Лекция 1

Введение

Проблема искусственного интеллекта (ИИ) является ключевой для решения широкого круга задач рассматриваемых в компьютерных науках. Внедрение элементов ИИ в системы автоматического управления, диагностики, прогнозирования и т. п. позволяет достичь существенного превосходства над аналогичными системами, не обладающими свойствами ИИ. Поэтому, с первых лет появления электронных вычислительных машин (ЭВМ) проблема разработки методов повышающих «интеллектуальность» вычислительной системы привлекает большое внимание разработчиков. До настоящего времени сложилась классическая парадигма ИИ, которая опирается на компьютерную технику традиционной последовательной архитектуры. В рамках данной парадигмы основным средством решения интеллектуальных задач являются экспертные системы (ЭС). Однако, объективные ограничения компьютерных систем фон-Неймановской архитектуры, не позволяют решать некоторые интеллектуальные задачи, сложность которых превышает реальные возможности таких систем. Поэтому, наравне с классическими моделями вычислений, в последнее время широкое распространение получили и нетрадиционные, обладающие рядом существенных преимуществ. Среди них, одной из основных стала модель, так называемых нейровычислений. По своей сути - это новое ответвление математики – *нейроинформатика*, предметом изучения которой являются *искусственные нейронные сети* и их вычислительные возможности. На основе теоретических работ в этой области создаются практические реализации нейронных сетей, которые бывают либо программными, либо аппаратными. Программная реализация – это, как правило, программа, написанная на универсальном языке программирования, которая обладает всеми функциональными возможностями нейронной сети. Аппаратная реализация нейронной сети может быть выполнена на нейрочипах (на СБИС, содержащих фрагменты нейронных сетей), на нейроускорителях (являются дополнительными модулями к стандартной ЭВМ, приспособлены для решения нейросетевых задач), оптоэлектронным способом. Нейрокомпьютером обычно называют вычислительную систему, приспособленную для выполнения операций, типичных для функционирующей модели нейросети. Нейрокомпьютеры принято относить к следующему поколению вычислительных систем.

Нейрокомпьютеры обладают рядом достоинств, являющихся причиной растущего интереса к ним со стороны специалистов:

1. Нейронные сети обладают возможностью воспроизведения сложных нелинейных зависимостей. Иными словами, нейронные сети являются универсальными аппроксимирующими устройствами. Можно также сказать, что нейронные сети дают стандартный способ решения многих нестандартных задач.

2. Вместо программирования – обучение. Принцип нахождения решения задачи существенно отличается от решения задач на современных ЭВМ. Отыскание решения задачи на нейрокомпьютере – это составление обучающих примеров. Таким образом, труд программиста замещается новым трудом – учителя.

3. Нейрокомпьютеры особенно эффективны там, где нет чётко сформулированных правил поиска решения и, возможно, нет строгой формальной постановки задачи (распознавание образов, перевод с одного естественного языка на другой и т.п.). Для таких задач обычно трудно сочинить явный алгоритм.

4. Простота элементной базы модели нейрокомпьютера: можно различными способами комбинировать простые составляющие нейрокомпьютеров – нейроны и связи между ними. За счет этого на одной элементной базе можно создавать совершенно различные машины.

5. Высокое быстродействие в случае реализации НС на аппаратном уровне. Это обусловлено распараллеливанием решаемой задачи на уровне данных. Математическое обеспечение на базе нейронных сетей, позволяет решить задачу распараллеливания для широкого класса задач.

6. Высокая помехоустойчивость: структурная избыточность НС позволяет продолжать функционирование НС даже в случае отказа части её элементов.

Разработки в области нейроинформатики ведутся по следующим направлениям:

- разработка нейроалгоритмов;
- программные реализации нейронных сетей: создание программного обеспечения для моделирования нейронных сетей;
- разработка специализированных процессорных плат для имитации нейросетей;
- электронные реализации нейронных сетей: создание нейрочипов;
- оптоэлектронные реализации нейронных сетей.

Для начала дадим неформальное определение *искусственных нейронных сетей.*

Искусственные нейронные сети (*НС*) можно определить как совокупность моделей биологических нейронных сетей. Представляют собой сеть элементов — искусственных нейронов — связанных между собой некоторым образом. Сеть обрабатывает входную информацию и, в процессе изменения своего состояния во времени, формирует совокупность выходных сигналов.

Работа сети состоит в преобразовании входных сигналов во времени, в результате чего меняется внутреннее состояние сети и формируются выходные воздействия. Обычно НС оперирует цифровыми, а не символьными величинами. Большинство моделей НС требуют обучения. В общем случае, *обучение* — такой выбор параметров сети, при котором сеть лучше всего справляется с поставленной проблемой. Обучение — это задача многомерной оптимизации, и для ее решения существует множество алгоритмов.

Каждый нейрон в свою очередь состоит из синапсов, на которые поступает входной сигнал, из устройства преобразующего сигнал и из аксонов, передающих сигнал на выход. Такое устройство нейрона, как основу возможных механизмов памяти и поведения, предложили МакКаллок и Питтс в 1943 году [1]. Затем, в 1949 году Хебб предложил модель обучения нейронов человеческого мозга [2]. В 1958 году Розенблатт с коллегами разработали модель работы мозга - персептрон, которая могла обучаться [3]. Однако после появления в 1969 году работы Минского и Пейперта «Персептроны» [6], исследования нейронных сетей были приостановлены. Дело в том, что в своей работе Минский и Пейперт проанализировали алгоритм обучения персептрона, предложенный Розенблаттом, и показали, что эта модель имеет существенные ограничения. В частности, персептрон Розенблатта в принципе не может решать проблему исключающего ИЛИ. Новые стимулы развития теории нейронных сетей были созданы работами Хопфилда в 1982 году [7]. В середине 1980-х началась новая волна развития теории искусственных нейронных сетей. Так, уже в 1987 году по нейронным сетям было опубликовано свыше 500 научных сообщений.

Сейчас искусственные нейронные сети применяются для решения очень многих задач обработки изображений, управления роботами и непрерывными производствами, для понимания и синтеза речи, для диагностики заболеваний людей и технических неполадок в машинах и приборах, для предсказания курсов валют и результатов скачек. Та часть работ, которая связана с разработкой устройств переработки информации на основе принципов работы естественных нейронных систем, относится к области *нейроинформатики* или *нейровычислений*.

Сведения из высшей математики

Традиционно используемым для описания нейронных сетей математическим языком является аппарат векторной и матричной алгебры.

Векторные пространства

Основным структурным элементом в описании способов обработки информации нейронной сетью является вектор - упорядоченный набор чисел, называемых компонентами вектора. В дальнейшем вектора будут обозначаться латинскими буквами $(\bar{a}, \bar{b}, \bar{c}, \bar{d})$, а скаляры - числа - греческими буквами (a, b, g, d). Для обозначения матриц будут применяться заглавные латинские буквы. Компоненты вектора могут быть действительными числами, целыми числами, булевыми числами "ноль-один" или "минус один - один". Компоненты вектора $\bar{x} = (x_1, x_2, x_3, \mathbf{K}, x_n)$ можно рассматривать, как его координаты в некотором n-мерном пространстве. В случае действительных компонент это пространство обозначается, как R^n и включает в себя набор всех возможных совокупностей из n действительных чисел. Говорят, что вектор $\bar{x}$ принадлежит пространству R^n $(\bar{x} \in R^n)$. В дальнейшем, если нам потребуется набор векторов, мы будем нумеровать их верхними индексами, чтобы не путать с нумерацией компонент: $\{\bar{x}^1, \bar{x}^2, \mathbf{K}, \bar{x}^k\}$.

Определение. Множество векторов с действительными компонентами является частным случаем более общего понятия, называемого *линейным векторным пространством* V, если для его элементов определены операции векторного сложения и умножения на вещественное число, удовлетворяющие перечисленным ниже соотношениям.

Для любых $x, y, z \in V$; $a, b \in R$

1) переместительный закон сложения: $x + y = y + x$;
2) сочетательный закон сложения: $(x + y) + z = x + (y + z)$;
3) существование нулевого элемента: Если $0 \in V \quad x \in V \Rightarrow 0 + x = x \in V$;
4) существование противоположного элемента: $\forall x \in V$ существует $-x \in V$, такой, что $x + (-x) = 0$
5) $1 \cdot x = x$,
6) Сочетательность умножения на скаляр: $(a \cdot b) \cdot x = a \cdot (b \cdot x)$;
7) Распределительный закон: $a \cdot (x + y) = a \cdot x + a \cdot y, (a + b) \cdot x = a \cdot x + b \cdot x$.

Элементами линейного пространства являются векторы.

Примером линейного векторного пространства является пространство R^n с покомпонентными операциями сложения и умножения.

Определение. Для двух элементов векторного пространства может быть определено их *скалярное (внутреннее) произведение*: $(x,y)=x_1\cdot y_1+x_2\cdot y_2+\mathbf{K}+x_n\cdot y_n$. Скалярное произведение обладает следующими свойствами:

1. $(x,y)=(y,x)$ - симметричность,
2. $(a\cdot x,y)=a\cdot(x,y)$ - линейность,
3. $(x+y,z)=(x,z)+(y,z)$ - аддитивность по каждому сомножителю, а также
4. $(x,x)\ge 0$, причем $(x,x)=0 \Leftrightarrow x=0$ - неотрицательность.

Линейное векторное пространство, в котором определено скалярное произведение, удовлетворяющее аксиомам 1-4, называется *евклидовым*.

Равенство нулю скалярного произведения двух векторов означает их взаимную ортогональность, сообразно обычным геометрическим представлениям.

Для векторов можно ввести понятие *нормы* – длины вектора. Пространство, в котором определена норма векторов называется *нормированным*.

Определение. Норма в пространстве E - это функция, определенная для каждого вектора пространства, обозначаемая символом $|\;|$ и обладающая следующими свойствами:

$\forall x,y\in E$

1. $|x|\ge 0$, причем $|x|=0\Leftrightarrow x=0$
2. $|a\cdot x|=|a|\cdot|x|$
3. $|x+y|\le|x|+|y|$

Пример. Для Евклидова пространства можно ввести норму как $|x|=\sqrt{(x,x)}$.

Для векторов, состоящих из действительных чисел, мы будем в дальнейшем иметь дело именно с Евклидовым пространством. В случае булевых векторов размерности n рассматриваемое пространство представляет собой множество вершин n-мерного гиперкуба с Хемминговой метрикой. Расстояние между двумя вершинами определяется длиной кратчайшего соединяющего их пути, измеренной вдоль ребер:

$$d_H(x,y)=\sum_{i=1}^{n}(x_i\oplus y_i),\quad x,y\in V^n=\{0,1\}^n.$$

Евклидова метрика для прямоугольной системы координат определяется формулой:

$$d_E(x,y)=|x-y|=\sqrt{(x-y,x-y)}=\sqrt{(x_1-y_1)^2+(x_2-y_2)^2+\mathbf{K}+(x_n-y_n)^2},\quad x,y\in R^n$$

Расстояние в данном случае может трактоваться как мера сходства двух образов.

Еще одно важное понятие, которое понадобиться в дальнейшем – это угол между двумя векторами. Углом между векторами x и y назовём число

$$j = \arccos \frac{(x, y)}{|x| \cdot |y|}.$$

Вектора, для которых $j = p$, то есть $(x,y)=0$ будем называть *ортогональными*.

Базис линейного векторного пространства

Определение. Вектора $x^1, x^2, \mathbf{K}, x^m$ считаются *линейно независимыми*, если их произвольная линейная комбинация $a_1 x^1 + a_2 x^2 + \mathbf{K} + a_m x^m$ не обращается в ноль, если только все константы $a_1, \mathbf{K}, a_m$ не равны одновременно нулю.

Определение. В линейном векторном пространстве L размерности n можно подобрать такой минимальный набор векторов $\{e^1, e^2, \mathbf{K}, e^n\}$, что любой вектор $x \in L$ можно представить в виде линейной комбинации векторов этого набора. Такой набор векторов называется *базисом*. Можно доказать, что базис может состоять из любой комбинации из n линейно независимых векторов, где n - размерность пространства.

Выберем некоторую систему линейно независимых векторов $x^1, x^2, \mathbf{K}, x^m$, $m < n$. Все возможные линейные комбинации этих векторов сформируют линейное пространство размерности m, которое будет являться *подпространством* или *линейной оболочкой* L исходного n-мерного пространства. Выбранная базовая система из m векторов является, очевидно, базисом в подпространстве L. Важным частным случаем линейной оболочки является подпространство размерности на единицу меньшей, чем размерность исходного пространства, называемое *гиперплоскостью*. В случае трехмерного пространства это обычная плоскость. Гиперплоскость делит пространство на две части. Совокупность гиперплоскостей разбивает пространство на несколько множеств, каждое из которых содержит вектора с близким набором признаков, тем самым осуществляется классификация векторов.

Для двух подпространств может быть введено понятие их взаимной ортогональности. Два подпространства L_1 и L_2 называются взаимно ортогональными, если всякий элемент одного подпространства ортогонален каждому элементу второго подпространства.

Ортогонализация набора векторов

Произвольно выбранные линейно независимые вектора необязательно являются взаимно ортогональными. Однако в ряде приложений удобно работать с ортогональными системами. Для этого исходные вектора требуется ортогонализовать. Классический процесс ортогонализации Грама-Шмидта состоит в следующем: по системе линейно независимых ненулевых векторов $x^1, x^2, \ldots, x^m$ рекуррентно строится система ортогональных векторов $h^1, h^2, \ldots, h^m$. В качестве первого вектора h^1 выбирается исходный вектор x^1. Каждый следующий (i-ый) вектор делается ортогональным всем предыдущим, для чего из него вычитаются его проекции на все предыдущие вектора:

$$h^i = x^i - \sum_{j=1}^{i-1} \frac{(x^i, h^j)}{\left|h^j\right|^2} h^j$$

При этом, если какой-либо из получившихся векторов h^i оказывается равным нулю, он отбрасывается. Можно показать, что, по построению, полученная система векторов оказывается ортогональной, т.е. каждый вектор содержит только уникальные для него признаки.

Матрицы и линейные преобразования векторов

Равно тому, как был рассмотрен вектор - объект, определяемый одним индексом (номером компоненты или признака), может быть введен и объект с двумя индексами, *матрица*. Эти два индекса определяют компоненты матрицы A_{ij}, располагаемые по строкам и столбцам, причем первый индекс i определяет номер строки, а второй j - номер столбца. В некоторых случаях упорядоченный набор элементов может трактоваться и как вектор и как матрица (например, последовательность двадцати пяти чисел).

Суммой двух матриц A и B одинаковой размерности $(n \times m)$ является матрица C той же размерности с компонентами, равными сумме соответствующих компонент исходных матриц: $C_{ij} = A_{ij} + B_{ij}$. Матрицу можно умножить на скаляр, при этом в результате получается матрица той же размерности, каждая компонента которой умножена на этот скаляр. Произведением двух матриц $A(n \times l)$ и $B(l \times m)$ является матрица $C(n \times m)$, компоненты которой даются соотношением:

$$C_{ij} = \sum_{k=1}^{l} A_{ik} B_{kj} .$$

Заметим, что размерности перемножаемых матриц должны быть согласованными - число столбцов первой матрицы должно равняться числу строк второй.

В важном частном случае, когда вторая матрица является вектором (т.е. матрицей с одной из размерностей, равной единице (m=1)), представленное правило определяет способ умножения матрицы на вектор:

$$c_i = \sum_{k=1}^{l} A_{ik} b_k \,.$$

В результате умножения получается также вектор $\bar{c}$, причем для квадратной матрицы $A(l \times l)$, его размерность равна размерности вектора-сомножителя $\bar{b}$. При произвольном выборе квадратной матрицы A можно построить произвольное линейное преобразование одного вектора $(\bar{x})$ в другой $(\bar{y})$ той же размерности: $\bar{y} = A\bar{x}$. Более точно, для того, чтобы преобразование T одного вектора в другой являлось линейным, необходимо и достаточно, чтобы для двух векторов $\bar{x}^1$ и $\bar{x}^2$ и чисел a и b выполнялось равенство: $T(a\bar{x}^1 + b\bar{x}^2) = aT(\bar{x}^1) + bT(\bar{x}^2)$. Можно показать, что всякому линейному преобразованию векторов соответствует умножение исходного вектора на некоторую матрицу.

Если в приведенной выше формуле для умножения матрицы A на вектор $\bar{x}$ компоненты этого вектора неизвестны, в то время, как A и результирующий вектор b известны, то о выражении $A\bar{x} = b$ говорят, как о системе линейных алгебраических уравнений относительно компонент вектора $\bar{x}$. Система имеет единственное решение, если вектора, определяемые строками квадратной матрицы A, являются линейно независимыми.

Часто используемыми частными случаями матриц являются диагональные матрицы, у которых все элементы вне главной диагонали равны нулю. Диагональную матрицу, все элементы главной диагонали которой равны единице, называют единичной матрицей I. Линейное преобразование, определяемое единичной матрицей, является тождественным: $I\bar{x} = \bar{x}$ для всякого вектора $\bar{x}$.

Для матриц определена, кроме операций умножения и сложения, также операция транспонирования. Транспонированная матрица A^T получается из исходной матрицы A заменой строк на столбцы: $A_{ij}^T = A_{ji}$. Матрицы, которые не изменяются при транспонировании, называют симметричными матрицами. Для компонент симметричной матрицы S имеет место соотношение $S_{ij} = S_{ji}$. Всякая диагональная матрица, очевидно, является симметричной.

Пространство квадратных матриц одинаковой размерности с введенными операциями сложения и поэлементного умножения на скаляр, является линейным пространством. Для него также можно ввести метрику и норму. Нулевым элементом служит матрица, все элементы которой равны нулю.

В заключении приведем некоторые тождества для операций над матрицами. Для всяких A, B и C и единичной матрицы I имеет место:

$$IA = AI = A,$$
$$(AB)C = A(BC),$$
$$A(B + C) = AB + AC,$$
$$(A^T)^T = A,$$
$$(A + B)^T = A^T + B^T,$$
$$(AB)^T = B^T A^T.$$

Литература

1. С. А. Терехов. Лекции по теории и приложениям искусственных нейронных сетей. http://alife.narod.ru/lectures/neural/
2. Курош А. Г. Лекции по общей алгебре. «Наука», Москва, 1973.
3. В. А. Ильин, Э. Г. Позняк. Линейная Алгебра, Москва, Наука, 1999.