*}(\mathbf{x}))\mathbf{x}\|_2] \leq \gamma\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_2.$$

Анализ сходимости EM-алгоритма II

Теорема 1

Для некоторого $r > 0$ и чисел $0 \le \gamma < \lambda$ предположим, что $Q(.|\boldsymbol{\Theta}^*)$ является λ-сильно вогнутой в $\mathbb{B}_r(\boldsymbol{\Theta}^*)$, а также выполнено условие $\mathrm{FOS}(\gamma)$ в $\mathbb{B}_r(\boldsymbol{\Theta}^*)$. Тогда для любого $\boldsymbol{\Theta} \in \mathbb{B}_r(\boldsymbol{\Theta}^*)$

$$\|M(\boldsymbol{\Theta}) - \boldsymbol{\Theta}^*\|_2 \le \frac{\gamma}{\lambda}\|\boldsymbol{\Theta} - \boldsymbol{\Theta}^*\|_2.$$

Доказательство: запишем условие оптимальности первого порядка:

$$\langle \nabla Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}^*), M(\boldsymbol{\Theta}) - \boldsymbol{\Theta}^* \rangle \le 0, \quad \langle \nabla Q(M(\boldsymbol{\Theta})|\boldsymbol{\Theta}), \boldsymbol{\Theta}^* - M(\boldsymbol{\Theta}) \rangle \le 0.$$
$$\Rightarrow \langle \nabla Q(M(\boldsymbol{\Theta})|\boldsymbol{\Theta}^*) - \nabla Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}^*), \boldsymbol{\Theta}^* - M(\boldsymbol{\Theta}) \rangle \le$$
$$\langle \nabla Q(M(\boldsymbol{\Theta})|\boldsymbol{\Theta}^*) - \nabla Q(M(\boldsymbol{\Theta})|\boldsymbol{\Theta}), \boldsymbol{\Theta}^* - M(\boldsymbol{\Theta}) \rangle.$$

Воспользуемся λ-сильной вогнутостью:

$$\langle \nabla Q(M(\boldsymbol{\Theta})|\boldsymbol{\Theta}^*) - \nabla Q(\boldsymbol{\Theta}^*|\boldsymbol{\Theta}^*), \boldsymbol{\Theta}^* - M(\boldsymbol{\Theta}) \rangle \ge \lambda \|\boldsymbol{\Theta}^* - M(\boldsymbol{\Theta})\|_2^2$$

Воспользуемся $\mathrm{FOS}(\gamma)$, а также неравенством КБШ:

$$\lambda \|\boldsymbol{\Theta}^* - M(\boldsymbol{\Theta})\|_2^2 \le \gamma \|\boldsymbol{\Theta}^* - M(\boldsymbol{\Theta})\|_2 \cdot \|\boldsymbol{\Theta}^* - \boldsymbol{\Theta}\|_2$$

Замечание: рассуждения верны для $M(\boldsymbol{\Theta}) \leftarrow \mathrm{proj}_{\mathbb{B}_r(\boldsymbol{\Theta}^*)}(M(\boldsymbol{\Theta}))$

Предположение 3

Пусть для заданного $\delta \in (0, 1)$ и объема выборки n с вероятностью хотя бы $1 - \delta$ выполнено $\sup\limits_{\boldsymbol{\Theta} \in \mathbb{B}_r(\boldsymbol{\Theta}^*)} \|M_n(\boldsymbol{\Theta}) - M(\boldsymbol{\Theta})\|_2 \leq \varepsilon(n, \delta)$.

Теорема 2

Пусть оператор M является сжимающим в $\mathbb{B}_r(\boldsymbol{\Theta}^*)$ с параметром $\kappa \in (0, 1)$. Пусть $\boldsymbol{\Theta}^0 \in \mathbb{B}_r(\boldsymbol{\Theta}^*)$, а также $\varepsilon(n, \delta) \leq (1 - \kappa)\|\boldsymbol{\Theta}^* - \boldsymbol{\Theta}^0\|$. Тогда с вероятностью хотя бы $1 - \delta$:

$$\|\boldsymbol{\Theta}^t - \boldsymbol{\Theta}^*\|_2 \leq \kappa^t \|\boldsymbol{\Theta}^0 - \boldsymbol{\Theta}^*\|_2 + (1 - \kappa)^{-1} \varepsilon(n, \delta).$$

Доказательство: Докажем по индукции, что с вероятностью хотя бы $1 - \delta$ выполнено $\|\boldsymbol{\Theta}^{t+1} - \boldsymbol{\Theta}^*\|_2 \leq \kappa \|\boldsymbol{\Theta}^t - \boldsymbol{\Theta}^*\|_2 + \varepsilon(n, \delta) \leq r$. База очевидна. Докажем переход:

$$\|\boldsymbol{\Theta}^{t+1} - \boldsymbol{\Theta}^*\|_2 = \|M_n(\boldsymbol{\Theta}^t) - \boldsymbol{\Theta}^*\|_2 \leq \|M_n(\boldsymbol{\Theta}^t) - M(\boldsymbol{\Theta}^t)\|_2 + \|M(\boldsymbol{\Theta}^t) - \boldsymbol{\Theta}^*\|_2 \leq$$
$$\varepsilon(n, \delta) + \kappa \|\boldsymbol{\Theta}^t - \boldsymbol{\Theta}^*\|_2 \leq r(1 - \kappa) + \kappa r \leq r.$$

$$\Rightarrow \|\boldsymbol{\Theta}^t - \boldsymbol{\Theta}^*\|_2 \leq \kappa^t \|\boldsymbol{\Theta}^0 - \boldsymbol{\Theta}^*\|_2 + \left(\sum_{s=0}^{t-1} \kappa^s \right) \epsilon(n, \delta) \leq \kappa^t \|\boldsymbol{\Theta}^0 - \boldsymbol{\Theta}^*\|_2 + \frac{\varepsilon(n, \delta)}{1 - \kappa}.$$

Сходимость EM-алгоритма для модели разделения смеси гауссиан

Теорема 3

Пусть для достаточно большого η выполнено $\dfrac{\|\boldsymbol{\Theta}^*\|_2}{\sigma} > \eta$. Тогда найдется универсальная константа $c > 0$ такая, что оператор M является сжимающим в шаре $\mathbb{B}_r(\boldsymbol{\Theta}^*)$, где $\kappa(\eta) \le e^{-c\eta^2}$, $r = \dfrac{\|\boldsymbol{\Theta}^*\|_2}{4}$.

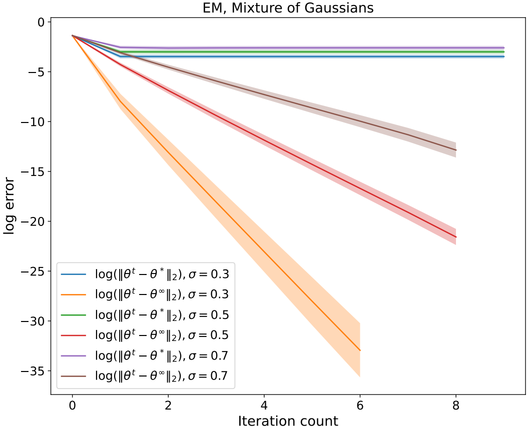
Теорема 4

Пусть выполнены условия теоремы 3. Пусть также $n \ge c_1 d \log(1/\delta)$. Тогда для любого $\boldsymbol{\Theta}^0 \in \mathbb{B}_{\|\boldsymbol{\Theta}^*\|/4}(\boldsymbol{\Theta}^*)$ с вероятностью хотя бы $1-\delta$ выполнено

$$\|\boldsymbol{\Theta}^t - \boldsymbol{\Theta}^*\|_2 \le \kappa^t \|\boldsymbol{\Theta}^0 - \boldsymbol{\Theta}^*\|_2 + \frac{c_2 \|\boldsymbol{\Theta}^*\|_2 \sqrt{\|\boldsymbol{\Theta}^*\|_2^2 + \sigma^2}}{1-\kappa} \sqrt{\frac{d}{n} \log(1/\delta)},$$

где c_1, c_2 – универсальные константы.

Сходимость EM-алгоритма на примере модели разделения смеси гауссиан



Параметры эксперимента

- $n = 10^4$
- $d = 10$
- $\|\mathbf{\Theta}^*\|_2 = 1$
- $\|\mathbf{\Theta}^0 - \mathbf{\Theta}^*\|_2 = \dfrac{\|\mathbf{\Theta}^*\|_2}{4}$

Замечание: для $\|\mathbf{\Theta}^t - \mathbf{\Theta}^\infty\|_2$ также можно показать линейную сходимость.

Литература

1. Bishop, Christopher M. "Pattern recognition and machine learning". Springer, New York (2006). Pp. 113-120, 161-171, 498-505.
2. Balakrishnan S., Wainwright M. J., Yu B. Statistical guarantees for the EM algorithm: From population to sample-based analysis. – 2017.