Нейронные сети. Краткий курс

Лекция 7

Модели на основе теории информации

Рассмотрим информационно-теоретические модели, которые приводят к самоорганизации. В этих моделях синаптические связи многослойной нейронной сети организуются так, чтобы максимизировать объем информации, которая сохраняется при преобразовании сигнала на каждой стадии обработки сигнала в сети. Главной функцией систем обработки сигнала является раскрытие и использование некоторой избыточности информации в той или иной форме.

Вспомним некоторые понятия из теории информации, а именно случайной величины из некоторого алфавита $X=\{x_1, x_2, ..., x_N\}$ каждому элементу которого соответствует вероятность с которой он встречается $P=\{p_1, p_2, ..., p_N\}$. Количество информации согласно Шеннону, которое переносит переменная x_k определяется как

$$I(x_k) = \log\frac{1}{p_k} = -\log p_k \quad . \tag{7.1}$$

Энтропия определяется как среднее значение количества информации $I(x_k)$ по ансамблю

$$H(X) = E[I(x_k)] = \sum_{k=1}^{N} p_k I(x_k) = -\sum_{k=1}^{N} p_k \log p_k \quad . \tag{7.2}$$

Энтропия это неотрицательная величина удовлетворяющая неравенству:

$$0 \le H(X) \le \log N \quad . \tag{7.3}$$

Равенство 0 в (7.3) выполняется когда $p_k = 1$ для некоторого k и равна 0 для остальных. В этом случае имеет место отсутствие неопределенности. Равенство $\log N$ в (7.3) имеет место когда $p_k = \frac{1}{N}$ для всех k , то есть все переменные равновероятны.

Дивергенция Кульбака-Лейблера определяется как

$$D_{p\|q} = \sum_{k=1}^{N} p_k \log\frac{p_k}{q_k} \quad , \tag{7.4}$$

где p и q это два различных распределения случайной величины X .

Теперь необходимо распространить величины определенные для дискретной случайной переменной на непрерывные переменные. Аналогично энтропии для дискретной случайной величины, определяется *дифференциальная энтропия* непрерывной случайной величины, которая характеризуется плотностью вероятности $f_X(x)$:

$$h(X) = -\int_{-\infty}^{+\infty} f_X(x)\log[f_X(x)]dx = -E[\log f_X(x)] \quad . \tag{7.5}$$

Эта величина легко распространяется на случай векторной случайной величины $\boldsymbol{X}=(X_1, X_2, ..., X_n)^T$:

$$h(\boldsymbol{X}) = -\int_{\mathbb{R}^n} f_X(\boldsymbol{x})\log[f_X(\boldsymbol{x})]d\boldsymbol{x} \quad , \tag{7.6}$$

где $f_X(\boldsymbol{x})$ - плотность совместной вероятности для компонент вектора $\boldsymbol{x}$. Дифференциальная энтропия обладает следующими свойствами:

1) $h(X+c) = h(X)$, где c - неслучайная константа.
2) $h(aX) = h(X)+\log|a|$, где a - масштабирующий множитель. Если обобщить это свойство на случай масштабирующей матрицы $\boldsymbol{A}$ и случайного вектора $\boldsymbol{X}$, то свойство можно записать как $h(\boldsymbol{A}\boldsymbol{X}) = h(\boldsymbol{X})+\log|det(\boldsymbol{A})|$

Принцип максимума энтропии

Пусть у нас есть некоторая стохастическая система с множеством известных состояний, но неизвестными вероятностями этих состояний. Процесс обучения заключается в выборе такого распределения (модели), которое является оптимальным в некотором смысле при наличии априорных знаний о модели. Подобрать распределение позволяет *принцип максимума энтропии*, применение которого означает нахождения максимума

$$h(X) = -\int_{-\infty}^{+\infty} f_X(x)\log\left[f_X(x)\right]dx \tag{7.7}$$

на всех функциях распределения $f_X(x)$ удовлетворяющих следующим условиям:

1) $f_X(x) \geq 0$,
2) $\int_{-\infty}^{+\infty} f_X(x)\,dx = 1$,

3) $\int_{-\infty}^{+\infty} f_X(x) g_i(x) dx = \alpha_i,\ i=1,2,...,m$, где $g_i(x)$ - некоторые функции от x .

Можно увидеть, что условие 3) можно трактовать как условие на моменты случайной величины, а 2) можно включить в 3) считая что $g_0(x)\equiv 1,\ \alpha_0=1$.

Итак имеем задачу условной оптимизации, для решения которой можно использовать *метод множителей Лагранжа*. Запишем функционал метода:

$$J(f_X) = \int_{-\infty}^{+\infty} \left[-f_X(x)\log f_X(x) + \lambda_0 f_X(x) + \sum_{i=1}^{m} \lambda_i g_i(x) f_X(x) \right] dx \quad . \tag{7.8}$$

Дифференцируя этот функционал, можно получить условие

$$-1 - \log f_X(x) + \lambda_0 + \sum_{i=1}^{m} \lambda_i g_i(x) = 0 \quad . \tag{7.9}$$

Из этого условия можно выразить неизвестное распределение

$$f_X(x) = \exp\left(-1+\lambda_0+\sum_{i=1}^{m} \lambda_i g_i(x) \right) \quad , \tag{7.10}$$

где множители Лагранжа определяются из условий 2) и 3).

Рассмотрим следующий пример. Если задаться случайной переменной X , которая характеризуется средним μ и дисперсией σ^2 , а это означает, что условия 3) можно записать как

$$\int_{-\infty}^{+\infty} f_X(x)\, x\, dx = \mu \quad \text{и} \quad \int_{-\infty}^{+\infty} f_X(x)\, (x-\mu)^2\, dx = \sigma^2 \quad , \tag{7.11}$$

то окажется, что функция $f_X(x)$ это нормальное распределение

$$f_x(X) = \frac{1}{\sqrt{2\pi}\,\sigma} \exp\left(-\frac{(x-\mu)^2}{2\sigma^2} \right) \quad . \tag{7.12}$$

Это означает, что гауссова случайная величина имеет наибольшую энтропию при заданных условиях 3) в виде (7.11), то есть если X - гауссова, а Y - любая другая, то $h(X)\geq h(Y)$, причем равенство достигается при $f_X\equiv f_Y$.

При создании самоорганизующихся систем главная цель – разработка такого алгоритма, который способен обучаться отображению входа нейронной сети на выход на основе только входного сигнала. В дальнейшем будем связывать случайную переменную X со входом нейронной сети, а случайную переменную Y с ее выходом. Вспомним теперь

определение *взаимной информации* из теории информации. Условная энтропия

$$H(X|Y) = H(X,Y) - H(Y) \tag{7.13}$$

определяет уровень оставшейся неопределенности относительно X после того как было получено наблюдение Y, причем эта величина неотрицательна и удовлетворяет неравенству $0 \le H(X|Y) \le H(X)$. Первое слагаемое справа в (7.13) это совместная энтропия

$$H(X,Y) = -\sum_{x \in X} \sum_{y \in Y} p_{X,Y}(x,y) \log p_{X,Y}(x,y) \quad , \tag{7.14}$$

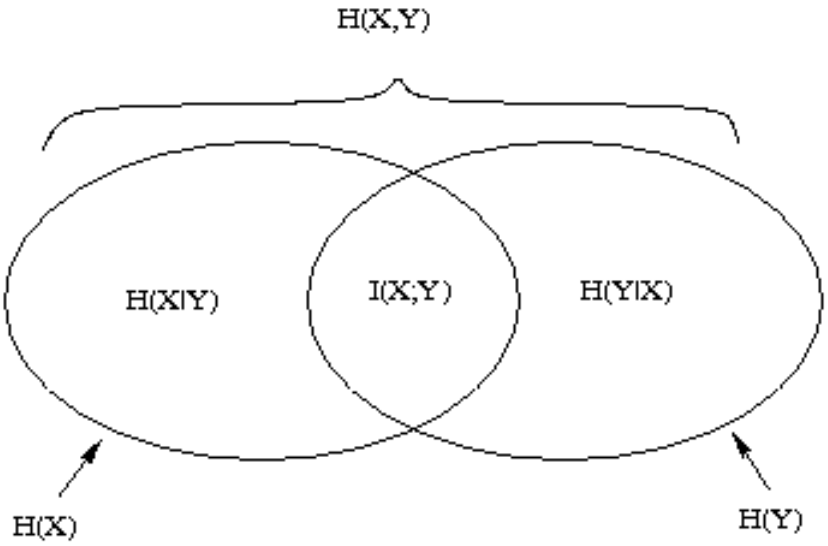
где $p_{X,Y}(x,y)$ - совместная вероятность. И наконец взаимная информация это

$$I(X;Y) = H(X) - H(X|Y) = \sum_{x \in X} \sum_{y \in Y} p_{X,Y}(x,y) \log \left(\frac{p_{X,Y}(x,y)}{p_X(x)\, p_Y(y)} \right) \quad . \tag{7.15}$$

Взаимная информация определенная в (7.15) обладает следующими свойствами:

1) симметрии, то есть $I(X;Y) = I(Y;X)$;
2) неотрицательна, то есть $I(X;Y) \ge 0$;
3) может быть записана через энтропию выхода, то есть $I(X;Y) = H(Y) - H(Y|X)$.

Величина $H(Y|X)$ несет информацию о помехах в системе, а не о входном сигнале. Ниже приведен рисунок представляющий собой визуальную интерпретацию соотношений между введенными выше величинами.



Теперь распространим величины из теории информации на непрерывные случайные величины, которые характеризуются функцией плотности вероятности. Взаимная информация в этом случае

$$I(X;Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\log\left(\frac{f_X(x|y)}{f_X(x)}\right)dx\,dy \quad , \tag{7.16}$$

где $f_X(x)$ – функция плотности вероятности, $f_{X,Y}(x,y)$ - функция плотности совместной вероятности, а $f_X(x|y)$ - функция плотности условной вероятности. Если вспомнить из теории вероятностей, что $f_{X,Y}(x,y)=f_X(x|y)f_Y(y)$ то выражение для взаимной информации можно переписать как

$$I(X;Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\log\left(\frac{f_{X,Y}(x,y)}{f_X(x)f_Y(y)}\right)dx\,dy \quad . \tag{7.17}$$

По аналогии со свойствами взаимной информации для дискретных величин, взаимная информация для непрерывных величин так же симметрична, неотрицательна и может быть выражена несколькими способами, то есть

$$I(X;Y) = h(X)-h(X|Y) = h(Y)-h(Y|X) = h(X)+h(Y)-h(X,Y) \quad . \tag{7.18}$$

Здесь $h(X)$ - дифференциальная энтропия, а условная энтропия $h(X|Y)$ определяется как

$$h(X|Y) = \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f_{X,Y}(x,y)\log[f_X(x|y)]\,dx\,dy \quad . \tag{7.19}$$

Взаимная информация $I(X;Y)$ равна нулю, тогда когда X и Y статистически независимы. В этом случае $f_{X,Y}(x,y)=f_X(x)f_Y(y)$ и $f_X(x|y)=f_X(x)$. Понятие взаимной информации можно обобщить и на случай векторных непрерывных случайных величин:

$$I(\boldsymbol{X};\boldsymbol{Y}) = \int_{\mathbb{R}^n}\int_{\mathbb{R}^m} f_{X,Y}(\boldsymbol{x},\boldsymbol{y})\log\left(\frac{f_X(\boldsymbol{x}|\boldsymbol{y})}{f_X(\boldsymbol{x})}\right)d\boldsymbol{x}\,d\boldsymbol{y} \quad , \tag{7.20}$$

где n и m размерности соответствующих пространств пространств.

Дивергенция Кульбака-Лейблера также распространяется на непрерывные величины, в том числе и векторные:

$$D_{f_X\|g_X} = \int_{\mathbb{R}^m} f_X(\boldsymbol{x})\log\left(\frac{f_X(\boldsymbol{x})}{g_X(\boldsymbol{x})}\right)d\boldsymbol{x} \quad . \tag{7.21}$$

Эта величина неотрицательна и равна нулю когда $f_X \equiv g_X$. Кроме того, она инвариантна к

изменениям порядка компонентов вектора $\boldsymbol{x}$, к масштабированию амплитуды и к монотонным нелинейным преобразованиям случайных векторов. Взаимная информация может быть интерпретирована в терминах дивергенции Кульбака-Лейблера:

$$I(\boldsymbol{X};\boldsymbol{Y}) = \int_{\mathbb{R}^n}\int_{\mathbb{R}^m} f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y}) \log\left(\frac{f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})}{f_{\boldsymbol{X}}(\boldsymbol{x}) f_{\boldsymbol{Y}}(\boldsymbol{y})}\right) d\boldsymbol{x}\, d\boldsymbol{y} = D_{f_{\boldsymbol{X},\boldsymbol{Y}} \| f_{\boldsymbol{X}} f_{\boldsymbol{Y}}} \quad . \tag{7.22}$$

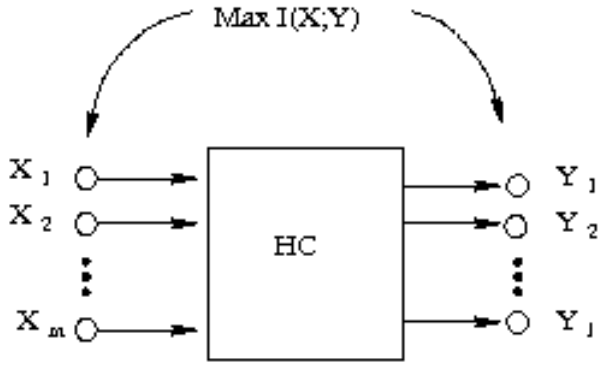
Таким образом взаимная информация между X и Y равна дивергенции Кульбака-Лейблера между плотностями вероятности $f_{\boldsymbol{X},\boldsymbol{Y}}(\boldsymbol{x},\boldsymbol{y})$ и $f_{\boldsymbol{X}}(\boldsymbol{x}) f_{\boldsymbol{Y}}(\boldsymbol{y})$.

Способы построения систем (общий взгляд).

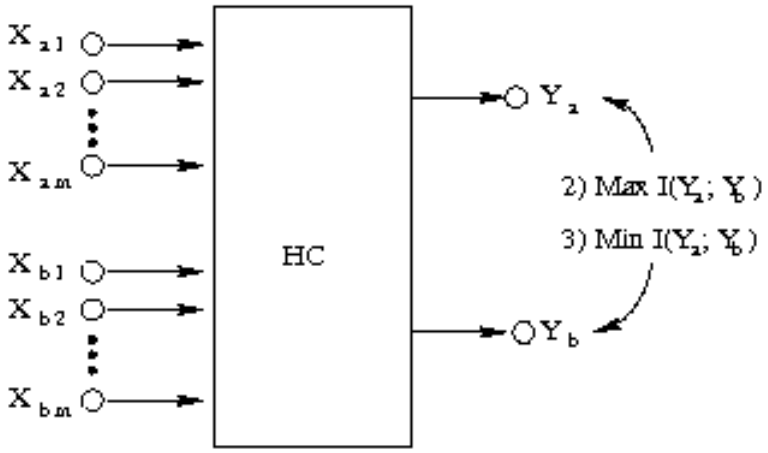
Рассмотрим нейронную сеть с некоторым количеством входов и выходов. Цель – добиться самоорганизации сети для выполнения поставленной задачи типа извлечения статистически значимых признаков или разделения сигналов. В таких системах взаимная информация является оптимизируемой целевой функцией для оптимизации которой подгоняются свободные параметры сети.

На практике можно идентифицировать четыре возможных сценария, которые можно описать следующим образом.

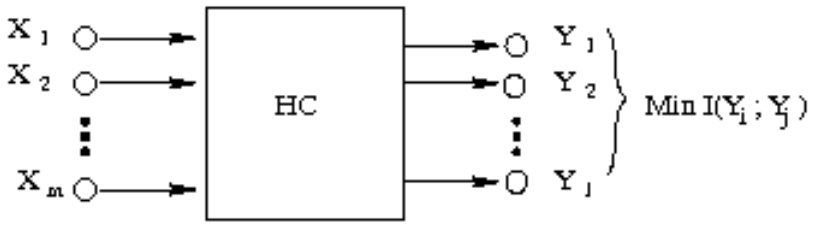
1) Входной вектор $\boldsymbol{X}=(X_1, X_2, ..., X_m)^T$, выходной вектор $\boldsymbol{Y}=(Y_1, Y_2, ..., Y_l)^T$, требуется максимизировать информацию о входе системы передаваемую на выход.



2) Пара входных векторов $\boldsymbol{X}_a$ и $\boldsymbol{X}_b$ порождена смежными, но не пересекающимися областями образа. Входы $\boldsymbol{X}_a$ и $\boldsymbol{X}_b$ производят скалярные выходы Y_a и Y_b . Требуется максимизировать информацию передаваемую из Y_a в Y_b .
3) То же, что и 2), но требуется минимизировать информацию передаваемую из Y_a в Y_b .



4) Входной и выходной векторы определяются как и в 1), но требуется минимизировать статистическую зависимость между компонентами выходного вектора.



Во всех описанных случаях главную роль играет взаимная информация. Однако способ ее формулирования зависит от конкретной задачи.