Машины Больцмана

Сергей Николенко

Академический Университет, весенний семестр 2011

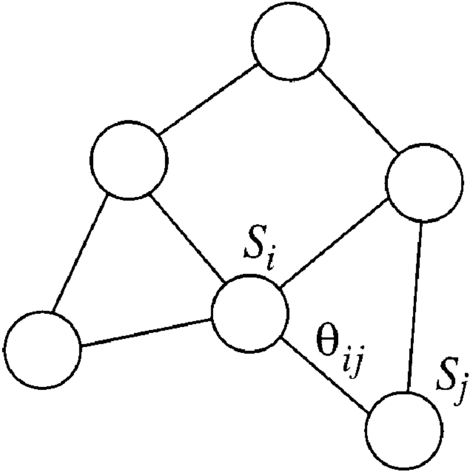
Outline

1. Машины Больцмана
   - Идея
   - Приближение
2. Блочная аппроксимация
   - Метод самосогласованного поля
   - Снова о машинах Больцмана

Машина Больцмана

- Машина Больцмана (Boltzmann machine) – это частный случай марковского случайного поля, т.е. ненаправленная графическая модель.
- Вершины – бинарные события S_i, набор функций-потенциалов ограничен.

Машина Больцмана



Машина Больцмана

- Фактор Больцмана (Boltzmann factor) – экспонента от квадратичного выражения от S_i.
- Потенциал каждой клики – произведение факторов, но мы предполагаем, что $e^{\theta_{ij} S_i S_j}$ встречается только в одной из клик.
- Поэтому совместное распределение выглядит как

$$p(S) = \frac{1}{Z} e^{\sum_{i<j} \theta_{ij} S_i S_j + \sum_i \theta_{i0} S_i},$$

где $\theta_{ij} = 0$ для несоседних S_i и S_j.

Машина Больцмана

- $E = -\sum_{i<j} \theta_{ij} S_i S_j - \sum_i \theta_{i0} S_i$ называется *энергией*.
- Вообще, совместное распределение $p(E) \sim e^{-\beta E}$ – это *распределение Больцмана* из статистической физики.

Машина Больцмана

- А смысл такой: если мы обучим веса θ_{ij} (это отдельный вопрос) и проведём вывод на каком-нибудь начальном распределении (evidence), то мы получим вероятности других, неизвестных вершин.
- Таким образом, можно дополнять частично известные распределения «по ассоциации».
- Вывод можно вести точно в некоторых частных случаях, но в общем случае он слишком сложен.

Стоящие перед нами задачи

- Мы хотим провести маргинализацию в распределении

$$p(S) = \frac{1}{Z} e^{\sum_{i<j} \theta_{ij} S_i S_j + \sum_i \theta_{i0} S_i}.$$

- Для маргинализации вида $p(H) = \sum_{\{S \setminus H\}} p(S)$ нам надо подсчитать сумму экспонент квадратичных функций.
- Для условных вероятностей вида $p(H \mid E) = \frac{p(H,E)}{p(E)}$ нужно подсчитать отношение таких сумм.
- Самая общая такая сумма – это, собственно, $Z = \sum_{\{S\}} p(S)$ (partition function); её и будем искать.

Стоящие перед нами задачи

- Метод такой: будем проводить вариационные преобразования одно за другим, оставаясь в рамках машины Больцмана.
- Один шаг:

$$Z = \sum_{\{S\}} e^{\sum_{j<k} \theta_{jk} S_j S_k + \sum_j \theta_{j0} S_j} =$$

$$= \sum_{\{S \setminus S_i\}} \sum_{S_i \in \{0,1\}} e^{\sum_{j<k} \theta_{jk} S_j S_k + \sum_j \theta_{j0} S_j}.$$

Нижняя оценка

- Легко показать, что внутренняя сумма лог-выпукла. Можно найти вариационную нижнюю оценку:

$$\ln\left[\sum_{S_i\in\{0,1\}} e^{\sum_{j<k}\theta_{ij}S_jS_k+\sum_j\theta_{j0}S_j}\right] =$$

$$= \sum_{\{j<k\}\neq i}\theta_{jk}S_jS_k + \sum_{j\neq i}\theta_{j0}S_j + \ln\left[1+e^{\sum_{j\neq i}\theta_{ij}S_j+\theta_{i0}}\right] \geq$$

$$\geq \sum_{\{j<k\}\neq i}\theta_{jk}S_jS_k+\sum_{j\neq i}\theta_{j0}S_j+\lambda_i^L\left(\sum_{j\neq i}\theta_{ij}S_j+\theta_{i0}\right)+H(\lambda_i^L),$$

т.к. $\ln(1+e^{-x}) \geq -\lambda x + H(\lambda)$.

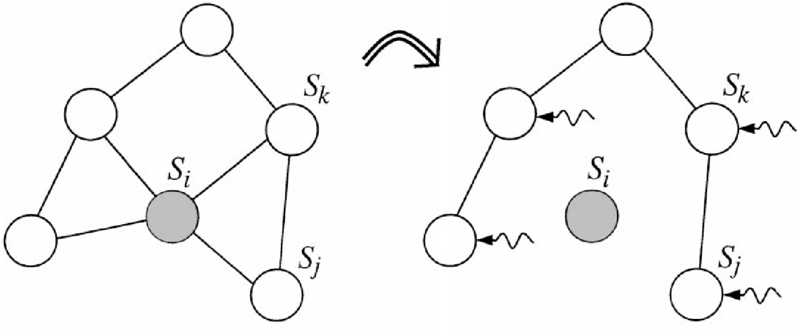
Нижняя оценка

- В графическом смысле мы отрезали S_i от соседей и добавили к соседям линейные члены:

$$\theta'_{jk} = \theta_{jk},$$
$$\theta'_{j0} = \theta_{j0} + \lambda_i^L \theta_{ij}.$$

- Но не соединили этих соседей, как получилось бы при точном выводе.
- Кроме того, добавился константный член $\lambda_i^L \theta_{i0} + H(\lambda_i^L)$.

Нижняя оценка



Верхняя оценка

- Можно аналогично найти и верхнюю оценку:

$$\ln(1+e^{-x}) = \ln(e^{\frac{x}{2}} + e^{-\frac{x}{2}}) + \frac{x}{2} \leq \lambda x^2 + \frac{x}{2} - g^*(\lambda).$$

- Соответственно,

$$\ln\left[\sum_{S_i\in\{0,1\}} e^{\sum_{j<k}\theta_{ij}S_jS_k + \sum_j \theta_{j0}S_j}\right] \leq \sum_{\{j<k\}\neq i}\theta_{jk}S_jS_k + \sum_{j\neq i}\theta_{j0}S_j +$$

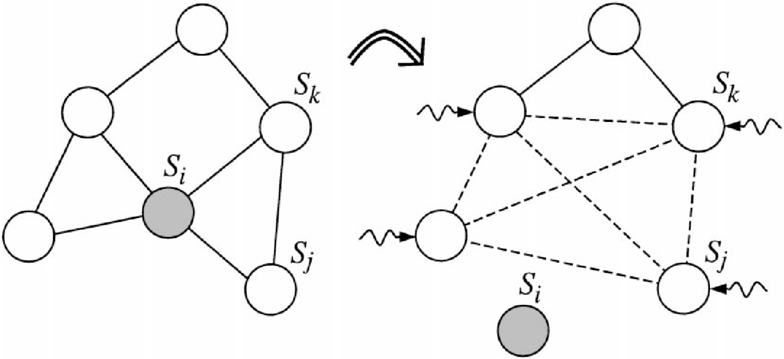
$$+\lambda_i^U\left(\sum_{j\neq i}\theta_{ij}S_j+\theta_{i0}\right)^2 + \frac{1}{2}\left(\sum_{j\neq i}\theta_{ij}S_j+\theta_{i0}\right) - g^*(\lambda_i^U).$$

Верхняя оценка

- Графический смысл теперь немножко другой, т.к. добавились новые связи между соседями S_i:

$$\theta'_{jk} = \theta_{jk} + 2\lambda_i^U,$$
$$\theta'_{j0} = \theta_{j0} + \lambda_i^L \theta_{ij}.$$

Верхняя оценка



Итого

- В итоге мы получили верхнюю и нижнюю оценки.
- Нижняя оценка удобнее: вершины отщепляются целиком, можно в любом удобном порядке аппроксимировать, довести до структуры (например, дерева), на которой уже легко вести точный вывод.
- Зато верхняя оценка точнее (практические результаты это показывают).

Outline

Блочная аппроксимация: идея

- Мы до сих пор удаляли вершины по одной.
- Но, может быть, если сразу несколько вершин выбрать, можно найти более точную аппроксимацию?
- Идея:
  1. выбрать подструктуру в графе, для которой можно провести точный вывод (дерево, набор цепочек и т.п.);
  2. рассмотреть семейство вероятностных распределений на этой подструктуре с вариационными параметрами;
  3. выбрать одно распределение из этого семейства, желательно оптимально аппроксимирующее.

Блочная аппроксимация: идея

- Формально: есть $p(S)$, мы хотим оценить $p(H \mid E)$.
- Введём семейство приближений $q(H \mid E, \lambda)$, где λ – вариационные параметры.
- Выберем из них одно, минимизирующее расстояние Кульбака–Ляйблера:

$$\lambda^* = \arg\min_{\lambda} \mathrm{KL}(q(H \mid E, \lambda) \| p(H \mid E)), \text{ где}$$

$$\mathrm{KL}(q \| p) = \sum_{\{S\}} q(S) \ln \frac{q(S)}{p(S)}.$$

Расстояние Кульбака–Ляйблера

- Почему именно KL? Помимо прочего, это естественная нижняя оценка на правдоподобие $p(E)$.
- По неравенству Йенсена:

$$\ln p(E) = \ln \sum_{\{H\}} q(H \mid E) \frac{p(H \mid E)}{q(H \mid E)} \geq$$

$$\geq \sum_{\{H\}} q(H \mid E) \ln \frac{p(H \mid E)}{q(H \mid E)},$$

  и разница между левой и правой частями – это как раз $\mathrm{KL}(q \| p)$.
- Поэтому справа стоит нижняя оценка на $p(E)$, и для оптимального λ^* это оптимальная оценка.

Расстояние Кульбака–Ляйблера

- В итоге получается:

$$\ln p(E) \geq \sum_{\{H\}} q(H \mid E) \ln p(H \mid E) - \sum_{\{H\}} q(H \mid E) \ln q(H \mid E).$$

Упражнение. Это можно было бы и вариационным методом получить. Попробуйте получить эту оценку вариационным методом, используя вектор вероятностей $q(H \mid E, \lambda)$ как вектор вариационных параметров.

Обучение параметров

- Эту оценку можно использовать в рамках EM-алгоритма для обучения параметров модели.
- Добавим теперь в наши обозначения параметры θ: теперь $p(S \mid \theta)$.
- Введём функцию

$$\mathcal{L}(q, \theta) =$$
$$= \sum_{\{H\}} q(H \mid E) \ln p(H \mid E, \theta) - \sum_{\{H\}} q(H \mid E) \ln q(H \mid E) \leq \ln p(E).$$

Обучение параметров

- Если мы разрешили бы $q(H \mid E)$ быть любым, оптимальное значение было бы $q(H \mid E) = p(H \mid E, \theta)$.
- Давайте применим такую форму EM-алгоритма:

Е-шаг $Q^{(k+1)} := \arg\max_Q \mathcal{L}\left(Q, \theta^{(k)}\right)$;
М-шаг $\theta^{(k+1)} := \arg\max_\theta \mathcal{L}\left(Q^{(k+1)}, \theta\right)$.

- Это просто покоординатный подъём для функции правдоподобия $\mathcal{L}(q, \theta)$.

Обучение параметров

- Если теперь $q(H \mid E)$ – это всё-таки аппроксимация, получается уже известный нам приём minorization–maximization.
- Мы на каждом шаге оптимизируем нижнюю оценку правдоподобия вместо него самого.
- Но в принципе алгоритм точно таким же остаётся.

Машины Больцмана

- Вернёмся к нашим машинам Больцмана:

$$p(S \mid \theta) = \frac{1}{Z} e^{\sum_{i<j} \theta_{ij} S_i S_j + \sum_i \theta_{i0} S_i}.$$

- Что будет в $p(H \mid E, \theta)$?
  - для $S_i \in E$ и $S_j \in E$ $\theta_{ij} S_i S_j$ – это константа, и она исчезает при нормализации;
  - для $S_i \in H$, $S_j \in E$ квадратичный член становится линейным и вписывается в S_i;
  - линейные члены для $S_i \in E$ исчезают.

Машины Больцмана

- Итого:
$$p(H \mid E, \theta) = \frac{1}{Z_c} e^{\sum_{i<j} \theta_{ij} S_i S_j + \sum_i \theta_{i0}^c S_i},$$
где суммы только по узлам из H, и $\theta_{i0}^c = \theta_{i0} + \sum_{j \in E} \theta_{ij} S_j$.
- Теперь
$$Z_c = \sum_{\{H\}} e^{\sum_{i<j} \theta_{ij} S_i S_j + \sum_i \theta_{i0}^c S_i},$$
и мы получили машину Больцмана на подмножестве H.

Метод самосогласованного поля

- Ещё один термин из физики – метод самосогласованного поля (mean field theory).
- Смысл в том, чтобы искать приближение среди *полностью* факторизуемых распределений.
- Иначе говоря, мы ищем $q(H \mid E, \mu)$ в виде

$$q(H \mid E, \mu) = \prod_{i \in H} \mu_i^{S_i} (1 - \mu_i)^{1 - S_i}.$$

Метод самосогласованного поля

- Теперь можно посчитать KL-расстояние:

$$\mathrm{KL}(q\|p) = \sum_{\{H\}} q(H \mid E, \mu) \ln \frac{q(H \mid E, \mu)}{p(H \mid E, \theta)} =$$

$$= \sum_i \left[\mu_i \ln \mu_i + (1-\mu_i)\ln(1-\mu_i)\right] - \sum_{i<j} \theta_{ij}\mu_i\mu_j - \sum_i \theta^c_{i0}\mu_i + \ln Z_c$$

(т.к. в распределении q S_i и S_j – независимые случайные величины со средними μ_i и μ_j).

- И мы хотим минимизировать это по q, т.е. по μ_i.

Метод самосогласованного поля

- Возьмём частные производные по μ_i и приравняем нулю; получим
$$\mu_i = \sigma\left(\sum_j \theta_{ij}\mu_j + \theta_{i0}\right),$$
где $\sigma(z) = 1/(1 + e^{-z})$ – сигмоид.
- Эти уравнения называются «уравнениями самосогласованного поля»; их можно решить итеративно.

Обучение машин Больцмана

- Как мы уже говорили, можно это применить и для обучения параметров θ_{ij}.
- Выпишем нижнюю оценку для метода самосогласованного поля:

$$\ln p(E \mid \theta) \geq$$
$$\geq \sum_{\{H\}} q(H \mid E) \ln p(H \mid E) - \sum_{\{H\}} q(H \mid E) \ln q(H \mid E) =$$
$$= \sum_{i<j} \theta_{ij}\mu_i\mu_j + \sum_i \theta^c_{i0}\mu_i - \ln Z - \sum_i \left[\mu_i \ln \mu_i + (1-\mu_i)\ln(1-\mu_i)\right].$$

Обучение машин Больцмана

- Теперь можно взять производные по θ_{ij} и запустить градиентный подъём. Но для этого нужно знать $\frac{\partial \ln Z}{\partial \theta_{ij}}$.

Упражнение. Покажите, что $\frac{\partial \ln Z}{\partial \theta_{ij}} = \langle S_i S_j \rangle$, где $\langle \cdot \rangle$ – ожидание по распределению $p(S \mid \theta)$.

- В итоге получается правило обучения:

$$\Delta \theta_{ij} \propto \mu_i \mu_j - \langle S_i S_j \rangle.$$

Обучение машин Больцмана

- К сожалению, $\langle S_i S_j \rangle$ точно подсчитать не получится (если вообще точный вывод нельзя провести).
- Есть и более серьёзная проблема: мы же хотим несколько ассоциаций заложить в модель, т.е. получить мультимодальное распределение.
- А приближаем мы его унимодальным распределением (после факторизации).
- Один возможный подход – рассматривать мультимодальные распределения q, например, смеси распределений рассмотреть. Этого мы делать уже не будем.

Thank you!

Спасибо за внимание!