Стохастичні методи навчання

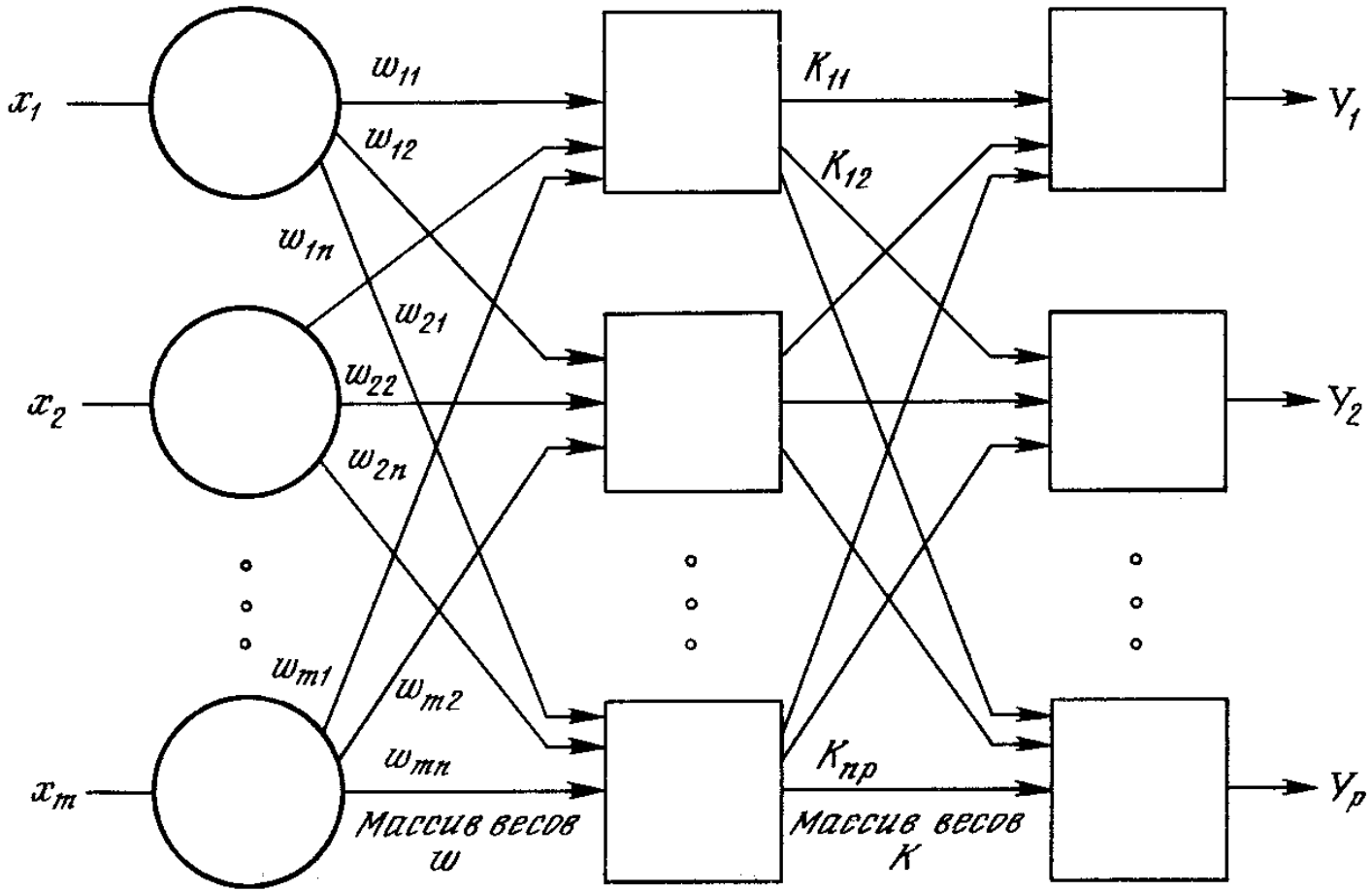
Стохастичні методи корисні як для навчання штучних нейронних мереж, так і для отримання виходу від вже навченої мережі. Стохастичні методи навчання приносять велику користь, дозволяючи виключати локальні мінімуми в процесі навчання. Але з ними також пов'язаний ряд проблем.

Штучна нейронна мережа навчається за допомогою деякого процесу,що модифікує її ваги. Якщо навчання успішне, то пред'явлення мережі безлічі вхідних сигналів призводить до появи бажаної множини вихідних сигналів. Є два класи навчальних методів : детерміністський і стохастичний.

Стохастичні методи навчання

Детерміністський метод навчання крок за кроком здійснює процедуру корекції ваг мережі, засновану на використанні їх поточних значень, а також величин входів, фактичних виходів і бажаних виходів. Навчання персептрона є прикладом подібного детерміністського підходу.

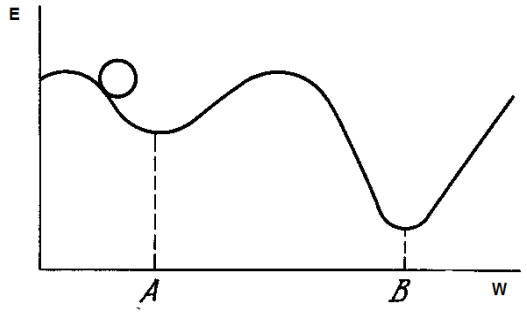
Стохастичні методи навчання виконують псевдовипадкові зміни величин ваг, зберігаючи ті зміни, які ведуть до поліпшень. Щоб побачити, як це може бути зроблено, розглянемо рис., на якому зображена типова мережа, в якій нейрони сполучені за допомогою ваг. Вихід нейрона є тут зваженою сумою його входів, яка, перетворена за допомогою нелінійної функції.


x_1
x_2
x_m
w_{11}
w_{12}
w_{1n}
w_{21}
w_{22}
w_{2n}
w_{m1}
w_{m2}
w_{mn}
Массив весов
w
K_{11}
K_{12}
K_{np}
Массив весов
K
Y_1
Y_2
Y_p

Процедура стохастичного навчання

1. Вибрати вагу випадковим чином і підкоригувати її на невелике випадкове Пред'явити безліч входів і вичислити виходи, що виходять.
2. Порівняти ці виходи з бажаними виходами і вичислити величину різниці між ними. Загальноприйнятий метод полягає в знаходженні різниці між фактичним і бажаним виходами для кожного елементу навченої пари, зведення різниць в квадрат і знаходження суми цих квадратів. Метою навчання є мінімізація цієї різниці, часто званою цільовою функцією.
3. Вибрати вагу випадковим чином і підкоригувати її на невелике випадкове значення. Якщо корекція допомагає (зменшує цільову функцію), то зберегти її, інакше повернутися до первинного значення ваги.
4. Повторювати кроки з 1 до 3 до тих пір, поки мережа не буде навчена достатньою мірою.

Пастки локальних мінімумів докучають усім алгоритмам навчання, заснованим на пошуку мінімуму, включаючи персептрон і мережі зворотного поширення, і представляють серйозну і широко поширену трудність, якої часто не помічають.



Стохастичні методи дозволяють розв'язати цю проблему. Стратегія корекції ваг, що змушує ваги набувати значення глобального оптимуму в точці B, можлива.

- Якщо коробку сильно потрясти в горизонтальному напрямі, то кулька швидко перекочуватиметься від одного краю до іншого. Ніде не затримуючись, в кожен момент кулька з рівною імовірністю знаходитиметься в будь-якій точці поверхні.
- Якщо поступово зменшувати силу струшування, то буде досягнута умова, при якій кулька на короткий час "застряватиме" в точці В. При ще слабкішому струшуванні кулька на короткий час зупинятиметься як в точці А, так і в точці В.
- При безперервному зменшенні сили струшування буде досягнута критична точка, коли сила струшування достатня для переміщення кульки з точки А в точку В, але недостатня для того, щоб кулька могла видертися з В в А.
- Таким чином, остаточно кулька зупиниться в точці глобального мінімуму, коли амплітуда струшування зменшиться до нуля.

- Штучні нейронні мережі можуть навчатися по суті тим же самим чином за допомогою випадкової корекції ваг. Спочатку робляться великі випадкові корекції із збереженням тільки тих змін ваг, які зменшують цільову функцію. Потім середній розмір кроку поступово зменшується, і глобальний мінімум врешті-решт досягається.
- Це дуже нагадує відпал металу, тому для її опису часто використовують термін "імітація відпалу". У металі, нагрітому до температури, що перевищує його точку плавлення, атоми знаходяться в сильному безладному русі.
- Як і в усіх фізичних системах, атоми прагнуть до стану мінімуму енергії (єдиному кристалу в даному випадку), але при високих температурах енергія атомних рухів перешкоджає цьому. В процесі поступового охолодження металу виникають усе більш низько енергетичні стани, поки врешті-решт не буде досягнуто найнижче з можливих станів, глобальний мінімум.
- В процесі відпалу розподіл енергетичних рівнів описується наступним співвідношенням:
    - $P(e) = \exp(-e/kT)$
- де *P(e)* - вірогідність того, що система знаходиться в змозі з енергією e; k - постійна Больцмана; T - температура за шкалою Кельвіна.
- При високих температурах P(e) наближається до одиниці для усіх енергетичних станів. Таким чином, високо енергетичний стан майже так же ймовірний, як і низько енергетичний. У міру зменшення температури вірогідність високо енергетичних станів зменшується в порівнянні з низько енергетичними. При наближенні температури до нуля стає дуже маловірогідним, щоб система знаходилася у високо енергетичному стані.

Навчання Больцмана.

Цей стохастичний метод безпосередньо застосований до навчання штучних нейронних мереж :

Визначити змінну Т, що представляє штучну температуру. Надати Т велике початкове значення.

Пред'явити мережі безліч входів і вичислити виходи і цільову функцію.

Дати випадкову зміну вазі і перерахувати вихід мережі і зміну цільової функції відповідно до зробленої зміни ваги.

Якщо цільова функція зменшилася (покращала), то зберегти зміну ваги.

Якщо зміна ваги призводить до збільшення цільової функції, то вірогідність збереження цієї зміни обчислюється за допомогою розподілу Больцмана :

$$P(c) = \exp(-c/kT)$$

де P(c) - вірогідність зміни з в цільовій функції; k - константа, аналогічна константі Больцмана, вибирана залежно від завдання; Т - штучна температура.

Вибирається випадкове число r з рівномірного розподілу від нуля до одиниці. Якщо P(c) більше, ніж r, то зміна зберігається, інакше величина ваги повертається до попереднього значення.

Це дозволяє системі робити випадковий крок в напрямі, що псує цільову функцію, дозволяючи їй тим самим вириватися з локальних мінімумів, де будь-який малий крок збільшує цільову функцію.

Для завершення больцманівского навчання повторюють кроки 3 і 4 для кожної з ваг мережі, поступово зменшуючи температуру T, поки не буде досягнуто допустимо низьке значення цільової функції. У цей момент пред'являється інший вхідний вектор і процес навчання повторюється.

Мережа навчається на усіх векторах навчальної множини, з можливим повторенням, поки цільова функція не стане допустимою для усіх них.

Величина випадкової зміни ваги на кроці 3 може визначатися різними способами. Наприклад, подібно до теплової системи вагова зміна w може вибиратися відповідно до розподілу Гауса :

$P(w) = \exp(-w2/T2)$

де P(w) - вірогідність зміни ваги на величину w, T - штучна температура.

Такий вибір зміни ваги призводить до системи, аналогічно.

Оскільки потрібна величина зміни ваги Δw, а не вірогідність зміни ваги, що має величину w, то метод Монте-Карло може бути використаний таким чином:

1. Знайти кумулятивну вірогідність, відповідну P(w). Це є інтеграл від P(w) в межах від 0 до w. Оскільки в даному випадку P(w) не може проінтегрувати аналітично, вона повинна інтегруватися чисельно, а результат потрібне затабулировать.
2. Вибрати випадкове число з рівномірного розподілу на інтервалі (0,1). Використовуючи цю величину як значення P(w}, знайти в таблиці відповідне значення для величини зміни ваги.

Властивості машини Больцмана широко вивчалися. У роботі показано, що швидкість зменшення температури має бути обернено пропорційна логарифму часу, щоб була досягнута збіжність до глобального мінімуму. Швидкість охолодження в такій системі виражається таким чином:

$$T(t) = \frac{T_0}{\log(1+t)}$$

де T(t) - штучна температура як функція часу; T0 - початкова штучна температура; t - штучний час.

Цей результат, що розчаровує, передбачає дуже повільну швидкість охолодження (і ці обчислення). Цей вивід підтвердився експериментально. Машини Больцмана часто вимагають для навчання дуже великого ресурсу часу.

Навчання Коші

У цьому методі при обчисленні величини кроку розподіл Больцмана замінюється на розподіл Коші. Розподіл Коші має, як показано на рис. , довші "хвости", збільшуючи тим самим вірогідність великих кроків.

Насправді розподіл Коші має нескінченну (невизначену) дисперсію. За допомогою такої простої зміни максимальна швидкість зменшення температури стає обернено пропорційній лінійній величині, а не логарифму, як для алгоритму навчання Больцмана. Це різко зменшує час навчання. Цей зв'язок може бути виражений таким чином:

Розподіл Коши має вигляд

$$T(t) = \frac{T_0}{1 + t}$$

де P(x) є вірогідність кроку величини x.

$$P(x) = \frac{T(t)}{T(t)^2 + x^2}$$

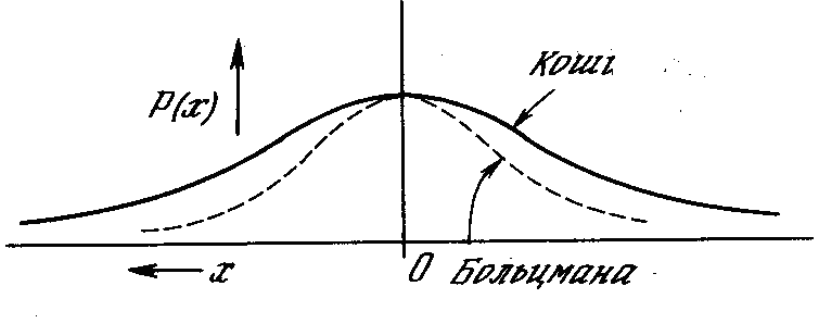


Рис. Розподіл Коші і розподіл Больцмана

У рівнянні P(x) можна про інтегрувати стандартними методами. Вирішуючи відносно x, отримуємо

$$xc = (T(t) tg(P(x)),$$

де xc - зміна ваги.

Тепер застосування методу Монте-Карло стає дуже простим. Для знаходження x в цьому випадку вибирається випадкове число з рівномірного розподілу на відкритому інтервалі $(-\pi/2, \pi/2)$ (необхідно обмежити функцію тангенса). Воно підставляється у формулу як P(x), і за допомогою поточної температури обчислюється величина кроку.

Обме́жена маши́на Бо́льцмана

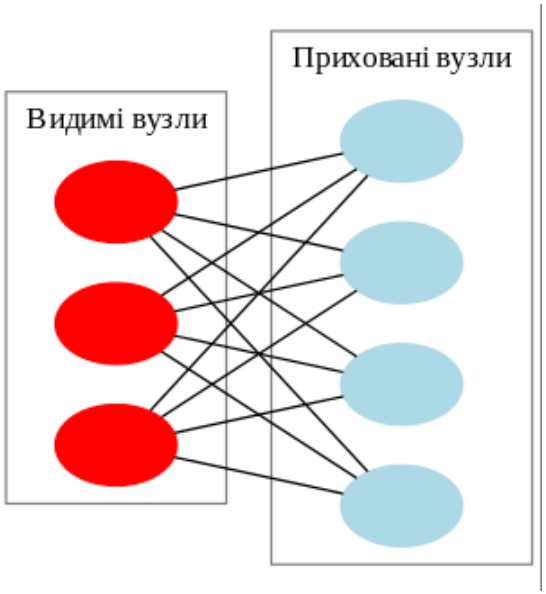
(ОМБ, англ. *restricted Boltzmann machine, RBM*) — це породжувальна стохастична штучна нейронна мережа, здатна навчатися розподілу ймовірностей над набором її входів.

ОМБ було спочатку винайдено під назвою **Гармоніум** (англ. *Harmonium* — фісгармонія) Полом Смоленським у1986 році, а популярності вони набули після винайдення Джефрі Хінтоном зі співавторами у середині 2000-х років алгоритмів швидкого навчання для них.

ОМБ знайшли застосування у зниженн розмірності, класифікації, колаборативній фільтрації, навчанні ознак та тематичному моделюванні. Їх може бути треновано як керованим, так і спонтанним чином, в залежності від завдання.

Обме́жена маши́на Бо́льцмана

Як випливає з їхньої назви, ОМБ є варіантом машин Больцмана, з тим обмеженням, що їхні нейрони мусять формувати двочастковий граф: пара вузлів з кожної з двох груп вузлів (що, як правило, називають «видимим» та «прихованим» вузлами відповідно) можуть мати симетричне з'єднання між ними, але з'єднань між вузлами в межах групи не існує.



Обме́жена маши́на Бо́льцмана

Обмежені машини Больцмана можуть також застосовуватися в мережах глибинного навчання. Зокрема, глибинні мережі переконань можуть утворюватися «складанням» ОМБ та, можливо, тонким налаштуванням отримуваної в результаті глибинної мережі за допомогою градієнтного спуску та зворотного поширення.

На противагу, «необмежені» машини Больцмана можуть мати з'єднання між прихованими вузлами. Це обмеження уможливлює ефективніші алгоритми тренування, ніж доступні для загального класу машин Больцмана, зокрема, алгоритм **порівня́льної розбі́жності** (англ. *contrastive divergence*) на основі градієнтного спуску.

Обме́жена маши́на Бо́льцмана

Стандартний тип ОМБ має двійковозначні (булеві/бернуллієві) приховані та видимі вузли, і складається з матриці вагових коефіцієнтів W = ($w_{i,j}$) (розміру *m*×*n*), пов'язаної зі з'єднанням між прихованим вузлом h_j та видимим вузлом v_i, а також вагових коефіцієнтів упереджень (зсувів) a_i для видимих вузлів, і b_j для прихованих вузлів.

З урахуванням цього, *енергія* конфігурації (пари булевих векторів) (*v*,*h*) визначається як

$$E(v,h) = -\sum_i a_i v_i - \sum_j b_j h_j - \sum_i \sum_j v_i w_{i,j} h_j$$

або, в матричному записі,

$$E(v,h) = -a^{\mathrm{T}} v - b^{\mathrm{T}} h - v^{\mathrm{T}} W h$$

Ця функція енергії є аналогічною до функції енергії мережі Хопфілда. Як і в загальних машинах Больцмана, розподіли ймовірності над прихованими та/або видимими векторами визначаються в термінах функції енергії:

$$P(v,h) = \frac{1}{Z} e^{-E(v,h)}$$

Обме́жена маши́на Бо́льцмана

де Z є статистичною сумою, визначеною як сума над усіма можливими конфігураціями (іншими словами, просто нормувальна стала для забезпечення того, щоби розподіл імовірності давав у сумі 1). Аналогічно, (відособлена) ймовірність видимого (вхідного) вектора булевих значень є сумою над усіма можливими конфігураціями прихованого шару:

$$P(v) = \frac{1}{Z} \sum_{h} e^{-E(v,h)}$$

Оскільки ОМБ має [illegible] ового графу, без з'єднань усередині шарів, активації прихованих вузлів є взаємно незалежними для заданих активацій видимих вузлів, і навпаки, активації видимих вузлів є взаємно незалежними для заданих активацій прихованих вузлів.

Обме́жена маши́на Бо́льцмана

Тобто, для m видимих вузлів та n прихованих вузлів умовною ймовірністю конфігурації видимих вузлів *v* для заданої конфігурації прихованих вузлів *h* є

$$P(v \mid h) = \prod_{i=1,m} P(v_i \mid h).$$

І навпаки, умовною ймовірністю *h* для заданої *v* є

$$P(h \mid v) = \prod_{j=1,n} P(h_j \mid v).$$

Окремі ймовірності активації задаються як

$$P(h_j = 1 \mid v) = \sigma\left(b_j + \sum_{i=1,m} w_{i,j} v_i\right),$$

$$P(v_i = 1 \mid h) = \sigma\left(a_i + \sum_{j=1,n} w_{i,j} h_j\right),$$

де σ позначає логістичну сигмоїду.

Незважаючи на те, що приховані вузли є бернуллієвими, видимі вузли ОМБ можуть бути багатозначними.

В такому випадку логістична функція для видимих вузлів замінюється багатозмінною логістичною функцією активації (*Softmax function*)

$$P(v_{ik} = 1 \mid h) = \exp\left(a_{ik} + \Sigma_j W_{ijk} h_j\right) \Sigma_{k=1,K} \exp\left(a_{ik} + \Sigma_j W_{ijk} h_j\right),$$

де *K* є кількістю дискретних значень, які мають видимі значення.