# ЛЕКЦИЯ 3. Глубокие машины Больцмана

## *Глубокие машины Больцмана*

Глубокая машина Больцмана (ГМБ) – еще один вид порождающих моделей. В отличие от глубокой сети доверия, она полностью неориентированная, а в отличие от ОМБ (ограниченной машины Больцмана), имеет несколько слоев латентных переменных (в ОМБ такой слой единственный). Но так же, как и в ОМБ, внутри слоя все переменные взаимно независимы при условии переменных из соседних слоев. Структура графа показана на рис. 1. Глубокие машины Больцмана применялись к различным задачам, в т. ч. к моделированию документов.

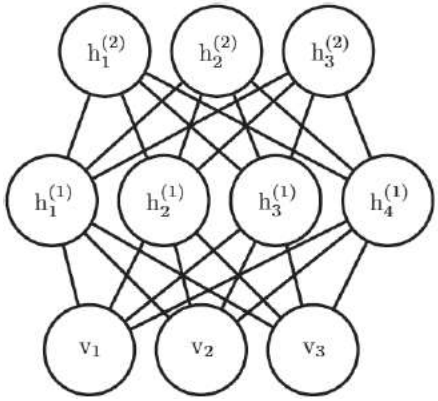


Рис. 1. Графическая модель глубокой машины Больцмана с одним видимым слоем (внизу) и двумя скрытыми слоями. Связи существуют только между блоками из соседних слоев. Внутри слоев никаких связей нет

Подобно ОМБ и ГСД, глубокие машины Больцмана обычно содержат только бинарные блоки (и мы придерживаемся такого предположения для простоты изложения), но включить в них вещественные видимые блоки не составляет труда.

ОМБ – энергетическая модель, а значит, совместное распределение вероятности переменных модели параметризовано функцией энергии E. В случае глубокой машины Больцмана с одним видимым слоем $v$ и тремя скрытыми слоями $h^{(1)}, h^{(2)}, h^{(3)}$ совместное распределение имеет вид:

$$P\left(v, h^{(1)}, h^{(2)}, h^{(3)}\right) = \frac{1}{Z(\theta)} exp\left(-E(v, h^{(1)}, h^{(2)}, h^{(3)}; \theta)\right) \quad (1)$$

Опускаем параметры смещения. Тогда функция ГМБ определяется формулой:

$$E(v, h^{(1)}, h^{(2)}, h^{(3)}; \theta) = -v^{\top} W^{(1)} h^{(1)} - h^{(1)\top} W^{(2)} h^{(2)} - h^{(2)\top} W^{(3)} h^{(3)} (3) \quad (2)$$

По сравнению с функцией энергии ОМБ (2), функция энергии ГМБ включает связи между скрытыми блоками (латентными переменными) в форме

матриц весов ($W^{(2)}$ и $W^{(3)}$). Наличие этих связей имеет важные последствия для поведения модели, а также для выполнения вывода.

По сравнению с полносвязными машинами Больцмана (в которых каждый блок связан со всеми остальными), ГМБ имеет ряд преимуществ, похожих на те, что свойственны ОМБ. Точнее говоря, как видно по рис. 2, слои ГМБ можно представить в виде двудольного графа, в котором нечетные слои принадлежат одной доле, а четные – другой. Отсюда сразу следует, что при условии переменных в четном слое переменные в нечетных слоях становятся условно независимыми. Разумеется, обусловливание переменными нечетных слоев делает условно независимыми переменные в четных слоях.

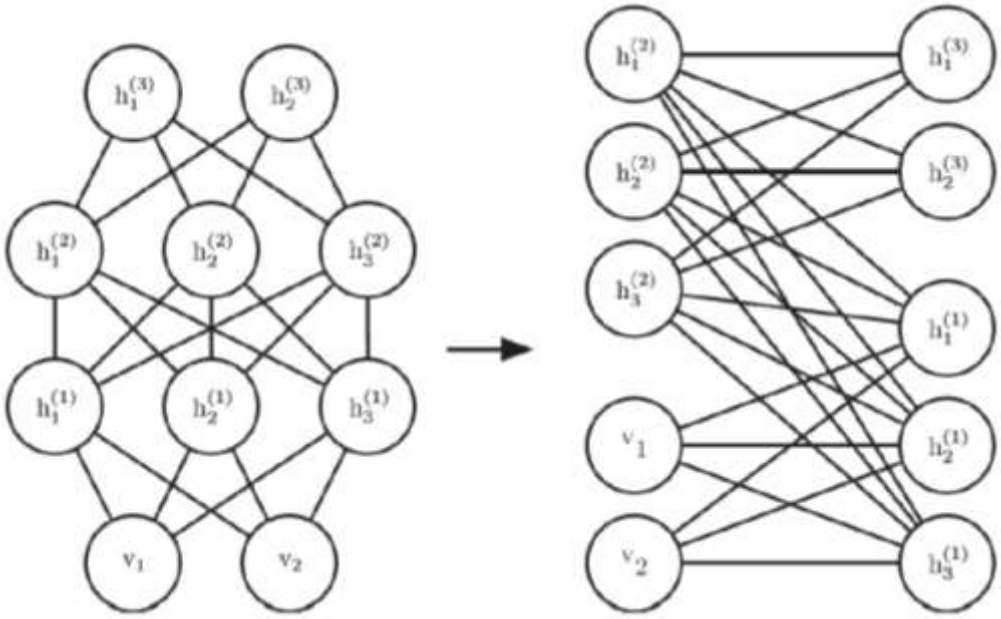


Рис. 2. Глубокая машина Больцмана, нарисованная так, чтобы выявить структуру двудольного графа

Из двудольной структуры ГМБ вытекает, что те же самые уравнения, которые мы раньше использовали для условных распределений ОМБ, применимы и к ГМБ. Блоки внутри одного слоя условно независимы друг от друга при условии значений в соседних слоях, поэтому распределения бинарных переменных можно полностью описать параметрами Бернулли, задающими вероятность активности каждого блока.

В примере с двумя скрытыми слоями вероятности активации равны

$$P(v_i = 1 \mid h^{(1)}) = \sigma(W_{i,:}^{(1)} h^{(1)}), \tag{3}$$

$$P(h_i^{(1)} = 1 \mid v, h^{(2)}) = \sigma(v^\top W_{:,i}^{(1)} + W_{i,:}^{(2)} \mathrm{h}^{(2)}) \tag{4}$$

$:, i$ – $i$-й столбец матрицы $W$ и $i,:$ – $i$-я строка матрицы $W$

и

$$P(h_k^{(2)} = 1 \mid h^{(1)}) = \sigma(h^{(1)\top} W_{:,k}^{(2)}) \quad (5)$$

Благодаря двудольной структуре глубокой машины Больцмана оказывается эффективной выборка по Гиббсу. При наивном подходе к выборке по Гиббсу обновляется по одной переменной за раз. ОМБ позволяет обновлять все видимые блоки в одной операции, а все скрытые – в другой. По наивности можно было бы предположить, что ГМБ с $l$ слоями требует $l + 1$ обновлений, так что на каждой итерации обновляются все блоки одного уровня. На самом же деле для обновления всех блоков достаточно всего двух итераций. Выборку по Гиббсу можно разбить на две группы обновлений: одна включает все четные слои (в т. ч. и видимый), другая – все нечетные. В силу двудольности графа связей в ГМБ распределение нечетных слоев при условии четных факторное, поэтому выборку из него можно произвести одновременно и независимо одной операцией. И так же можно произвести одновременную и независимую выборку из распределения четных слоев при условии нечетных. Эффективная выборка особенно важна для обучения с помощью алгоритма стохастической максимизации правдоподобия.

*Интересные свойства*

У глубоких машин Больцмана много интересных свойств.

ГМБ были разработаны после ГСД (глубокая сеть доверия). По сравнению с ГСД, их апостериорное распределение $P(h|v)$ проще. В противоречие с интуицией, благодаря простоте апостериорного распределения возможны его улучшенные аппроксимации. В случае ГСД мы выполняем классификацию, применяя эвристическую процедуру приближенного вывода, в которой высказываем гипотезу, что разумное значение математического ожидания среднего поля для скрытых блоков можно получить проходом вверх по сети, образованной МСП (многослойным персептроном) с сигмоидными функциями активации и такими же весами, как у исходной ГСД. Для получения вариационной нижней границы логарифмического правдоподобия можно использовать любое распределение $Q(h)$. Поэтому такая эвристическая процедура дает возможность получить нижнюю границу. Однако эта граница явно никак не оптимизировалась и потому может быть далеко не точной. В частности, эвристическая оценка $Q$ игнорирует взаимодействия между скрытыми блоками в одном слое, а также влияние нисходящей обратной связи между скрытыми блоками более глубоких слоев и слоев, расположенных ближе к входу. Поскольку эвристическая процедура вывода на основе МСП не может учесть этих взаимодействий в ГМБ, результирующее распределение $Q$, скорее всего, далеко от оптимального. В ГМБ все скрытые блоки внутри одного слоя условно независимы при условии других слоев. Благодаря отсутствию внутрислойного

взаимодействия открывается возможность использовать уравнения неподвижной точки для оптимизации вариационной нижней границы и нахождения истинно оптимальных математических ожиданий среднего поля (в пределах некоторого допуска).

Использование надлежащего среднего поля позволяет процедуре приближенного вывода в ГМБ уловить влияние нисходящей обратной связи. Это делает ГМБ интересными для нейробиологии, поскольку известно, что человеческий мозг задействует много нисходящих обратных связей. Благодаря этому свойству ГМБ использовались в качестве вычислительных моделей реальных нейробиологических явлений.

Один из недостатков ГМБ – относительная сложность выборки из них. В ГСД выборку МСМС-методами (методы Монте-Карло по схеме марковских цепей) необходимо использовать только в двух верхних слоях. Остальные слои используются лишь в конце процесса выборки, в одном эффективном проходе предковой выборки. Чтобы произвести выборку из ГМБ, необходимо применять МСМС-методы во всех слоях, т. е. каждый слой модели принимает участие во всех переходах марковской цепи.

*Вывод среднего поля в ГМБ*

Условное распределение одного слоя ГМБ при условии соседних слоев факторное. В примере ГМБ с двумя скрытыми слоями это будут распределения $P(v|h^{(1)})$, $P(h^{(1)}|v,h^{(2)})$ и $P(h^{(2)}|h^{(1)})$. Распределение всех скрытых слоев обычно не является факторным из-за взаимодействий между слоями. В примере с двумя скрытыми слоями $P(h^{(1)},h^{(2)}|v)$ не факторизуется из-за весов $W^{(2)}$ взаимодействия между $h^{(1)}$ и $h^{(2)}$, вследствие чего эти переменные оказываются взаимно зависимыми.

Как и в случае с ГСД (глубокая сеть доверия), нам остается искать способы аппроксимации апостериорного распределения ГМБ. Но, в отличие от ГСД, апостериорное распределение скрытых блоков ГМБ, хотя и сложное, легко аппроксимируется вариационной аппроксимацией, а конкретно – приближением среднего поля. Приближение среднего поля – это простая форма вариационного вывода, когда мы ограничиваемся только факторными аппроксимирующими распределениями. В контексте ГМБ уравнения среднего поля улавливают двусторонние взаимодействия между слоями.

Вариационный подход к приближенному выводу предполагает аппроксимацию конкретного целевого распределения – в нашем случае апостериорного распределения скрытых блоков при условии видимых блоков – некоторым достаточно простым семейством распределений. В случае

приближения среднего поля в качестве такого семейства берется множество распределений, для которых скрытые блоки условно независимы.

Теперь разработаем подход на основе среднего поля для примера с двумя скрытыми слоями. Пусть $Q(h^{(1)}, h^{(2)}|v)$ – аппроксимация. Из предположения среднего поля следует, что

$$Q\left(h^{(1)}, h^{(2)}\middle|v\right) = \prod_j Q\left(h_j^{(1)}|v\right) \prod_k Q\left(h_k^{(2)}|v\right) \quad (6)$$

Приближение среднего поля пытается найти член этого семейства распределений, который наилучшим образом аппроксимирует истинное апостериорное распределение $P(h^{(1)}, h^{(2)}|v)$. Важно отметить, что процесс вывода следует запускать снова для нахождения другого распределения $Q$ всякий раз, как используется новое значение $v$.

Можно придумать много способов измерить качество аппроксимации $P(h|v)$ распределением $Q(h|v)$. Подход на основе среднего поля предполагает минимизацию

$$KL(Q||P) = \sum_h Q\left(h^{(1)}, h^{(2)}\middle|v\right) \log\left(\frac{Q(h^{(1)}, h^{(2)}|v)}{P(h^{(1)}, h^{(2)}|v)}\right) \quad (6)$$

где $KL(Q||P)$ – расхождение Кульбака–Лейблера между $P$ и $Q$

Вообще говоря, мы не обязаны предоставлять параметрическую форму аппроксимирующего распределения, а только гарантировать выполнение предположений о независимости. Процедура вариационной аппроксимации сама способна восстановить функциональную форму приближенного распределения. Однако в рассматриваемом случае предположения скрытого поля о бинарных скрытых блоках фиксирование параметризации модели заранее не приводит к потере общности.

Параметризуем $Q$ в виде произведения распределений Бернулли, т. е. ассоциируем параметр с вероятностью каждого элемента $h^{(1)}$. Точнее, для каждого $j$ $\hat{h}_j^{(1)} = Q(h_j^{(1)} = 1|v)$, где $\hat{h}_j^{(1)} \in [0, 1]$, и для каждого $k\hat{h}_j^{(2)} = Q(h_j^{(2)} = 1|v)$, где $\hat{h}_j^{(2)} \in [0, 1]$.

Таким образом, имеем следующую аппроксимацию апостериорного распределения:

$$Q\left(h^{(1)}, h^{(2)}\middle|v\right) = \prod_j Q\left(h_j^{(1)}|v\right) \prod_k Q\left(h_k^{(2)}|v\right) \quad (7)$$

$$= \prod_j \left(\hat{h}_j^{(1)}\right)^{h_j^{(1)}} \left(1 - \hat{h}_j^{(1)}\right)^{\left(1-h_j^{(1)}\right)} \times \prod_k \left(\hat{h}_j^{(2)}\right)^{h_k^{(2)}} \left(1 - \hat{h}_j^{(2)}\right)^{\left(1-h_j^{(2)}\right)} \quad (8)$$

Разумеется, эта параметризация приближенного апостериорного распределения очевидным образом обобщается на ГМБ с большим числом слоев,

нужно только воспользоваться двудольной структурой графа и одновременно обновить сначала все четные слои, а затем все нечетные, применяя такую же схему, как в выборке по Гиббсу.

После того как семейство аппроксимирующих распределений $Q$ определено, остается задать процедуру выбора того члена этого семейства, который лучше всего соответствует $P$. Самое простое – воспользоваться уравнениями среднего поля.

На практике в большинстве случаев не приходится решать вариационные задачи самостоятельно. Вместо этого имеется общее уравнение для обновления неподвижной точки среднего поля. Если принять аппроксимацию среднего поля

$$g(h|v) = \prod_i g(h_i|v) \tag{9}$$

и зафиксировать $q(h_j|v)$ для всех $j \neq i$, то оптимальное распределение $q(h_i|v)$ можно получить нормировкой ненормированного распределения

$$\tilde{q}(h_i|v) = exp(\mathbb{E}h_{-i\sim q(h_{-i}|v)} log\ \tilde{p}(v,h)) \tag{10}$$

при условии что $p$ не назначает вероятность 0 ни одной совместной комбинации переменных. Перенос математического ожидания внутрь уравнения дает корректную функциональную форму $q(h_i|v)$. Непосредственный вывод функциональных форм $q$ методами вариационного исчисления необходим только в случае, когда цель – разработать новый вид вариационного обучения; уравнение (10) дает аппроксимацию среднего поля для любой вероятностной модели. Это уравнение неподвижной точки, которое следует итеративно применять для каждого значения $i$ до достижения сходимости. Однако этим оно не исчерпывается. Оно сообщает нам функциональную форму оптимального решения вне зависимости от того, найдено оно из уравнения неподвижной точки или иным способом. Это означает, что мы можем взять функциональную форму из этого уравнения, но рассматривать некоторые значения в ней как параметры, которые можно оптимизировать с помощью любого алгоритма по своему выбору.

При выводе этих уравнений мы искали, в каких точках обращаются в нуль производные вариационной нижней границы. Они абстрактно описывают, как оптимизировать вариационную нижнюю границу для любой модели, просто взяв математические ожидания относительно $Q$.

Применив эти общие уравнения, получим правила обновления (опять-таки члены смещения игнорируются):

$$\hat{h}_j^{(1)} = \sigma\left(\sum_i v_i W_{i,j}^{(1)} + \sum_{k'} W_{i,k'}^{(2)} \hat{h}_{k'}^{(2)}\right), \forall j \tag{11}$$

$$\hat{h}_k^{(1)} = \sigma\left(\sum_{j'} W_{j',k}^{(2)} \hat{h}_{j'}^{(1)}\right), \forall k \tag{12}$$

В неподвижной точке этой системы уравнений мы имеем локальный максимум вариационной нижней границы $\mathcal{L}(Q)$. Следовательно, эти уравнения обновления неподвижной точки определяют итеративный алгоритм, в котором обновление $\hat{h}_j^{(1)}$ (по формуле 11) чередуется с обновлением $\hat{h}_k^{(2)}$ (по формуле 12). В небольших задачах типа MNIST (черно-белая база данных образцов рукописного написания цифр) достаточно всего десяти итераций, чтобы найти приближенный градиент положительной фазы для обучения, а пятидесяти обычно хватает для получения высококачественного представления одного конкретного примера, используемого для классификации с высокой верностью. Обобщение приближенного вариационного вывода на более глубокие ГМБ не составляет труда.

*Обучение параметров ГМБ*

Вариационный вывод допускает построение распределения $Q(h|v)$, аппроксимирующего неразрешимое распределение $P(h|v)$. Затем максимизируется $\mathcal{L}(v, Q, \theta)$ – вариационная нижняя граница неразрешимого логарифмического правдоподобия, $logP(v; \theta)$.

Для глубокой машины Больцмана с двумя скрытыми слоями функция $\mathcal{L}$ имеет вид:

$$\mathcal{L}(Q, \theta) = \sum_i \sum_{j'} v_i W_{i,j}^{(1)} \hat{h}_{j'}^{(1)} + \sum_{j'} \sum_{k'} \hat{h}_{j'}^{(1)} W_{j',k'}^{(2)} \hat{h}_{k'}^{(2)} - \log Z(Q) + \mathcal{H}(Q) \qquad (13)$$

Это выражение все еще содержит логарифм статистической суммы $logZ(\theta)$. Поскольку глубокая машина Больцмана состоит из ограниченных машин Больцмана, то результаты, касающиеся трудности вычисления статистической суммы и выборки в ограниченных машинах Больцмана, применимы и к ГМБ. Это означает, что для вычисления функции вероятности машины Больцмана необходимы приближенные методы, например выборка по значимости с отжигом. Аналогично для обучения модели требуется аппроксимировать градиент логарифма статистической суммы. ГМБ обычно обучаются с помощью алгоритма стохастической максимизации правдоподобия (рассмотренный на предыдущей лекции). Однако для псевдоправдо-подобия необходимо уметь вычислять ненормированные вероятности, а не просто получать для них вариационную нижнюю границу. Метод сопоставительного расхождения слишком медленный для глубоких машин Больцмана, потому что они не допускают эффективной выборки из распределения скрытых блоков при условии видимых, поэтому метод сопоставительного расхождения должен был бы прирабатывать марковскую цепь всякий раз, как необходим новый пример в отрицательной фазе.

*Послойное предобучение*

К сожалению, обучение ГМБ методом стохастической максимизации правдоподобия со случайными начальными параметрами не годится. В одних случаях модель не может обучиться адекватному представлению распределения, в других ГМБ представляет распределение хорошо, но не удается получить более высокое правдоподобие, чем дала бы простая ОМБ. ГМБ, для которой веса очень малы во всех слоях, кроме первого, представляет приблизительно то же распределение, что и ОМБ.

Оригинальный и самый популярный метод решения проблемы совместного обучения ГМБ – жадное послойное предобучение. В этом случае каждый слой ГМБ обучается изолированно – как ОМБ. Первый слой обучается моделированию входных данных, а каждый последующий – моделированию примеров, выбранных из апостериорного распределения предыдущей ОМБ. После того как все ОМБ обучены, их можно объединить в ГМБ. Затем ГМБ можно обучить методом PCD (устойчивого сопоставительного расхождения (стохастической максимизации правдоподобия)). Обычно такое обучение вносит лишь небольшое изменение в параметры модели и в ее качество, измеряемое по логарифмическому правдоподобию, присвоенному данным, или по способности модели классифицировать входы. Процедура обучения иллюстрируется на рис. 3.

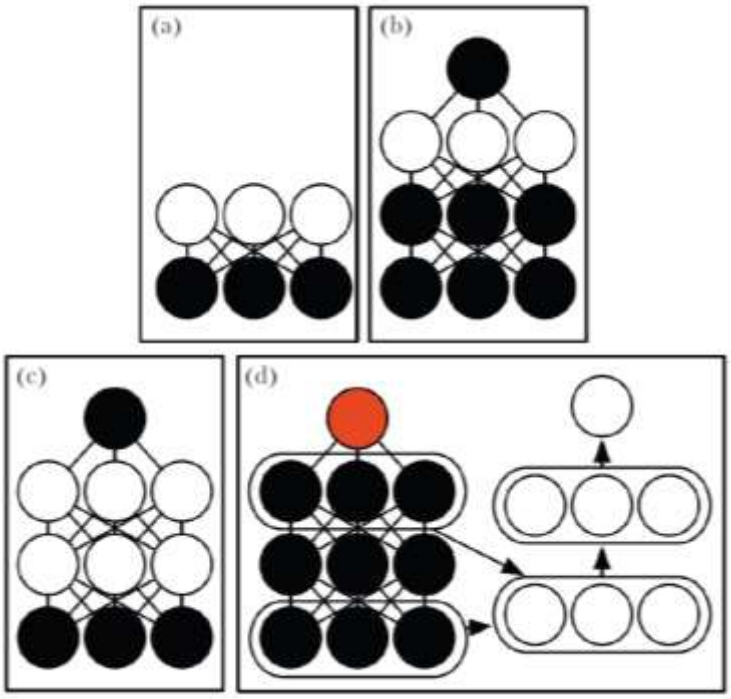


Рис. 3. Процедура обучения глубокой машины Больцмана, использованной для классификации набора данных MNIST

(а) Обучить ОМБ, применив алгоритм CD для приближенной максимизации $logP(v)$.

(b) Обучить вторую ОМБ, которая моделирует $h^{(1)}$ и целевой класс y, применив алгоритм CD-k для приближенной максимизации $logP(h^{(1)}, y)$, где $h^{(1)}$ – выборка из апостериорного распределения первой ОМБ при условии данных. Увеличивать $k$ от 1 до 20 в процессе обучения.

(c) Объединить обе ОМБ в ГМБ. Обучить ее приближенной максимизации $logP(v, y)$, применив алгоритм стохастической максимизации правдоподобия с $k = 5$.

(d) Удалить $y$ из модели. Определить новый набор признаков $h^{(1)}$ и $h^{(2)}$, полученных путем выполнения вывода среднего поля в модели без $y$. Использовать эти признаки в качестве входа МСП, структура которого такая же, как структура дополнительного прохода среднего поля, с дополнительным выходным слоем для оценки $y$. Инициализировать веса МСП (многослойного персептрона) весами ГМБ. Обучить МСП приближенной максимизации $logP(y|v)$, применив алгоритм стохастического градиентного спуска и прореживание.

Алгоритм 1. Алгоритм вариационной стохастической максимизации правдоподобия для обучения ГМБ с двумя скрытыми слоями

Установить размер шага ε равным малому положительному числу.

Установить число шагов выборки по Гиббсу $k$ достаточно большим для приработки марковской цепи $p(v, h^{(1)}, h^{(2)}; \theta + \varepsilon\Delta_\theta)$.

Инициализировать три матрицы $\tilde{V}$, $\tilde{H}^{(1)}$ и $\tilde{H}^{(2)}$ с $m$ строками каждая случайными значениями.

while не сошелся (цикл обучения) do

Выбрать мини-пакет m примеров из обучающих данных и организовать его в виде строк матрицы плана $V$.

Инициализировать матрицы $\widehat{H}^{(1)}$ и $\widehat{H}^{(2)}$, возможно, маргиналами модели.

while не сошелся (цикл вывода среднего поля) do

$\widehat{H}^{(1)} \leftarrow \sigma(\mathrm{V}\mathrm{W}^{(1)} + \widehat{H}^{(2)}\mathrm{W}^{(2)\top})$

$\widehat{H}^{(2)} \leftarrow \sigma(\widehat{H}^{(1)}\mathrm{W}^{(2)})$

end while

$\Delta_{\mathrm{W}^{(1)}} \leftarrow (1/\mathrm{m})\mathrm{V}^\top\widehat{H}^{(1)}$

$\Delta_{\mathrm{W}^{(2)}} \leftarrow (1/\mathrm{m})\widehat{H}^{(1)\top}\widehat{H}^{(2)}$

for $l = 1$ to $k$ (выборка по Гиббсу) do

Блочная выборка по Гиббсу 1:

$\forall i, j, \tilde{V}_{\mathrm{i,j}}$ выбирается из $P(\tilde{V}_{\mathrm{i,j}} = 1) = \sigma(W^{(1)}_{j,:}(\tilde{H}^{(1)}_{i,:})^\top)$.

$\forall i, j, \widetilde{\mathrm{H}}_{i,j}^{(2)}$ выбирается из $P(\widetilde{H}_{i,j}^{(2)} = 1) = \sigma(\widetilde{H}_{i,:}^{(1)} W_{:,j}^{(2)})$.

Блочная выборка по Гиббсу 2:

$\forall i, j, \widetilde{\mathrm{H}}_{i,j}^{(1)}$ выбирается из $P(\widetilde{H}_{i,j}^{(1)} = 1) = \sigma(\tilde{V}_{i,:} W_{:,j}^{(1)} + \widetilde{H}_{i,:}^{(2)} W_{j,:}^{(2)\top})$

end for

$\Delta_{\mathrm{W}^{(1)}} \leftarrow \Delta_{\mathrm{W}^{(1)}} - (1/\mathrm{m})\mathrm{V}^{\top} \widehat{H}^{(1)}$

$\Delta_{\mathrm{W}^{(2)}} \leftarrow \Delta_{\mathrm{W}^{(2)}} - (1/\mathrm{m})\widehat{H}^{(1)\top} \widehat{H}^{(2)}$

$W^{(1)} \leftarrow W^{(1)} + \varepsilon \Delta_{\mathrm{W}^{(1)}}$ (это упрощенная иллюстрация, на практике применяется более эффективный алгоритм, например импульсный с убывающей скоростью обучения)

$W^{(2)} \leftarrow W^{(2)} + \varepsilon \Delta_{\mathrm{W}^{(2)}}$

end while

Эта процедура жадного послойного обучения – не просто покоординатное восхождение. Она действительно напоминает покоординатное восхождение, потому что на каждом шаге мы оптимизируем одно подмножество параметров. Но оба метода отличаются, поскольку в процедуре жадного послойного обучения на каждом шаге используется другая целевая функция.

Жадное послойное предобучение ГМБ отличается от жадного послойного предобучения ГСД. Параметры каждой отдельной ОМБ можно копировать в соответствующую ГСД непосредственно. В случае же ГМБ параметры ОМБ необходимо модифицировать перед включением в ГМБ. Слой в середине стека ОМБ обучается только на входных данных, поступающих снизу, но после того как стек собран в ГМБ, этому слою данные поступают снизу и сверху. Чтобы учесть этот эффект предлагается делить пополам веса всех ОМБ, кроме нижней и верхней, перед тем как вставлять их в ГМБ. Кроме того, нижнюю ОМБ следует обучать с использованием двух «копий» каждого видимого блока со связанными, равными между собой весами. Это означает, что на восходящем проходе веса, по сути дела, удваиваются. Аналогично верхнюю ОМБ следует обучать с использованием двух копий верхнего слоя.

Для получения не уступающих лучшим образцам результатов с помощью глубоких машин Больцмана необходимо модифицировать стандартный алгоритм стохастической максимизации правдоподобия, а именно использовать небольшую толику среднего поля в отрицательной фазе шага совместного обучения методом PCD (устойчивого сопоставительного расхождения (стохастической максимизации правдоподобия)). Точнее говоря, математическое ожидание градиента энергии следует вычислять относительно распределения среднего поля, в котором все блоки независимы. Параметры этого распределения среднего поля

следует получать, выполнив всего одну итерацию уравнений неподвижной точки среднего поля.

*Совместное обучение глубоких машин Больцмана*

Для классической ГМБ требуется жадное предобучение без учителя, а чтобы она хорошо выполняла классификацию, необходим отдельный основанный на МСП классификатор поверх выделенных ей скрытых признаков. У этой схемы есть нежелательные свойства. Трудно следить за качеством в процессе обучения, поскольку мы не можем вычислить свойства полной ГМБ во время обучения первой ОМБ. Поэтому сказать, насколько хорошо выбраны гиперпараметры, можно только, когда процесс обучения зайдет достаточно далеко. Программным реализациям ГМБ нужно много различных компонент: для обучения отдельных ОМБ методом CD (алгоритм сопоставительного расхождения), обучения полной ГМБ методом PCD (алгоритм устойчивого сопоставительного расхождения) и обучения на основе обратного распространения через МСП. Наконец, МСП, построенные поверх машины Больцмана, теряют многие преимущества ее вероятностной модели, например способность выполнять вывод, когда часть входных значений отсутствует.

Существуют два основных способа решить проблему совместного обучения глубокой машины Больцмана. Первый – *центрированная глубокая машина Больцмана*, когда модель перепараметризуется так, чтобы гессиан функции стоимости был лучше обусловлен в начале процесса обучения. В результате получается модель, которую можно обучить без этапа жадного послойного предобучения. Эта модель достигает отличного логарифмического правдоподобия на тестовом наборе и порождает примеры высокого качества. К сожалению, она попрежнему не может конкурировать с правильно регуляризированным МСП (многослойным персептроном) в роли классификатора. Второй способ – использовать *многопредсказательную глубокую машину Больцмана*. В этой модели применяется альтернативный критерий обучения, который позволяет использовать алгоритм обратного распространения, чтобы избежать проблем с МСМС (методами Монте-Карло по схеме марковских цепей)-оценками градиента. К сожалению, новый критерий не приводит к хорошему правдоподобию или выборкам, но, по сравнению с МСМС-методами, производит более качественную классификацию и может хорошо рассуждать об отсутствующих входных данных.

Центрирование машины Больцмана проще всего описать, вернувшись к общему взгляду на машину Больцмана как на множество блоков $x$ с матрицей весов $U$ и смещениями $b$. Функция энергии имеет вид

$$E(x) = -x^{\top}Ux - b^{\top}x. \qquad (14)$$

Применяя различные паттерны разреженности в матрице весов $U$, мы можем реализовать такие структуры машин Больцмана, как ОМБ или ГМБ, с разным числом слоев. Для этого нужно разбить $x$ на видимые и скрытые блоки и обнулить элементы $U$, соответствующие блокам, которые не взаимодействуют. В центрированной машине Больцмана вводится вектор $\mu$, вычитаемый из всех состояний:

$$E'(x;\ U, b)\ = -(x - \mu)^{\top} U(x - \mu) - (x - \mu)^{\top} b \qquad (15)$$

Обычно $\mu$ является гиперпараметром и фиксируется в начале обучения. Как правило, он выбирается, так чтобы $x - \mu \approx 0$ на этапе инициализации модели. Такая перепараметризация не влияет на множество распределений вероятности, представимых моделью, но изменяет динамику стохастического градиентного спуска в применении к правдоподобию. Точнее говоря, во многих случаях перепараметризация дает лучше обусловленную матрицу Гессе. Экспериментально подтверждено, что обусловленность гессиана улучшается, и отмечено, что центрирование эквивалентно другой технике обучения машин Больцмана – расширенному градиенту. Благодаря улучшенной обусловленности гессиана обучение может успешно завершиться даже в трудных случаях типа обучения глубокой машины Больцмана с несколькими слоями.

Другой подход к совместному обучению глубоких машин Больцмана – многопредсказательная глубокая машина Больцмана (МП-ГМБ), идея которой – рассматривать уравнения среднего поля как определение семейства рекуррентных сетей для приближенного решения любой возможной проблемы вывода. Вместо того чтобы обучать модель максимизации правдоподобия, мы обучаем ее, так чтобы каждая рекуррентная сеть получала верный ответ на соответствующую проблему вывода. Процесс обучения показан на рис. 4. Он состоит из трех частей: случайная выборка обучающего примера, случайная выборка подмножества входов сети вывода и обучение сети вывода предсказывать значения остальных блоков.

В строках (рис.4) показаны разные примеры из мини-пакета для одного и того же шага обучения, а в столбцах – временной шаг процесса вывода среднего поля. Для каждого примера мы выбираем подмножество переменных, которое будет служить входом для процесса вывода. Эти переменные закрашены черным, чтобы показать обусловливание. Затем выполняется процесс вывода среднего поля, стрелки показывают влияние одних переменных на другие. В практических приложениях среднее поле разворачивается на несколько шагов, здесь же таких шагов всего два. Штриховые стрелки означают, что процесс можно было бы развернуть еще на несколько шагов. Переменные, которые не подавались на вход

процесса вывода, становятся метками, они закрашены серым цветом. Процесс вывода для каждого примера можно рассматривать как рекуррентную сеть. Мы пользуемся градиентным спуском и обратным распространением, чтобы обучить эти рекуррентные сети порождать правильные метки при известных входах. Тем самым процесс среднего поля для МП-ГМБ обучается давать верные оценки.

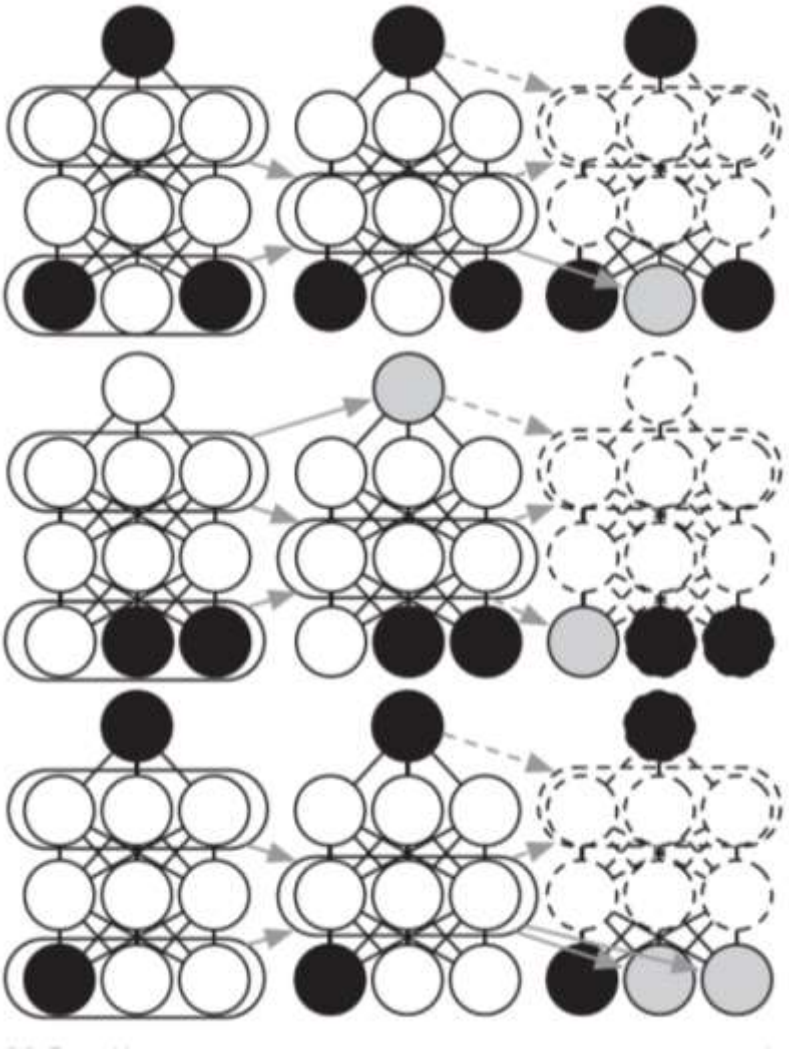

Рис. 4. Иллюстрация многопредсказательного процесса обучения глубокой машины Больцмана.

Этот общий принцип обратного распространения по графу вычислений для приближенного вывода применялся и к другим моделям. И в этих моделях, и в МП-ГМБ окончательная потеря не является нижней границей правдоподобия, а обычно основана на приближенном условном распределении отсутствующих значений, индуцируемом сетью приближенного вывода. Это значит, что в обоснование обучения таких моделей выдвигаются чисто эвристические аргументы. Если исследовать распределение p(v), представленное машиной Больцмана, которая была обучена как МП-ГМБ (многопредсказательная глубокая

машина Больцмана), то оно окажется несовершенным в том смысле, что выборка по Гиббсу дает плохие примеры.

У обратного распространения по графу вывода есть два основных преимущества. Во-первых, он обучает модель так, как она реально используется, – с помощью приближенного вывода. Это означает, что приближенный вывод с целью, например, восполнить отсутствующие входные данные или выполнить классификацию, несмотря на отсутствие части данных, будет более верным при использовании МП-ГМБ, чем оригинальной ГМБ. Оригинальная ГМБ не является верным классификатором сама по себе; наилучшие результаты достигаются, когда на признаках, извлеченных ГМБ, обучается отдельный классификатор, а не когда вывод применяется для вычисления распределения меток классов. Вывод среднего поля в МП-ГМБ хорошо работает в роли классификатора даже без специальных модификаций. Второе преимущество обратного распространения по графу приближенного вывода состоит в том, что вычисляется точный градиент потери. Для оптимизации это лучше, чем приближенные градиенты, вычисляемые алгоритмом стохастической максимизации правдоподобия, которые подвержены как смещению, так и дисперсии. По всей видимости, это объясняет, почему МП-ГМБ можно обучать совместно, тогда как для ГМБ требуется жадное послойное предобучение. Недостаток обратного распространения по графу приближенного вывода – в том, что оно не позволяет оптимизировать логарифмическое правдоподобие, а дает лишь эвристическую аппроксимацию обобщенного псевдоправдоподобия.

Существуют связи между МП-ГМБ и прореживанием. Прореживание означает разделение параметров между несколькими графами вычислений, различие между которыми заключается в том, включает граф некоторый блок или нет. В МП-ГМБ параметры также разделяются между графами вычислений. Но различие между графами состоит в том, наблюдается некоторый входной блок или нет. Если блок не наблюдается, то МП-ГМБ, в отличие от прореживания, не удаляет его полностью, а рассматривает как латентную переменную, подлежащую выводу. Можно было бы представить себе применение прореживания к МП-ГМБ посредством удаления некоторых блоков, вместо того чтобы делать их латентными.