Пролегомены к высокому искусственному интеллекту

Конспект лекций

Автор конспекта:

Король Михаил

Содержание

Лекция 1. Вводная

Теги, ассоциирующиеся с высоким искусственным интеллектом:

- Многозадачность (Теория бифуркаций(интуиция - теория катастроф), Теория самоорганизации)
- Обучение (Теория адаптивных систем, Теория многоагентных систем)
- Самосознание (Theory of self, теория самоорганизации)
- Сложность (Теория сложности)
- Структурность (Теория сложных сетей)

Теория сложных сетей. Важные вопросы:

1) Размеры

2) Эволюция

3) Распределение графов

Степенное распределение (heavy tail distribution)

$$P(X) = C \cdot X^{-\alpha}$$

Где C - константа нормализации, обеспечивающая требование

$$\int_{-\infty}^{+\infty} P(X)\ dx = 1$$

С математической точки зрения степенные распределения порождаются неким аналогом Центральной предельной теоремы при нарушении формальных требований независимости, конечных математического ожидания и дисперсии. Более того, результат взаимодействия бесконечного числа взаимодействующих случайных величин дает нам четырех-параметрическое семейство функций плотности, который при одном конкретном наборе параметров даст нам нормальное распределение, а при всех остальных значениях параметров ассимптотически при $X \to \infty$ дадут нам одно из степенных распределений при том или ином зачении α.

Какими свойствами обладает данное распределение?

$$\mathbb{E}(X) = \int_{-\infty}^{+\infty} xP(x)\ dx$$

$$Var(X) = \mathbb{E}(X - \mathbb{E}(X))$$

Существуют такие значения α, при которых $\mathbb{E}(X) \to \infty$, более того, если мы снизим требования к α, то мы войдем в область, где математическое ожидание конечно, а дисперсия бесконечна.

Лекция 2. Свойства сложных сетей.

Первое свойство носит название гигантской связанной компоненты.

Наблюдение за реальными сложными сетями указывает, что они не просто эволюционируют (меняют количество ребер и вершин) но и имеют тенденцию к росту. Это позволило применить к ним классический прием естественно научного исследования, который носит название «переход к термодинамическому пределу» или «континуализация». А именно мы исследуем что происходит с объектом, если число составляющих его элементов (в данном случае вершин графа) стремится к бесконечности.

Мы, разумеется, понимаем, что реальные сложные сети конечны, но вместе с тем, мы предполагаем, что, начиная с некоторого большого N, мы можем говорить о некоторых асимптотических свойствах, то есть, начиная с некоторого достаточно большого N сложная сеть будет сохранять те же свойства, что и сеть, обладающих «бесконечным количеством вершин».

Здесь было установлено, что для всех сложных сетей мы наблюдаем несвязанность сложных сетей как графов. Сложные сети состоят из некоторого (иногда достаточно большого) количества несвязанных компонент. Но вместе с тем, из этих компонент выделяется одна, число вершин в которой по порядку совпадает с числом вершин во всем графе.

$$N_{GCC} = O(N), N \to \infty$$

В случае ориентированных графов, мы должны модифицировать понятие гигантской связанной компоненты. Она разбивается на четыре составляющих:

1) Гигантская сильно связанная компонента. Здесь предполагается, что из любых вершин i и j мы можем достигнуть из вершины i вершину j, из вершины j вершину i.

2) Гигантская выходная компонента. Это множество вершин, в которые мы можем попасть из вершин гигантской сильно связанной компоненты.

3) Гигантская входная компонента. Это множество вершин, из которых мы можем попасть в вершины гигантской сильно связанной компоненты.

4) Так называемые усы, специальная структура, которая представляет собой линейно упорядоченную последовательность вершин, исходящих из гигантской сильно связанной компоненты.

Более того, возвращаясь к неориентированным графам, мы получаем, что для характеризации сложных сетей мы должны ввести свойство его разреженности. Традиционно, разреженность графа характеризуют как отношение фактического числа ребер к максимально возможному.

$$\rho = \frac{E}{(N(N-1))/2}$$

При этом, мы пользуемся той же идеей перехода к термодинамическому пределу, мы смотрим, как ведет себя величина ρ не для данной конкретной сложной сети, но для последовательности сетей, с увеличивающимся размером, при $N \to \infty$.

Очевидно, что если граф полносвязный, или близкий к полносвязному (неразреженный), тогда величина ρ будет вести себя как $O(1)$, поскольку $E \sim O(N^2)$.

С другой стороны, если мы имеем дело с чем-то вроде минимального остовного дерева, где $E \sim O(N)$, то $\rho \to 0$. Если мы будем наблюдать промежуточную ситуацию, где $E \to O(N^\alpha), 1 < \alpha < 2$, то мы говорим о разреженном графе.

Все сложные сети являются разреженными графами.

Второе свойство носит название Малого мира.

Путем между вершинами i_0 и i_n называется последовательность ребер $(i_0, i_1), (i_1, i_2), ..., (i_{n-1}, i_n)$ такая, что первое ребро инцидентно вершине i_0, а последнее вершине i_n. Кратчайшим путем между вершинами i_0 и i_n является путь, содержащий минимальное число ребер. Далее, на основании этих конструкций мы должны построить некоторые характеристики, которые характеризуют не отдельную пару вершин, но граф в целом. А именно:

1) Диаметром графа называется максимальный из путей, где $l_{i,j}$ – длина кратчайшего пути, соединяющего вершины i и j

$$d_G = \max_{i \neq j} l_{ij}$$

2) Эксцентриситетом вершины i мы будем называть максимальную длину кратчайшего пути, соединяющий вершины j и k, не проходящей через вершину i:

$$ec(i) = \max_{j,k \in V \setminus \{i\}} l_{j,k}$$

Эксцентриситетом вершины называется расстояние до самой дальней вершины

$$ec(i) = \max_{i \neq j} l_{i,j}$$

3) Тогда радиусом графа G будет минимальный эксцентриситет.

$$r_G = \min_i ec(i)$$

4) Самое ходовое и самое эффективное на практике – средняя длина кратчайшего пути в графе

$$< l >= \frac{1}{N(N-1)/2} \sum_{i \neq j}^{N} l_{ij}$$

Если мы наблюдаем что-то вроде полносвязности ($< l > \sim O(1)$) – это простая сеть.

Если мы возьмем что-то похожее на кристалическую решетку, это тоже будет простая сеть порядка $O\left(n^{1 \cdot d}\right)$, где d – размерность.

Оказалось, что если расстояние ведет себя как $O\left(n^{\beta}\right)$, то речь идет о какой-то вариации простой сети.

Классическим примером сложной системы являются системы, у которых среднее расстояние – это величина порядка логарифма числа вершин.

$$< l > \sim O(\ln N)$$

Для реализации такого рода системы нам необходимо существование специальных вершин – хабов, которые характеризуются тем, что через них проходит много кратчайших путей, эти вершины обеспечивают связность графа.

В отношении хабов, как всегда в математике, мы можем ставить две задачи:

1) Отыскание, обнаружение. Это прямая задача теории хабов.

2) Обратная задача, которая заключается в конструировании сети таким образом, что удаление даже значительного числа его хабов не приводит ни к потери связанности, ни даже к нарушению нормального функционирования сети, протекания потоков.

Если для сети выполняется такое свойство (свойство 2), то мы будем говорить, что сеть структурно устойчива (resilent). В настоящее время именно организация структурно устойчивых бесхабовых сетей является одной из наиболее значимых.

Все задачи в математике делятся на три больших класса:

1) Прямые задачи, есть некоторое описание реального процесса, структура, уравнение и подобное.

2) Обратные задачи, имеется некоторое множество наблюдений реального процесса, мы пытаемся по этим наблюдениям восстановить процесс, который имеет место в реальном мире.

3) Задача управления – имеется возможность каким-то образом воздействовать на объект, с которым мы работаем, и мы должны добиться того, чтобы наше воздействие приводило к желательному результату.

Лекция 3. Ассортативность и дисассортативность.

Все сложные сети делятся на два больших класса, которые отличаются взаимоотношением хабов друг с другом.

В ассортативных сетях хабы имеют тенденцию быть связанными друг с другом непосредственно, как например это имеет место в интернете.

Диссортативные сети характеризуются тем, что их хабы имеют тенденцию быть связанными друг с другом через цепочку не хабов (вершин с малыми степенями). К таким типам сетей относятся экологические и тропические сети (волк и тигр имеют много связей, являются хабами, но при этом напрямую друг с другом не связаны).

Это отличие (фундоментальная дихотомия) является первым вопросом, на который мы должны ответить, приступая к изучению сложной сети.

Чтобы ответить на вопрос, является ли сеть ассортативной или дисассортативной, мы должны построить следующее распределение условной вероятности:

$$P(k \mid k_1, ..., k_n)$$

где k – степень просматриваемой вершины, n – число ее соседей, $k_1, ..., k_n$ – их степени.

Проблема заключается в том, что любая характеристика, которую мы хотим использовать на практике, должна отвечать нескольким требованиям. Это касается не только сильного искусственного интеллекта.

1) Она должна измерять ту категорию, которую мы рассматриваем, при этом мы должны наблюдать не только корреляцию (в статистическом смысле) этой величины и исследуемой категории.

2) Мы должны уметь предъявлять логически прозрачный и ясный механизм, который объясняет, почему наша измеримая характеристика действительно описывает теоретическое понятие. Чтобы избежать конструкций в стиле «влияние лунного света на рост телеграфных столбов».

3) Характеристика должна быть практически измерима. Здесь мы отбрасываем ситуации, что мы можем посчитать эту статистику, но нам потребуется время вычисления суперкомпьютера сопоставима со временем существования вселенной.

4) Характеристика должна быть робасной (в первом приближении вычислительно устойчивой) (если мы немного изменим выборку, то значение характеристики тоже должно немного измениться)

Почему $P(k \mid k_1, ..., k_n)$ нам не подходит? Пусть в нашей сложной сети не больше ста соседей, то вообще говоря, общее число вариантов, зашитое в этой вероятности $102^{101} \sim 10^{1000}$, таким образом, это невычислимо. Допустим мы сделали 10^{1000} наблюдений, и даже в этой ситуации мы только один раз попадем в соответствующую область вероятности, а для хорошей оценки нужно попасть в каждую область вероятности несколько сотен, тысяч раз. Из этого следует вычислительная неустойчивость. Если мы в соответствующую область попали один раз, то если мы попадем во второй раз, мы можем получить что-то другое, что является статистически неустойчивой ситуацией.

Нам нужно сделать понятие ассортативности вычислимым.

Что такое математическая статистика? Мы с вами говорим, что в прикладной математике задачи делятся на прямые, обратные, и задачи управления. Вся теория вероятности по своему построению, по своему существу, является прямой задачей. Мы постулируем вероятности неких элементарных событий и пытаемся ответить на вопрос, каковы же вероятности каких-то не элементарных событий. Математическая статистика является обратной задачей для теории вероятности. Мы пытаемся по наблюдениям оценить вероятности событий, которые представляют для нас интерес. Общая схема математической статистики состоит в формировании статистического критерия (теста), который представляет из себя алгоритм, позволяющий с ошибкой, не превосходящей заданного небольшого уровня (уровня значимости) отвечать на вопрос, верна ли некоторая статистическая гипотеза.

1) Коэффициент парной корреляции является механизмом проверки гипотезы, что две различные выборки коррелируют друг с другом.

Формально, если я имею случайную величину X и выборку $x_1, ..., x_n$ ей порожденной и случайную величину Y и выборку $y_1, ..., y_n$ ей порожденной, то выдвинув нулевую гипотезу H_0: коррелируют друг с другом против гипотезы H_1: не коррелируют / слабо коррелируют, мы должны построить величину эмперического коэффициента парной корреляции z_{xy}

$$z_{xy} = \frac{Cov(x, y)}{\sqrt{Var(X)Var(Y)}}, -1 \leq z_{xy} \leq 1$$

$$Cov(x, y) = \frac{1}{n} \sum_{i=1}^{n} (x_i - \overline{x})(y_i - \overline{y})$$

Для того, чтобы установить, является сеть ассортативной или дисассортативной, была предложена простая и сильная идея. Пусть в качестве случайной величины X выступает степень вершины рассматриваемой сети. А в качестве случайной величины Y степень вершины, с которой данная вершина связана через любое ребро. Мы выбираем все ребра сложной сети и записываем степени инцидентных им вершин. В этой ситуации, если сеть является ассортативной, то есть хабы связаны с хабами, а малостепенные вершины связаны с малостепенными вершинами, то коэффициент парной корреляции будет большим и близким к единице. Если же напротив сеть дисассоративна, то есть хабы связаны с малостепенными вершинами, то коэффициент парной корреляции близок к -1. Если же сеть случайна, то есть она не представляет собой отражение реального физического объекта, а представляет собой некую математическую генерацию, то, соответственно, коэффициент парной корреляции будет близок к нулю.

Замечание: Такая чудесная характеристика обладает двумя недостатками

1) z_{xy} является мерой линейной связи между двумя величинами, никто не обещал, что соответствующая связь между степенью одной и другой вершины должна быть линейной. Здравый смысл подсказывает, что связь должна быть не линейной и сложной. Но это не главная проблема. С линейностью связи можно побороться, заменив коэффициент парной корреляции на коэффициенты, предназначенные для статистической оценки нелинейных связей (коэффициент корреляционного отношения и т. д.).

К сожалению оказалось, что все эти коэффициенты не робасны. Нужно придумать что-то другое.

Более робасной оказалась следующая величина:

Рассмотрим i-ую вершину сети. Пусть $V(i)$ — множество соседних вершин, тогда

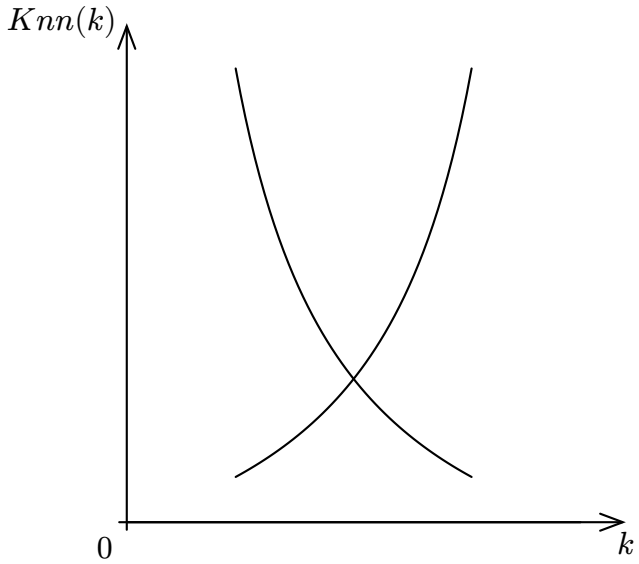
$$Knn(i,k) = \frac{1}{k_i} \sum_{j \in V(i), k_j = k} k_j$$

Здесь k_i — степень вершины i, k_j – степени соседних с ней вершин, но суммирование идет не по всем соседним вершинам, но только по тем, которые имеют степень k. nn — nearest neighbours. Тогда получаем следующее:

$$Knn(k) = \frac{1}{N} \sum_{i=1}^{n} Knn(i,k),$$

где суммирование идет по всем вершинам сети.

Эта величина позволяет построить график



На котором ассортативные сети дадут возрастающую функцию (в идеале монотонную), а дисассортативные дадут убывающую функцию (в идеале монотонную).

Другим эффективным подходом к установлению ассортативности / дисассортативности сети является Коэффициент Клуба Богатых.

Обозначим через $N_{>k}$ число вершин, степень которых превышает k. А через $E_{>k}$ число ребер, соединяющих две вершины, каждая из которых превышает k. Тогда:

$$\varphi(k) = \frac{E_{>k}}{N_{>k}(N_{>k} - 1)\ /2}$$

В чистом виде такой характеристики оказывается недостаточно, и обычно используют нормированную величину:

$$\rho(k) = \frac{\varphi(k)}{\varphi_0(k)}$$

где $\varphi_0(k)$ – это коэффициент клуба богатых для случайного графа.

Определение степенных распределений в сложных сетях.

Мы говорим, что базовой характеристикой, отличающей сложные сети от других типов графов, является то, что всевозможные распределения характеристик являются степенными функциями распределения. Соответственно, для практической работы со сложными сетями необходимо уметь отличать степенные распределения от других распределений.

Лекция 4. Идентификация степенных распределений.

Базовой задачей, при установлении того факта, что граф, с которым мы имеем дело является сложной сетью, является задача идентификация степенных распределений, а именно установление того «простого» факта, что данная выборка порождена степенным распределением. Математически такая задача является задачей математической статистики. Но работа со степенными распределениями не входит в стандартный курс.

Мы имеем выборку, то есть набор н.о.р.с.в $(\xi_1, ..., \xi_n)_{iid}$. Мы можем ставить в отношение этой выборки два вопроса:

1) Мы предполагаем, что выборка этих величин порождена неким конкретным распределением, класс которого нам известен. Например, это выборка из нормального распределения. Но мы не знаем параметры этого распределения. Мы хотим проверить гипотезу о параметрах этого распределения.

$$N(a, \sigma^2); H_0 : a = a_0$$

Другой, более важный для нас вопрос:

2) В реальных сложных системах мы обычно не знаем класс распределения, которым порождена наша выборка. Соответственно, второй вопрос, к какому классу распределений принадлежит распределение, породившее нашу выборку.

Мы можем «попытаться» проверить статистическую гипотезу о том, что распределение, породившее выборку, это некое конкретное распределение, за этой выборкой стоит некий конкретный вероятностный закон.

$$H_0 : F = F_0$$

При этом мы не ограничиваем себя каким-то конкретным классом распределений. Любая функция, удовлетворяющая требованиям функции распределения.

В первом случае говорят о параметрической статистике, потому что выдвигаемые гипотезы касаются параметров распределения, класс распределения мы знаем.

Во втором случае говорят о непараметрической статистике, потому что выдвигаемые гипотезы касаются распределения как такового. Иногда употребляют англоязычный термин goodness-of-fit test, проверка гипотезы, насколько данная выборка соответствует данному распределению.

При работе со степенными распределениями является первым и более важным является ответ на второй вопрос. Должны ли мы работать с этой выборкой как с выборкой из степенного распределения, должны ли мы предполагать, что мы можем каким-то образом оценить параметры степенного распределения, исходя из того, что мы действительно имеем дело со степенным распределением

Оказалось, что даже классических методов непараметрической статистики недостаточно. Сколько нибудь эффективные методы работы со степенными распределениями появились в последние 15 лет, поэтому, они обычно не входят в классический курс математической статистики.

Вспомним, что такое степенное распределение:

$$P(x) = C \cdot x^{-\alpha}$$

Где α является параметром распределения, а C - константой нормализации, гарантирующей нам, что:

$$\int_{-\infty}^{+\infty} P(X)\ dx = 1$$

Соответственно, когда мы говорим о оценке параметров степенного распределения, мы на самом деле оцениваем один параметр – α, а C просто определяется из условия.

Какие подходы мы можем указать для решения задачи идентификации, является ли наша выборка выборкой из степенного распределения?

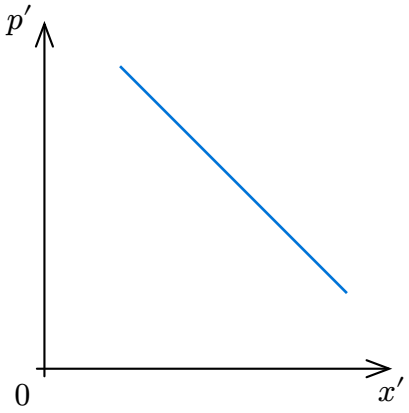
Первый метод носит название Метод Хилла (Hill), он базируется на переходе к двойному логарифмическому масштабу. Если мы прологарифмируем выражение плотности для степенного распределения, мы получим:

$$P(x) = C \cdot x^{-\alpha}$$

$$\underbrace{\ln P}_{p'} = \underbrace{\ln C}_{C'} - \alpha \underbrace{\ln x}_{x'}$$

$$p' = C' - \alpha x'$$

Соответственно, если мы нарисуем в этих новых координатах нашу зависимость, то это должен быть отрезок прямой с отрицательным коэффициентом наклона, в том случае, если верна наша нулевая гипотеза о том, что мы имеем дело со степенным распределением.



Более формально, можно предложить следующее развитие метода Хилла: давайте оценим параметры C' и α с помощью МНК или метода максимального правдоподобия.

МНК: из выборки имеем $p'_1, ..., p'_n, x'_1, ..., x'_n$, тогда давайте посчитаем минимум следующей функции:

$$\frac{1}{n}\sum_{i=0}^{n}(p'_i - C' - \alpha x'_i)^2 \to \min$$

где n – размеры выборки. Отсюда, дифференцируя по C' и α (p'_i и x'_i нам известны) мы находим выражения для оценки C' и α, которые минимизируют это выражение. Если окажется, что полученная таким образом оценка (оценка методом наименьших квадратов) действительно делает этот квадратичный функционал малым, то это означает, что, во-первых, мы нашли хорошие оценки этих двух параметров, а во-вторых, что у нас действительно имеет место степенной закон распределения.

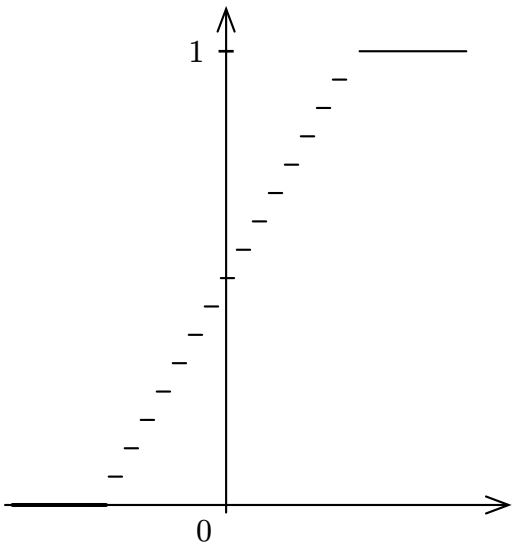
Следующий подход принадлежит трем американским математикам Clauset, Shalizi, Newman, которые, в прочем, опирались на работы двух российских математиков Колмогорова и Смирнова. Так называемая KS-статистика.

По выборке н.о.р.с.в $(\xi_1, ..., \xi_n), \xi_i \sim F(x)$, мы можем построить так называемую

Эмпирическую функцию распределения. В одномерном случае алгоритм построения эмпирической функции распределения $\hat{F}_n(x)$ выглядит просто:

Мы сортируем выборку $(\xi_1, ..., \xi_n)$, по возрастанию, и получаем из нее так называемый вариационный ряд $(\xi_1^*, ..., \xi_n^*)$

$$\hat{F}_n(x) = \begin{cases} 0, & x < \xi_1^* \\ \frac{l}{n}, & x = \xi_i^* \\ 1, & x > \xi_n^* \end{cases}$$



Функция $\hat{F}_n(x) = 0$ для всех $x < \xi_1^*$, $\hat{F}_n(x) = 1$ для всех $x > \xi_n^*$, и в каждой точке ξ_j^* она совершает скачок на величину $\frac{1}{n}$, если существует только одно значение ξ_j^* (нет равных ей) и скачок на $\frac{l}{n}$, если в вариационном ряде встречается l одинаковых значений ξ_j^*.

Определенная функция является функцией распределения по определению. через нее мы проведем истинную функцию, которую мы аппроксимировали такой эмпирической функцией. Чем больше выборка, тем точнее такая аппроксимация будет приближаться к истинной функции. Про эмпирическую функцию распределения было доказано два предельно мощных утверждения.

Первое утверждение носит название теорема Гливенто-Кантелли.

При увеличении выборки до бесконечности, случайная величина $F_n(x)$ сходится по вероятности к $F(x)$

$$p \lim_{n \to \infty} F_n(x) = F(x)$$

Второе утверждение носит название теоремы Колмогорова. Если мы рассмотрим статистику KS вида

$$KS = \sqrt{n} \sup_{-\infty < x < +\infty} |F(x) - \hat{F}_n(x)|$$

То полученная случайная величина будет иметь одно и то же распределение для всех функций $F(x)$, так называемое распределение Колмогорова.

На основании этой теоремы Колмогорова и его ученика Смирнова был сформулирован, пожалуй, первый критерий в непараметрической статистике.

Мы выдвигаем нулевую гипотезу, что F – конкретно заданная функция F_0:

$$H_0 : F = F_0$$

Тогда, если наша гипотеза верна, то F_0 и $\hat{F}_n$, восстановленная по выборке, должны мало отклоняться друг от друга. Причем, мы можем оценить степень этой малости, а именно мы должны сравнить KS с квантилем распределения Колмогорова.

$$KS \;\; ? \;\; K_{\alpha;n}$$

Где α – уровень значимости, n – количество степеней свободы.

Если эта величина действительно мала ($KS < K_{\alpha;n}$), то мы не отклоняем нулевую гипотезу. В противоположном случае мы отклоняем нулевую гипотезу и принимаем альтернативную.

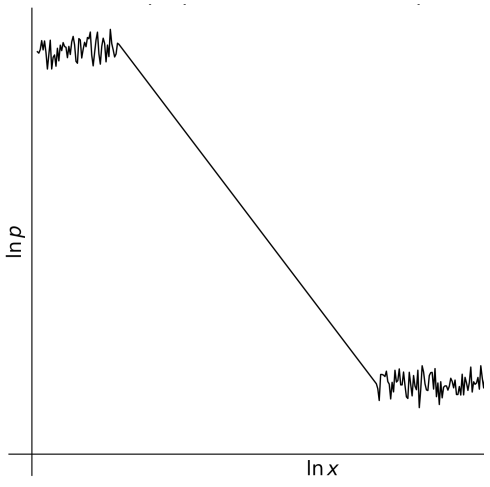
Каковы же недостатки критерия KS? Для того, чтобы применять критерий KS к выборкам в реальных задачах мы должны знать точное значение параметра α, а для того, чтобы сколько-нибудь адекватно оценить параметр α, мы должны быть уверены, что выборка, с которой мы имеем дело, порождена степенным распределением. На практике мы получаем логический круг. Но Clauset, Shalizi и Newman придумали способ, как из него выбраться.

Они предложили следующую идею:

1) На практике, степенное распределение в чистом виде встречается редко. Обычно реальные распределения имеют вид

$$P(X) = \begin{cases} ?, & x < x_{\min} \\ C \cdot x^{-\alpha}, & x_{\min} \le x \le x_{\max} \\ ?, & x > x_{\max} \end{cases}$$

В двойном логарифмическом масштабе это выглядит как

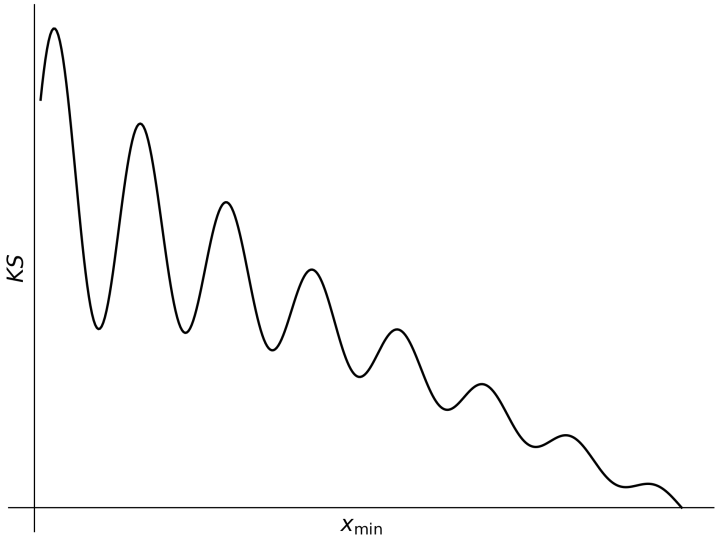


Физически это связано с тем, что сложные системы имеют некие характерные масштабы, для которых и выполняются законы поведения сложных сетей.

Давайте возьмем $x_{\min}$ в некотором разумном диапазоне значений с некоторым разумным шагом. Для каждого конкретного значения $x_{\min}$ с помощью метода Хилла, либо любого другого метода параметрической статистики, мы оценим значение α, удержав в вариационном ряде только те значения, которые больше текущего $x_{\min}$. Получив оценку для α, мы тем самым в точности специфицируем функцию F_0 из критерия Колмогорова Смирнова.

$$F_0 = C \cdot X^{-\hat{\alpha}_{\text{Hill}}}$$

Тогда мы можем для такой функции вычислить значение KS статистики, для функции от $\hat{\alpha}$ и в конечном случае от $x_{\min}$. Нарисуем график такой зависимости:



Он будет убывающим по очевидной причине – чем меньше у нас выборка, тем точнее мы можем приблизить нашу функцию данным распределением. Он будет убывающим не монотонно, тогда в качестве значения $x_{\min}$ выбирается значение, при котором функция $KS(x_{\min})$ достигает своего первого локального минимума. Это некая общая идея, иногда выбирается не первый локальный минимум, а второй или третий, что обычно связано с той ситуацией, что первый локальный минимум является некоторой флуктуацией на убывающем участке такой зависимости.

Тем самым мы получаем оптимальные в некотором смысле оценки $\hat{x}_{\min}, \hat{\alpha}$.

К сожалению, этого оказалось недостаточно для проведения goodness-of-fit теста для степенного распределения.

Замечание: для $x_{\max}$ процедура аналогичная, но мы двигаемся справа налево. Однако, обычно $x_{\max}$ мало влияет на результат, а вот $x_{\min}$ может быть критичен.

Clauset, Shalizi и Newman предложили решение данной проблемы проверки на степенность. Они предложили на ряду с исходной выборкой рассмотреть еще значительное число синтетических выборок, а именно синтетических выборок, порожденных законом распределения:

$$C \cdot X^{-\hat{\alpha}}, x > \hat{x}_{\min}$$

как раз с теми оценками, которые мы получили движением по графику KS от $x_{\min}$.

Мы берем некоторое степенное распределение, и порождаем выборки. У нас есть много выборок. Для каждой выборки мы считаем KS, в том числе для исходной (обрезав $x_{\min}$). В подавляющем большинстве случаев, значение KS для нашей выборки будет больше, чем для синтетических выборок, которые мы сделали таким образом, чтобы они были максимально близки к нашему реальному распределению.

KS статистика показывает, насколько истинное распределение близко к нашей эмпирической функции распределения.

Если исходная выборка не является степенной, например на самом деле она является нормальной, то KS статистика для будет иметь гигантское значение, намного большее, чем значение KS статистики для синтетических выборок, так как мы будем пытаться сравнивать что-то восстановленное по нормальному распределению с истинно степенным распределением.

Мы замерим процент случаев p, для которых:

$$KS_0 < \min_j KS_j$$

где KS_0 – KS статистика для нашей выборки, KS_j – для синтетических. Обычно берут порядка тысячи ($j < 1000$).

Если верна нулевая гипотеза о том, что F – степенное распределение, что за нашей наблюдаемой выборкой стоит степенное распределение, то $p > 10\%$.

Этот полуэмпирический критерий служит способом проверки выборки на степенность. Если $p < 10\%$, то есть, p близко к нулю, то мы имеем дело не со степенным распределением.

В рамках одного метода первое достоинство – мы умудряемся и оценить параметр распределения, и ответить на вопрос, действительно ли распределение является степенным. Второе достоинство – метод работает. Недостатки метода: нет теоретического обоснования границы в 10%, все остальное теоретически обоснованно.

Второй недостаток: время-емкий процесс.

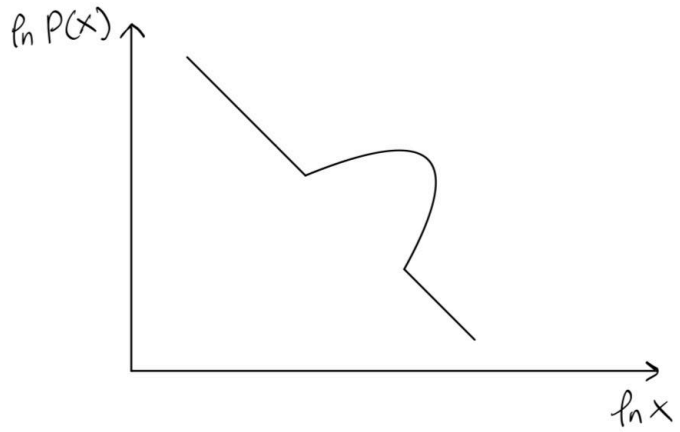
Лекция 5. Метод Прюснера, Mind-Brain problem.

Метод Прюснера (Pruessner) (self-organized criticality) заключается в том, что при специально выбранных координатах различные выборки, порожденные одним и тем же распределением, дают один и тот же вид распределения. Графически на экране монитора эти зависимости совпадают друг с другом, и мы получаем явление, которое в статистике называется Collapse данных. Соответственно, этот метод выходят далеко за пределы анализа степенных распределений, но Прюснер рассматривал конкретно Collapse данных степенных распределений.

Более того, Прюснер работал с распределениями вида

$$P(x) = \begin{cases} .. , & x < x_{\min} \\ C \cdot X^{-\alpha} g(x/x_c), & x \geq x_{\min} \end{cases}$$

где C – константа нормализации, $x_{\min}$ и α – нижнее отсечение и параметр степенного распределения, x_c – характерный размер элементов выборки (максимальное значение элемента в выборке), $g(x/x_c)$ – функция горба (hunch) и присутствует в большинстве реальных степенных распределений. Если вы возьмете реально степенное распределение и запишите его в двойном логарифмическом масштабе, вы получите что-то вроде:



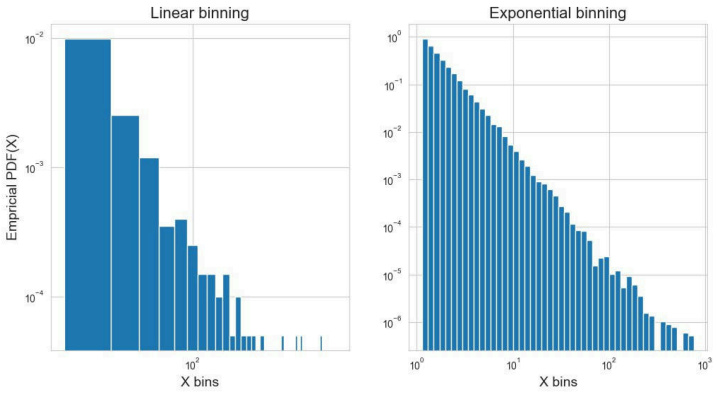
Мы получаем характерный горб, математически его существование вызвано наличием функции $g(x/x_c)$, а физически его существование связано с тем, что степенное распределение все-таки предполагает, что у нас бесконечность возможных значений, и невероятные события все-таки существуют. Так как выборка конечно, мы этот бесконечный интервал ужимаем до $x_{\min}, x_{\max}$.

Метод Прюснера лучше прошлых методов, так как может работать с горбами.

Метод Прюснера базируется на двух китах:

1-ый кит) Экспоненциальный биннинг - когда мы говорим о биннинге распределений данных, мы привыкли к равномерному биннингу, все бины имеют одинаковую длину. Для степенных распределений это не самая лучшая идея, для них лучше применять бины не одинаковой длины, а именно бины, длина которых растет экспоненциально с увеличением значения X.

Это достаточно понятно, потому что у степенного распределения большая часть данных будет сконентрирована в окресности максимума, но какой-то объем данных, в силу того, что хвост у него тяжелый, будет сконцентрирован дальше.

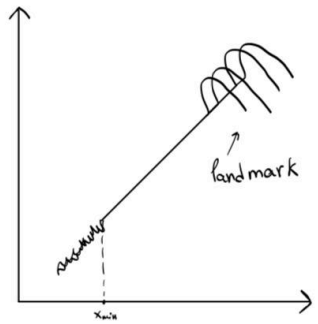


2 кит) Collapse данных. Допустим мы угадали $\mathrm{x}_{\max}, \alpha$ и x_c, то для различных выборок, порожденных одним и тем же распределением, наши зависимости в координатах $p' = P(X) \cdot x^{-\alpha}, x/x_c = x'$ будут иметь один и тот же вид.

$$p' = C \cdot g(x')$$

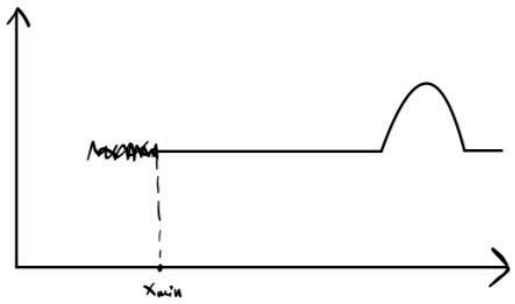
Прюснер работал с модельными данными, то есть с выборками, порожденными моделями самоорганизованных критичных систем (самыми сложными из сложных систем). В реальных ситуациях, когда у нас есть всего одна выборка, то, если она достаточно большая, мы берем подвыборки из нее (случайно выбранные, 95%, 90%, 80%), и на этих подвыборках моделируем ситуацию с многими выборками, распределениями. При этом надо ясно понимать, что для каждой из этих выбранных выборок x_c будет разным, что обеспечивает нам Collapse данных.

Мы берем разные выборки (экспоненциально отбиннингованные), и оцениваем (одним из предыдущих методов) параметр α. Мы будем обозначать эту оценку α', она является очень примерной. Поэтому мы строим наши зависимости в наших координатах p', x'. Если нулевая гипотеза о том, что мы действительно имеем дело со степенным распределением верна, то мы получаем следующую картинку:



У нас есть участок ниже $x_{\min}$ с шумом, у каждой выборки он свой, у нас есть наклонный участок прямой, который отвечает $C \cdot x^{-\alpha}$, и у нас есть характерный изгиб, который Плюснер назвал Landmark. В силу того, что выборки у нас разные, мы получаем разные положения всех трех участков (они сдвинуты друг на друга). С другой стороны, по-скольку α' получаена с помощью какой-то грубой оценки, то мы имеем дело с наклонным участком прямой, но то, что этот участок есть, является первый признак того, что распределение, все же, степенное.

Третий шаг: мы должны оценить истинное значение α и x_c. Если нам это удастся, то вместо множества графиков, как на предыдущем рисунке, мы получим один график характерного вида, иными словами все графики коллапсируют в один.



В этом графике будет участок меньше $x_{\min}$, далее горизонтальный участок прямой, указывающий на то, что расрпеделение является степенным, и характерный landmark, отвечающий функции горба g. Для осуществления коллапса, перехода от

верхнему графику к нижниму, нужно совершить две операции: повернуть график таким образом, чтобы график стал горизонтальным, тем самым получая истинный α, и выбрать характерный масштаб таким образом, чтобы все горбы совпали друг с другом (они действительно совпадут).

Это можно делать вручную, но Прюснер рекомендует использовать МНК, где в качестве данных выступают положения максимумов горбов и значения в этих максимумах.

Достоинства метода:

Метод работает с реалистичными степенными распределениями, включающими функцию g.

Метод позволяет оценить не просто параметр такого степенного распределения $(x_{\min}, \alpha, x_c)$, но и проверить нулевую гипотезу о том, что распределение действительно является степенным. Появление этого горизонтального участка прямой является критерием проверки.

Метод позволяет провести goodness-of-fit test, что дорогого стоит.

Недостатки:

Поскольку он базируется на разных подвыборках, он требует весьма больших выборок, что не всегда возможно в реальных задачах.

Замечание: теория работы со степенными распределениями является развивающийся областью статистики, и в общем-то задача проверки распределения на степенность является открытой задачей. При практическом использовании целесообразно использовать несколько методов проверки распределения на степенность и делать выводы о степенности, если все три метода дадут положительный ответ с более или менее одинаковыми значениями $x_{\min}, x_{\max}, \alpha$.

Конетком и Когнитом – это понятия, введенные в теорию сильного интеллекта академиком Константином Анохиным.

Конетком – совокупность нейронов головного мозга человека вместе со совокупностью их аксонно дендридных связей, носят название конеткома (не трудно догадаться, что это некая сложная сеть со всеми особенностями, присущими сложным сетям, с которыми мы уже знакомы)

Аксонно дендридные связи – каждый нейрон человека состоит из тела (сомо) и длинного хвоста (аксона), который ведет к другим нейронам. сигналы в головном мозге передаются от одного нейрона к другому через такие аксоно дендритные связи. между ними есть синноптическая щель, которая заполнена нейромедиаторами, сила связи между двумя нейронами определяется концентрацией нейромедиаторов в этой щели. Радость – повышение нейромедиаторов в щелях.

Реальное обучение в реальном головном мозге это не изменение концентрации нейромедиаторов, это изменение самой структуры. Все эмоции – отмирание нейронов и построение новых связей.

Когнитом – мы можем в том или ином смысле исследовать конетком, но внутреннему наблюдению нам доступны только ментальные состояния человека. Эти состояния также образуют сложную сеть, которая носит название когнитома, и уже принадлежит области психического (не материального).

Здесь мы сталкиваемся с вечной проблемой нейро-физиологии, которая носит название Mind-Brain problem. Проблема заключается в том, что взаимодействие вполне реальных вещей не вполне понятным нам образом порождает психические явления, то есть явления, относящиеся к сфере духовного.

Скажем одно, гипотеза Анохина заключается в том, что когнитом представляет собой совокупность когитов, то есть, некоторых временно возникающих совокупностей нейронов головного мозга. Ансамблей нейронов. С математической точки зрения это приводит к понятию сложной гипер-сети, то есть, сложной сети (графа) вершинами которой являются другие сложные сети. ее верхний уровень - когнитом (ментальные состояния), ее самый нижний уровень – конетком (просто связи между нейронами), но есть промежуточные уровни, которых от 1 до 4, точно неизвестно.

В целом, идея конеткома когнитома приводит нас к другому нейро-физиологическому вопросу, вопросу пространственной локализации. Существуют две противоположных точки зрения. Согласно первой точки зрения, высшие когнитивные функции человека локализованы в конкретных участках головного мозга (напри-

мер, зрительная кора, отвечающая за распознование образов, зоны, отвечающие за лексику и грамматику)

Вторая позиция – мозг это единое целое, все связано со всем, но в этом едином целом возникают объекты, которые никак геометрически не связаны с зонами, то есть, расположены в разных местах.

В целом, при математическом описании геометрии в сложных системах, и при ответе на майнд-брейн-проблем, мы можем использовать следующие методы: сложные сети как таковые, при этом обычно применяются Community-detection алгоритмы, второй подход: гиперсети (гиперграфы), третий подход: так называемые симплициальные компле'ксы – базовое понятие топологии, но в том варианте, в котором оно нам нужно, нам его понять достаточно легко. Мы знаем, что такое граф – G(V, E), пусть кроме ребер мы стали рассматривать вершины более высокого уровня, например грани, то есть мы начинаем рассматривать структуру, которая представляет собой $E \times E \times E$, это называется симплициальный комплекс.

Последний подход: Графоны – континуальное обобщение графов.

К критической самоорганизованной системе относится естественный язык и человеческий мозг, соотственно сильный искуственный интеллект тоже будет таковой. К самоорганизованной критической системе относятся системы, которые удволетворяют трем следующим требованиям:

1) Они состоят из гигантского числа взаимодействующих элементов, причем правила взаимодействия между этими элементами сравнительно простые.

2) В этих системах должны возникать так называемые лавины, когда активация одного элемента влечет активацию второго элемента, второго влечет активацию третьего и так далее. Лавины захватывают существенную часть системы, сопоставимую с ее размерами, или даже систему целиком.

3) Размеры лавин подчиняются степенным законам распределениям.

Если это так, то система является самоорганизованной критичной системой. Что же касается языков, то здесь элементарными элементами называются либо сами люди, либо семы – некий элементарный элемент когнитивного пространства.

Лавина – любой текст, произнесенный или сказанный представляет собой лавину в языке.

Лекция 6. Сложность

Самый важный тег для нас – сложность. Если мы говорим о сильном искусственном интеллекте, то мы должны говорить о сложности.

Понятие сложности в математике может определяться по-разному.

Существует три центральных подхода:

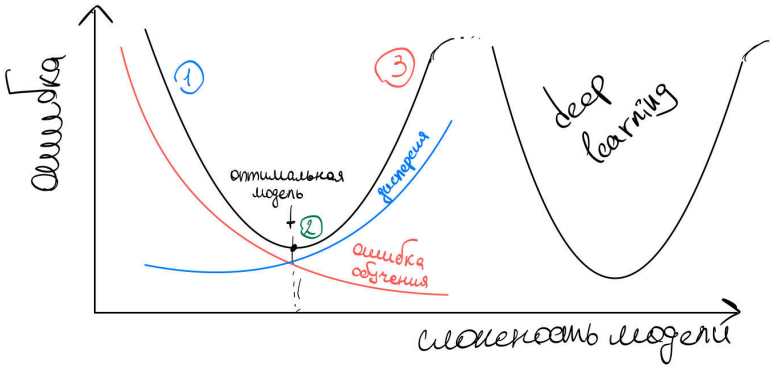
1) Сложно то, что сложно сгенерировать (Колмогоровская сложность). Если для данного массива данных мы должны построить такую машину Тьюринга, что (какая-то цель) – определение сложности. Практически слабо применимо.

2) Сложно то, что сложно предсказать.(Предсказательная сложность). Базовая идея, что если мы не можем предсказать поведение системы, то она сложна.

3) Сложно то, что сложно выучить. (Размерность Вафника-Червоненкиса, VC).

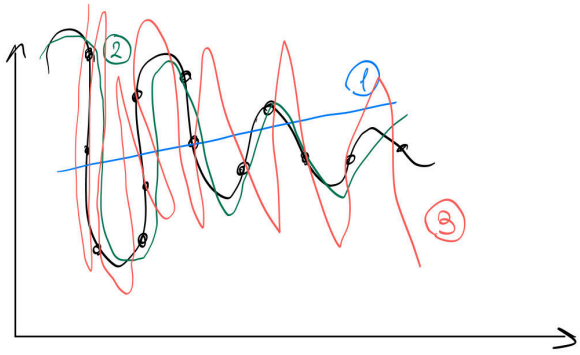
Эти все рассуждения о сложности не являются чисто теоретически. Золотое правило интеллекта: сложность обрабатываемых данных должна равняться сложности системы, которая их обрабатывает. Для того, чтобы оно работало, необходимо хоть как-то определить сложность.

Когда речь идет об обычных нейронных сетях, мы видим следующую картину:



Если вы увеличиваете сложность, то до какого-то момента ошибка обобщения падает (ошибка обобщения – оишбка на тестовой выборке, ошибка обучения – ошибка на тренировочной выборке), есть некий оптимум сложности, который наилучшим образом соответствует данным, с которыми мы работаем. Если мы продолжаем увеличивать сложность, мы получаем возростающий участок. Второй кусочек связан с революцией в машинном обучении в связи с появлением глубокого обучения. Но с чем это связано теоретически пока никто не понимает. Возможно это связано с теорией самоорганизации, о которой мы поговорим позже. Как интерпретировать восходящие и снисходящие участки?

Предположим, что у нас есть какая-то истинная функция:



Мы не имеем всей функции, у нас есть выборка, снятая с этой функции. Допустим, сложность системы проще сложности данных, нашей системе просто не хватит ресурсов, чтобы более менее точно соответствовать системе (функция 1). Из-за чего получаем большую ошибку.

Если мы попадаем куда-то в окрестность оптимума, то мы сможем более менее точно пройтись по кривой и более менее точно ей соответствовать. (функция 2)

Переобучение – контринтуитивное свойство, в данном случае мы получаем так называемые wild-functions, они достаточно хорошо проходят через каждый элемент выборки, но засчет большого количество степеней свободы мы получаем такой забор (функция 3). Такая функция не имеет никакого отношения к истинной функции, поэтому ошибка обобщения становится гигантской.

Прежде чем переходить к предсказательной сложности (predictive_complexity), стоит поговорить про информацию и информационную энтропию.

Шэноновская информация. Давайте рассмотрим два вероятностных события:

Событие А: Профессор N вошел в кабинет.

Событие Б: Профессор N убил студента.

Какое событие несет больше информации? Почему?

Вероятность второго события существенно ниже. После этого Шэнон сказал:

1) $I = f\left(\frac{1}{p}\right)$

Где I - информация, p — вероятность.

2) Пусть есть два независимых события, тогда $I = I_1 + I_2, f\left(\frac{1}{p_1}\frac{1}{p_2}\right) = f\left(\frac{1}{p_1}\right)f\left(\frac{1}{p_2}\right)$

Функция логарифма удовлетворяет такому свойству. С другой стороны можно доказать, что единственная функция, которая удовлетворяет такому свойству – логарифм. Тогда $I = \ln\frac{1}{p} = -\ln p$

Предположим, что у нас есть система, которая может находиться в N состояниях. И в этих состояниях она находится с вероятностями $p_1, ..., p_n$. Тогда энтропия по Шэнону есть средняя информация, которую мы знаем о такой системе.

$$H = <I> = \sum_{i=1}^{N} p_i I_i = -\sum_{i=1}^{N} p_i \ln p_i$$

Что это означает? Свойства:

Предположим, что мы знаем о системе все, тогда мы можем утверждать, что с вероятностью $p_i = 1$ система находится в определенном состоянии, остальные вероятности равны нулю. Тогда энтропия равна нулю. Значит если мы все знаем о системе, то ее информационная энтропия равна нулю.

В обратной ситуации, когда мы не знаем ничего о системе, то p_i равновероятны.

$$p_i = \frac{1}{N} \Rightarrow H = \ln N \Rightarrow \ln N = \max(H)$$

Из этих двух свойств мы делаем вывод: информационная энтропия – мера неопределенности системы. Допустим мы получили какую-то информацию, тогда энтропия упадет, и мера полученной информации равна этой разнице.

Как генерировать хаотические ряды? $X_{n+1} = 1 - \lambda X_n^2$ – логистическая зависимость. Это мы обсуждать не будем, просто пример ряда, который генерирует информацию с каждой итерацией. И за небольшой период времени он сгенерирует тонну информации, хотя по сути, никакой смысловой нагрузки она не имеет. В этом заключается одна из проблем шэноновской энтропии.

Введенная таким образом Шэновская энтропия является аддитивной в силу выполнения свойства 2 (про независимые события), более того, можно показать, что она является единственной аддитивной энтропией. Между тем в сложных системах обычно имеют место так называемые не аддитивные энтропии, для которых свойство 2 не выполняется по банальной причине: считается, что не существует двух подлинно информационно изолированнох подсистем.

Поэтому в сложных системах целесообразно применять не аддитивные энтропии, то есть энтропии, для которых неверно $I = I_1 + I_2$

Этих энтропий бывает много, но основые: энтропия Реньи(Renyi)

$$H_q = \frac{1}{1-q} \ln \sum_{i=1}^{n} p_i^q$$

Энтропия Цависа(Tsalis), энтропия Каниадакиса(Kaniadakis), энтропия Шарма-Митталя и так далее.. Открытая проблема, к какой сложной системе какую не аддитивную энтропию применять.

Общее методическое замечание: Сложные системы делятся на классы универсальности, классы смежности. Сложным системам, пренадлежащим одному классу, свойственны некие универсальные свойства.

Предсказательная сложность. Предположим, что у нас идет некоторый поток данных в неком времени, не обязательно метрологическом, и мы можем выделить в нем некоторый момент времени $t = 0$ и две части данных: прошлое и будущее. Тогда данные, относящиеся к прошлому мы будет обозначать как X_{past}, а к будущему X_{future}, совместную вероятность прошлого обозначим за $P\big(X_{past}\big)$, совместную вероятность будущего за $P\big(X_{future}\big)$, тогда взаимной информацией между будущем и прошлым будет:

$$I\big(X_{future}, X_{past}\big) = < \log_2 \frac{P\big(X_{future} \mid X_{past}\big)}{P\big(X_{future}\big)} >$$

Где усреднение производится по совместному распределению $P\big(X_{future}, X_{past}\big)$.

Эта величина носит название Девергенция Кульбаха Лейблера (Kulback Leibler). Если мы хотим сравнить между собой, например, две матрицы, мы посчитаем какую-нибудь норму или какое-нибудь расстояние. Это наше действие по умолчанию. А если мы хотим сравнить два вероятностных распределения, то действием по умолчанию является сравнение по Девергенции Кульбаха Лейблера.

По сути мы сравниваем информацию о прошлом (которая позволяет нам говорить о будущем) с $P\big(X_{future}\big)$. То есть, действительно предсказательную сложность.

Вот математическая мера того, насколько нам сложно предсказать будущее по прошлому. Если мы умножим числитель и знаменатель на $P\big(X_{past}\big)$, то получим

$$I\big(X_{future}, X_{past}\big) = < \log_2 \frac{P\big(X_{future} \mid X_{past}\big)}{P\big(X_{future}\big)} > =$$

$$= - < \log_2 P\big(X_{past}\big) > - < \log_2 P\big(X_{future}\big) > - \ \big[- < \log_2 P\big(X_{future}, X_{past}\big) > \big]$$

Пусть $past = T'$, $future = T$; тогда по определению нами рассмотренной энтропии:

$$= H(T') + H(T) - H(T + T')$$

Таким образом мы свели понятие взаимной информации к некой комбинации информационных энтропий.

Замечание 1:

Выражение, введенное для совместной информации, введеное таким образом, является симметричным. В этом смысле равносложными являются задачи как предсказания, так и подсказания, когда мы прошлое определяем по будущему.

В формуле H всегда есть линейная состовляющая и еще одна одна величина.

$$H(T) = H_0 \cdot T + \underbrace{H_S(T)}_{o(T)}$$

Первая, линейная, называется экстенсивной энтропией, она связана с течением времени. То есть, любая система подчиняется второму началу термодинамики. В этом смысле энтропия растет всегда, причем растет пропорционально протекшему времени. Это имеет место абсолютно для любой системы. И это в общем то неинтересно.

Вторая – субэкстенсивная энтропия, которая ведет себя как $o(T)$ и определяет поведение системы, которая отличает ее от всех остальных.

Мы постараемся показать, что разные системы дают нам принципиально различные функции суб-экстенсивной энтропии как функции времени. Тип этой функции и дает нам класс сложности соответствующей системы. Более того, типов этих функций существует не так уж и много. И это дает нам некоторый первый подход к классификации сложных систем. Пока же нам достаточно отметить, что экстенсивые состовляющие энтропии сокращаются. Соответственно, совместная информация определяется только суб-экстенсивной состовляющей энтропии.

$$I\big(X_{future}, X_{past}\big) = H_s(T') + H_s(T) - H_s(T + T')$$

Предположим, что мы знаем о прошлом все, тогда мы можем устремить T' к бесконечности:

$$\lim_{T' \to +\infty} I(T \mid T') = H_S(T) = I_{pred}(T)$$

Тогда мы получим по сути суб-экстенсивную составляющую того, что мы называем будущим. Эта величина и носит название предсказательной сложности.

Нужно обратить внимание на следующий факт: эта величина симметрична относительно обращения прошлого и будущего. Задача предикции и постдикции – одинакова. Можно устремить не T', а T, получится то же самое выражение.

Здесь можно сказать, что, когда мы говорим о такой хорошей, фундоментальной величине, она должна удовлетворять двум, противоречивым требованиям:

1) она следует из базовых принципов

2) ее нужно уметь измерить экспериментально

Метрика хорошо измеряется.

Лекция 7. Плоскость Энтропия-Сложность

Классической постоновкой является постановка следующего вида:

Мы наблюдаем временной ряд $y_0, y_1, ..., y_t, ..., y_t \in \mathbb{R}^S$ (вообще говоря, дальнейшние рассуждения работают для любого s, но для простоты мы будем рассматривать случай $s = 1$) и завершаем его наблюдение в момент времени t. Мы хотим получить прогноз данного временного ряда для момента времени $t+1, t+2, ..., t+K$, где $K -$ некоторая константа, число шагов вперед, на которые мы хотим получить прогноз. Мы хотим получить оценки $\hat{y}_{t+1}, \hat{y}_{t+2}, ..., \hat{y}_{t+k}$ неизвестных нам наблюдений $y_{t+1}, y_{t+2}, ..., y_{t+k}$ таким образом, чтобы

$$\sum_{j=1}^{k} \mathbb{E}\left[y_{t+j} - \hat{y}_{t+j}\right]^2 \to \min$$

И абсолютно все методы, от простейшего линейного МНК до сверхсложных алгоритмов прогнозирования на основе кластеризации и нейросетевых моделей, укладываются в эту постановку.

Проблема заключается в том, что все возможные ряды укладываются в две большие категории: регулярные ряды и ряды хаотические. Базовым свойством, отличающим хаотические ряды от регулярных, является наличие горизонта прогнозирования. То есть, числа шагов вперед, на которые мы в принципе в состоянии спрогнозировать ряд любым из возможных и невозможных методов. Горизонт прогнозирования – это физическое свойство наблюдаемой системы.

Для регулярных рядом горизонт прогнозирования равен $+\infty$, для хаотических рядом горизонт прогнозирования конечен, его можно посчитать по ряду, например, алгоритмом Розенштейна.

Допустим, у нас есть регулярный ряд, например, зашумленная синусоида, понятно, если нам удалось как-то оценить его амплитуду / фазу, допустим, с помощью того же МНК, то, теоретически, мы можем давать прогноз до бесконечности. (пока система вообще существует)

Однако, регулярные ряды в природе не встречаются. Если мы хотим работать с действительно сложными системами, то мы должны учиться работать именно с хаотическими рядами. А в хаотических рядах мы сразу натыкаемся на горизонт прогнозирования. Ошибка прогнозирования здесь растет экспоненциально.

Для регулярных и хаотических рядов существуют принципиально различные методы прогнозирования.

Допустим, перед нами есть ряд. По его графику трудно понять, регулярной он или хаотический. Возникает естественный вопрос, а можем ли мы отличить регулярный ряд от хаотического? Это базовый вопрос теории прогнозируемых рядов. Регулярные ряды – простые, хаотические ряды – сложные. Простые системы, обычные системы искусственного интеллекта порождают регулярные ряды, сильный искусственный интеллект порождает хаотические. Это недостаточный, но необходимый признак.

Плоскость Энтропия-Сложность: рассмотрим наблюдаемую часть временного ряда

$$y_0, y_1, ..., y_t, ...$$

и разобьем его на отрезки длины k. В теории их называют z-вектора, а на жаргоне – чанки (chunks). k — достаточно небольшая величина.

$$z_0 = (y_0, y_1, ..., y_{k-1})$$

$$z_1 = (y_1, y_2, ..., y_k)$$

И так далее. Мы предполагаем, что элементы каждого z-вектора строго не равны друг другу, поэтому относительно двух соседних элементом мы можем сказать, что y_i строго больше / меньше y_{i+1}; тогда мы можем ввести понятие канонического расположения наблюдений в z-векторе, например, будем считать, что $y_i < y_{i+1}$ для всех наблюдений в z-векторе. Мы имеем монотонно возрастающий набор элементов. Мы можем ввести понятие перестановки, которая приводит актуально наблюдаемую последовательность к каноническому виду.

При достаточно большом объеме наблюдаемой части ряда мы можем считать частоту появления перестановки того или иного типа (z-вектора того или иного типа) хорошей мерой, хорошей оценкой его вероятности (ЗБЧ). Тем самым, каждому временному ряду мы можем поставить в соответствие набор вероятностей $p_1, ..., p_n$ появления перестановки того или иного типа.

Авторами этого метода (бразильцы Martin, Plastino, Rosso) было предложено, основываясь на этих вероятностях, посчитать две величины, характеризующие исходный временной ряд. Первая величина – это привичная нам энтропия, но нормированная на ее максимальное значение ($\log m$), затем, чтобы нормированная энтропия лежала в пределах от нуля до единицы.

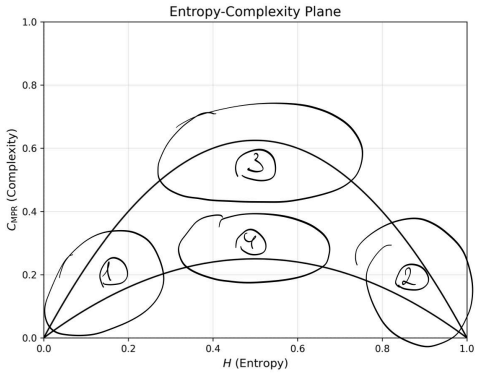
$$0 \leq H \leq 1$$

Одной характеристики оказалось недостаточно. Вторая характеристика носит название сложности, а если быть точным, MPR-сложности (фамилии авторов).

$$C_{\text{MPR}} = Q_0 \cdot H \cdot \|P - P_e\|$$

где P_e — равномерное распределение, то есть: $P_e = \left\{p_j = \frac{1}{N}\right\}$, H — энтропия, Q_0 — нормализирующая константа, которая гарантирует, что $0 \leq C_{\text{MPR}} \leq 1$, $\|P - P_e\|$ показывает, насколько уклоняется актуальное распределение от распределения равномерного.

Далее мы нашему ряду ставим в соответствие два числа, которые лежат в единичном квадрате.



Если ряд относится к простым детерменированным процессам (допустим, синус), то соответствующая точка попадет в левый нижний угол (область 1) нашего геометрического места точек (точка не может попасть ниже нижней «параболы», выше верхней «параболы»).

Если ряд является чисто случайным процессом, то соответствующая точка (пара энтропии и сложности) попадет в правый нижний угол этой фигуры (область 2).

Если же ряд относится к хаотическим рядам, то он попадет в окрестность вершины верхней «перевернутой параболы» (область 3).

Если речь идет о цветных шумах, то речь идет об оставшейся области(область 4).

(Если кто-то хочет заниматься hft, то ваша область будет лежать между областями 3 и 4, то есть, хаотический ряд с ярко выраженным цветным шумом)

(Интуиция + понимание: первое утверждение верно, так как большинство вероятностей будет зануляться, например на примере синуса, вероятность большинства перестановок будет равна нулю, так как синус подвергается четкому закону и многие перестановки никогда не встретятся. таким образом будет очень маленькая энтропия, которая занулит сложность, так как напрямую содержится в ее формуле.

У случайного процесса распределение будет близким к равномерному, тогда норма в сложносте будет близка к нулю, при этом будет высокая энтропия, так как чем ближе мы к равномерному, тем больше у нас степень неопределенности.

Заметим, что у хаотического ряда энтропия равна примерно одной второй, то есть, не все последовательности возможны. Мы говорим, что это детерминиро-

ванный процесс, но часть последовательностей запрещена, причем, порядка 50%. С другой стороны этот процесс максимально сложный. То есть этот процесс максимально удален от равномерного распределения (смотреть на норму).

Все хаотические ряды – только сингулярные распределения. То есть такие распределения, которые не являются ни дискретными, ни равномерными. И все сложные системы имеют именно сингулярные распределения.)

Рассмотрим классическую задачу регриссионного анализа. Мы имеем две выборки: $x_1, ..., x_n$ и $y_1, ..., y_n$, и предполагаем, что выборка x состоит из детерминированных величин (это не всегда так), а y — величины случайные, причем связь между x и y дается соотношением:

$$y = f(x, \alpha) + \varepsilon$$

Где f — некоторая, вообще говоря, нелинейная функция, $\alpha \in \mathbb{R}^S$ — вектор параметров, $\varepsilon = (\varepsilon_1, ..., \varepsilon_n)$ — вектор случайных состовляющих. В таком случае y действительно случайная величина.

Задача – нужно найти такие значения $\alpha = \alpha^*$, что математическое ожидание меры уклонения вектора y от вектора $f(x, \alpha)$ было минимальным в той или иной мере. При этом обычно стремятся к тому, чтобы случайные величины, которые являются оценкой α, были «хорошими» в статистическом смысле. (*Несмещенность, состоятельность, эффективность.*)

$$\frac{1}{n} \sum \mathbb{E}[y_i - f(x_i, \alpha)]^2 \to \min$$

Данная идея хороша, только если ε хорош в статистическом смысле. Иначе полученный результат будет плохим. Для того, чтобы получать хорошие оценки параметра α с помощью интуитивно понятных теоретико-вероятностных методов типа МНК или метода максимального правдаподобия, необходимо, чтобы случайные состовляющие ε удовлетворяли условиям Гаусса-Маркова. Условие заключется в том, чтобы $\varepsilon_1, ..., \varepsilon_n$ были н.о.р.с.в. (iid). Дополнительно к этому условию добавляют условие нормальности, то есть, $\varepsilon_i \sim N(0, \sigma^2)$

Лекция 8. Решение ОДУ. Теория бифуркаций.

Обыкновенные дифференциальные уравнения. Естественным продолжением, и более того, причиной появления того, что называется интегрально-дифференциальное исчисление (математический анализ) является необходимость решения задач, которые описываются обыкновенными дифференциальными уравнениями или уравнениями в частных производных.

Формально, обыкновенные дифференциальные уравнения – это уравнения, которые зависят от неизвестной функции и какого-то числа ее производных.

$$F\big(y(x), y'(x), ..., y^{(n)}(x)\big) = 0$$

Решением обыкновенного дифференциального уравнения являтеся $f(y(x))$

Замечание: по отношению к обыкновенному дифференциальному уравнению мы можем ставить два типа задач:

1) Задача Коши(Cauchy) : предполагает, что все дополнительные условия сосредоточены в одной точке.

$$\begin{cases} y(x_0) = y_0 \\ y'(x_0) = y_0' \\ ... \\ y^{(n)}(x_0) = y_0^{(n)} \end{cases}$$

Где $y_0, y_0', ..., y_0^{(n)}$ – это конкретные числа.

2) Краевая задача : дополнительные условия сосредоточены в нескольких точках. (Простейший пример: двухточечная краевая задача)

$$\begin{cases} y(x_0) = y_0 \\ y'(x_0) = y_0' \\ ... \\ y^{(k)}(x_0) = y_0^{(k)} \\ y^{(k+1)}(x_1) = y_0^{(k+1)} \\ ... \\ y^{(n)}(x_1) = y_0^{(n)} \end{cases}$$

Обычно для начальной задачи x_0 совпадает с левым краем области определения функции $y(x)$, с его началом. Отсюда название начальная задача. Мы знаем что-то в некий начальный момент времени. И хотим понять, как оно развивалось дальше. Это не является догмой. Формальное определение – условие сосредоточено в одной точке.

В случае краевой задачи x_0 совпадает с левым краем промежутка определения $y(x)$, с его началом, а x_1 совпадает с концом, правым краем, отсюда название –

краевая задача. Это тоже не является догмой. Кроме того, в краевой задаче может быть больше точек и больше условий, могут быть более сложные условия.

Пример:

Простейшим и очевидным демографическим законом является то, что число рожденных (в некоторый промежуток) детей пропорционально числу живущих в данной стране людей. Столь же очевидно, что число умерших людей будет пропорционально той же величине, числу живущих в данной стране людей.

Давайте обозначим через $y(t)$ число людей, популяцию, в момент времени t. И попытаемся вывести демографическое уравнение, то есть, получить ответ на вопрос, как будет меняться количество людей с течением времени.

За некий период времени Δt изменение количества людей Δy равно:

$$\Delta y = y(t+\Delta t) - y(t) = \big[K_p \cdot y(t) - K_y \cdot y(t)\big]\Delta t$$

$$\frac{y(t+\Delta t) - y(t)}{\Delta t} = \big(K_p - K_y\big)y(t)$$

Пусть $K = K_p - K_y$. Применим к обоим частям операцию предельного перехода:

$$\lim_{\Delta t \to 0} \frac{y(t+\Delta t) - y(t)}{\Delta t} = \lim_{\Delta t \to 0} Ky(t)$$

$$\dot{y}(t) = Ky(t)$$

K_p – коэфициент рождаемости, K_y – коэффициент смертности, $\dot{y}$ – производная по времени. Получили обыкновенное дифференциальное уравнение.

Кроме того, к этому ОДУ мы можем сформулировать задачу Коши:

$$y(t_0) = y_0$$

В некий момент t_0 число людей в стране – y_0.

Одними и теми же уравнениями мы можем описывать принципиально разные процессы. Пусть $y(t)$ – не население страны, а масса радиактивного вещества. Тогда можем применить теорему полураспада – объем распадающегося радиактивного вещества пропорционален объему радиактивного вещества. Получаем то же самое уравнение без K_p, так как ничего не пребывает, только убывает.

Решим полученное уравнение коэфициента рождаемости:

$$\frac{dy}{dt} = Ky \quad \Rightarrow \quad \frac{dy}{y} = Kdt$$

Применим к обеим частям равенства операцию интегрирования.

$$\int \frac{dy}{y} = \int Kdt$$

$$\ln|y(t)| + C_1 = Kt + C_2$$

Мы имеем две неопределенные константы, можем сказать, что $C = C_2 - C_1$.

$C' = e^C$. Пропотенциируем данное выражение:

$$y(t) = e^{Kt} \cdot C'$$

$$y(0) = C' = y_0$$

$$y(t) = y_0 \cdot e^{Kt}$$

Получили решение задачи Коши для данной системы.

Давайте предположим, что $K_y > K_p$. Это означает, что $\left(K_p - K_y\right) < 0$. Тогда если устремим t к бесконечности, то $y(t) \to 0$. Если демографический коэфициент меньше нуля, то население страны вымрет. Экспонента – очень быстро убывающая функция. Так что население вымрет очень быстро.

С другой стороны, если $K_y < K_p$, то при $t \to \infty$ $y(t) \to \infty$. Эта скорость даст очень быстрый прирост населения.

В силу замкнутости, например, земного шара, к бесконечности такая величина стремиться не может. Это не значит, что решение неверное, но это лишь часть правды. Допустим, мы не знаем точную функцию, по которой меняется население (или другая величина).

$$\Delta y = y(t + \Delta t) - y(t) = \underbrace{f(y(t))}_{\text{неизвестная функция}} \Delta t$$

Разложим по Тейлору:

$$y(t + \Delta t) - y(t) = f(y(t))\Delta t = \left[\underbrace{f(0)}_{0} + \underbrace{f'(0)}_{k} y(t) + \underbrace{\frac{f''(0)}{2}}_{-k_2}\left(y(t) - \underbrace{y(0)}_{0}\right)^2\right]\Delta t$$

Что касается второй производной, было эмперически доказано, что эта величина отрицательна. Тогда применим предельный переход:

$$\lim_{\Delta t \to 0} \frac{y(t + \Delta t) - y(t)}{\Delta t} = \lim_{\Delta t \to 0}\left[ky - k_2 y^2\right]$$

$$\begin{cases} \dot{y}(t) = ky(t) - k_2 y^2 \\ y(0) = y_0 \end{cases}$$

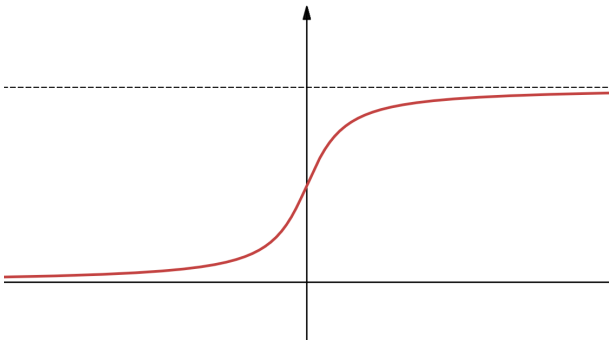
$$\frac{dy}{ky - k_2 y^2} = dt$$

Применяем интегрирование:

$$\int \frac{dy}{ky - k_2 y^2} = \int dt = t + C$$

$$\int \frac{dy}{y(k - k_2 y)} = \int \left[\frac{A}{y} + \frac{B}{k - k_2 y}\right] dy$$

Получим арктангес, который даст нам следующую картину:



Что еще важно: у ОДУ кроме обычных решений существуют еще так называемые «особые решения», которые, несмотря на их тривиальность, дают очень много для понимания поведения реальной системы, которая описывается ОДУ. Одним из типов особых решений для уравнения вида:

$$\dot{y} = f(y)$$

Это решение вида (точечное решение):

$$y(t) = y_0 = \text{const}, y_0 : f(y) = 0$$

Константа y_0 должна согласовываться с начальным условием. Почему $y(t)$ равно константе будет решением? Подставим такое решение в наше дифференциальное уравнение, тогда левая часть обратится в ноль, так как производная константы равна нулю, а правая часть обнулиться так как мы искали y_0 таким образом, чтобы $f(y) = 0$. Начальные условия так же выполнены.

Какие же типы особых решений есть у наших демографических уравнений?

$$\dot{y} = ky = 0 \Rightarrow y(t) \equiv 0$$

Если людей не было, то они и не появятся.

Тогда кроме нашего арктангенса получаем еще два решения, на самом деле это верхняя и нижняя граница нашего арктангенса. Если мы рассмотрим другие решения при $t \to \infty$, мы заметим, что при $K < 0$ все наши решения будут стремиться к тривиальному решению, то есть к нулю. А если $K > 0$, они будут уходить от этого решения.

Какие особые решения есть у более сложного уравнения, полученного через Тейлора?

$$\dot{y}(t) = ky(t) - k_2 y^2 = 0 \Rightarrow y \equiv 0; y(t) \equiv \frac{k}{k_2}$$

Если людей мало, то рост населения экспоненциальный. Та простая модель была верной, но только на определенном промежутке.

Когда число людей превышает определенный порог, то число людей выходит на определенное плато (а именно $\frac{k}{k_2}$ – емкость системы, в данном случае система на определенной стадии развития может прокормить определенное количество людей). Это называется порог системы. Переход от экспоненциального роста к плато называется предельным переходом. Предел роста. У каждой системы есть такой предел.

Когда мы пробиваем предел, система какими-то жесткими механизмами возвращают состояние системы на наше плато. Психологически сложно осознать, что система достигла своего предела и нужно что-то менять.

Замечание: ОДУ делятся на линейные и нелинейные. В линейных правая часть зависит от неизвестной функции и ее производных линейно, во втором случае она зависит нелинейно. Это порождает два лагеря математиков.

Простота линейных в том, что они решаются простой формулой. Для нелинейных таковой нет. Бельгийский ученый (Илья Пригожин) сформулировал гипотезу: мир нелинеен, темпорален, случаен. Линейные уравнения – лишь приближение к нелинейным в некоторой узкой области параметров.

Лекция 9. Качественная теория ОДУ. Нейродифференциальные уравнения

Нужно идею дифференциальных уравнений развить до тех идей, которые будут использоваться в сильном искусственном интеллекте.

Качественные и количественные теории ОДУ.

Количественная теория ОДУ предполагает, что базовым объектом исследования является само по себе ОДУ, и представляет собой некую сводку правил аналитического решения такого рода уравнения. Классический подход того, как излагать ОДУ. Этот подход сталкивается с двумя возражениями, когда речь идет о реальных, практических задачах, связанных с ОДУ.

Первое возражение: Подавляющее число ОДУ не имеет аналитического решения. (не мы не можем его найти, а его просто нет, доказан факт его отсутствия) И в этом случае мы выходим за рамки классических ОДУ и приходим в другую математическую дисциплину: численные методы или вычислительная математика. Это рассказ о том, как решать ОДУ с помощью компьютера. Термин носит название численно интегрировать ОДУ.

Какая базовая проблема ОДУ на примере примера с прошлой лекции? При решении у нас возник коэфициент рождаемости, что, на самом деле, сложно вычислимая вещь. Коэфициент рождаемости еще звучит реально, но когда мы уходим за рамки простых физических процессов, мы обнаруживаем, что эти коэфициенты определяются нестрого. Когда речь идет о химических-физических задачах, предметы подвержены каким-то простым законам и нам легко их описать, а если система сложная (а сильный интеллект относится к сложной системе), то законов тоже много и они сложны.

В математической экологии есть классическое диференциальное уравнение хищник-жертва (модель Лотки-Вольтерры)

Если $x(t), y(t)$ это количество жертв и хищников, то выполняется следующее

$$\dot{x}(t) = \alpha x(t) - \gamma x(t) y(t)$$

$$\dot{y}(t) = \beta y(t) + \gamma x(t) y(t)$$

Так же в математической экологии есть десятки версий таких уравнений для разных фаун, однако они все улавливают фундоментальные, качественные принципы поведения системы.

Отсюда следует следующее:

При исследовании реальных процессов, описываемых ОДУ, нам важен не вид конкретного дифференциального уравнения, а качественные свойства, которые эту модель описывают. Глобально эта идея носит название мягкого моделирования. А применительно к ОДУ, она дает нам качественную теорию ОДУ. Качественная теория работает с качественными свойствами и базовым ее объектом являются так называемые потоки ОДУ, о которых мы поговорим позже.

Если мы хотим работать со сложными системами, то мы должны работать, во-первых, с численными методами, во-вторых, с качественной теорией ОДУ. Это наша базовая интенция.

Качественная теория: не останавливаемся на решении одного уравнения, нам интересно изучать свойства набора уравнений

Численные методы решения ОДУ. В подавляющем большинстве аналитических решений не существует и нужно их искать численно. Здесь возможны два варианта: Задача Коши :

$$\begin{cases} \frac{dy}{dx} = f(x,y) : y \in \mathbb{R}^1 \\ y(0) = y_0 \end{cases}$$

И краевая задача. Рассмотрим, как решать задачу Коши в одномерном случае, но в общем случае алгоритм такой же, просто более сложный, матричный.

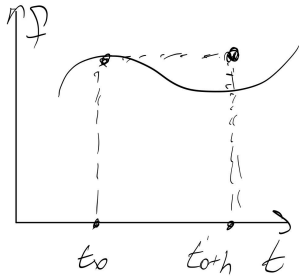
Разобьем область определения $y \in [0, x_{\max}]$ на маленькие кусочки длины h. Стоит подчеркнуть, что h являются малыми, но не бесконечно малыми, около $10^{-2}, 10^{-3}$ в сравнении с $x_{\max}$. Рассмотрим первый из таких промежутков, лежащий от 0 до h. Перепишем наш диффур в виде $dy = f(x,y)dx$. Сначала применим оператор интегрирования $\int_0^h$. Применим к нашему диффуру формулу Ньютона-Лейбница на этом участке.

$$y(h) - y(0) = \int_0^h f(x,y)dx$$

$$y(h) = y(0) + \int_0^h f(x,y)dx$$

Если нам удастся оценить этот интеграл, то мы сможем найти $y(h)$.

Самый простейший метод оценки – посчитать площадь трапеции под функцией.



Давайте рассмотрим площадь прямоугольника, она будет примерно равна

$$h \cdot f(0, y(0))$$

Это выражение мы можем вычислить, значит мы знаем $y(h)$. Давайте рассмотрим следующий кусочек $[h, 2h]$. Такой же логикой мы можем найти $y(2h)$, зная $y(h)$. В результате последовательного применения численного интегрирования мы получаем последовательность значений исходной функции, удовлетворяющей граничным условиям и дифференциальному уравнению. Находим решение для этой функции в конечном множестве точек. Так как h по сути параметр, мы можем сделать эти значения сколь угодно частыми. Если мы нарисуем график, то увидим, что аналитическое и численное решения совпадают. Таким образом, нет большого смысла искать аналитическое решение в реальных задачах.

Следует иметь в виду, что любой численный метод обладает погрешностью, вносимой на каждом шаге численного интегрирования. Чем дальше мы двигаемся, тем большую погрешность мы внесли в решение. Для рассмотренного метода Эйлера погрешность на каждом шаге составляет $O(h)$. Значит в самом неудачном случае двигаясь по промежутку мы можем внести погрешность сравнимую с значением функции.

К счастью, существуют гораздо более эффективные методы численного интегрирования. Самым базовым является метод Рунге-Кутты (4-го порядка). Он тоже основывается на достаточно логичных принципах, мы функцию разности между истинной и аппроксимированной функцией расскладываем в функцию Тейлора. Погрешность, которую дает этот метод на каждом шаге: $O(h^s)$, где $s -$ взятый порядок.

Отступление. Нейродифференцальные уравнения.

Какое-то время назад в нейронных сетях произошла революция, когда все перешли с малых моделей на глубокое обучение. Следующей же революция будет связана с нейродифференцальными уравнениями.

У нас некая нейронная сеть, например Resnet, $x_{n+1} \in \mathbb{R}^s = x_n + f(x_n)$ – состояние следующего слоя. Состояние следующего слоя – функция активации от текущего слоя. В теории нейродифференцальных уравнений мы говорим: давайте считать, что процесс протекания информации от входа в нейронную сеть к выходу – процесс, разворачивающийся в дискретном времени с шагом единица.

Для первого слоя мы находимся в $n = 0, n = 1, ..., n = N, \Delta t = 1$

Это означает, что у нас есть две величины: одна $N -$ число слоев, а вторая, $T -$ (физическое) время, которое протекло от момента подачи информации до момента выхода.

До этого мы их не различали, потому что считали, что $T = N \cdot \underbrace{\Delta t}_{=1}$. Дальнейшая логика изложения: что такое переход от неглубокого обучения к глубокому? Ранее N было небольшим, позже мы обсудим почему, но произошла некая революция, модели глубокого обучения, где мы сказали, что $N -$ весьма большая величина. Нейродиффурщики сказали: давайте доведем процесс до конца. Давайте не считать время T фиксированным, а N устремим к бесконечности. $T -$ константа, $N \to \infty$, следовательно, $\Delta t \to 0$. Перепишем состояния слоев:

$$\frac{x(t + \Delta t) - x(t)}{\Delta t} = f(x(t))$$

Давайте применим оператор предельного перехода к обоим частям равенства:

$$\lim_{\Delta t \to 0} \frac{x(t + \Delta t) - x(t)}{\Delta t} = f(x(t))$$

Получаем ОДУ: $\frac{dx}{dt} = f(x(t))$. Получаем обычную задачу Коши, остается только добавить, что $x(0) = x_0$. Следуя традиции классического машинного обучения мы смотрим не на истинный объект (нейродиффур), а какое-то приближение:

$$x_{n+1} = x_n + f(x_n)$$

Если внимательно посмотреть на формулу Эйлера, то мы обнаружим, что на самом деле мы для исследования истинного объекта используем старый метод Эйлера. Сейчас идет стремительный переход к нейродифференцальным уравнениям, мы поговорим о них далее.

То, что мы сейчас расписали – это лишь то, что называется прямой ход нейронной сети. Мы написали уравнение, которое описывает изменение состояния. Здесь у нас слоев бесконечно количество, но сути это не меняет. Понятно, что здесь мы не рассмотрели главного: в нашей нейронной сети нет весов. При рассмотрении нейродифференцальных уравнений мы, конечно же, должны рассматривать и вектор-функцию весов $w(t)$, где $w(t)$ есть некий аналог матрицы (тензора) весов нейронной сети, который получается из него тем же предельным переходом $\Delta t \to 0$. Тогда мы имеем, что наш диффур зависит не только от состояния $x(t)$, но и от $w(t)$, которая и подлежит определению.

$$\frac{dx}{dt} = f(x(t), w(t))$$

Мы будем определять ее следующим образом:

У нас есть некая функция ошибки (потери), которая характеризует состояние сети на выходном слое, или в наших новых терминах состояние сети в последний момент времени T и некие указания учителя x^*, причем эту функцию мы стремимся минимизировать.

$$L(x(T), x^*) \to \min$$

Например квадратичная функция ошибки:

$$L = \frac{1}{2}[x(T) - x^*]^2$$

$$dx = f(x(t), w(t))dt$$

$$\int_0^T dx = \int_0^T f(x(t), w(t))dt$$

$$x(T) = \underbrace{x(0)}_{=x_0} + \int_0^T f(x(t), w(t))dt \Rightarrow$$

$$\Rightarrow L = \frac{1}{2}\left[x_0 + \int_0^T f(x(t), w(t))dt - x^*\right]^2 \to \min$$

$w(t)$ — наша степень свободы. Наша задача найти такое $w(t)$, чтобы наша функция ошибки была минимальной.

Хотим найти минимум этого функционала при ограничениях:

$$\begin{cases} \frac{dx}{dt} = f(x(t), w(t)) \\ x(0) = x_0 \end{cases}$$

Это делается вариационными вычислениями, тема следующей лекции.

Лекция 10. Вариационное исчисление.

В прошлый раз мы показали, что всякая нейросетевая модель представляет собой дискретный вариант нейродифференциального уравнения. Точнее, ее прямой ход – это нейродифференциальное уравнение, ее обратный ход, отыскание функции весов сети, суть есть задача вариационного исчисления (оптимального управления – когда вариация и сама функция могут быть разрывны в конечном числе точек).

То, в каком виде у вас сформулирована задача, представляет собой ситуацию, когда мы подаем на вход сети ровно одно наблюдение (x_0), что является неестественной ситуацией. Нормальная ситуация заключается в том, что мы подаем на вход некоторую выборку $(\xi_1, ..., \xi_n)$, полученные из некоторого распределения $P_0(x)$, а получаем некоторую выборку $(\eta_1, ..., \eta_n)$, полученной из распределения $P_1(x)$.

В отличии от теории вероятности в статистике рассматриваются не только теоретические функции распределения, но и их эмперические аналоги.

Собственно говоря, вся статистика заключается в том, что мы пытаемся построить некий аналог теоретической величины (рассмотренной в теории вероятности), такой, что этот аналог восстановим по выборке (является функией выборки) и он в заданном смысле близок к неизвестной теоретической величине. Он хорош в статистическом смысле.

Можем ли мы восстановить распределение выборки? Отсортируем выборку $\xi_1, ..., \xi_n$ по возростанию и получим вариационный ряд $\xi_1^*, ..., \xi_n^*$. Построим следующую функцию:

$$\hat{F}_n(x) = \begin{cases} 0, & x < \xi_1^* \\ \frac{l}{n}, & x = \xi_i^* \\ 1, & x > \xi_n^* \end{cases}$$

Подробнее см. лекцию 4, где мы строили эмперическую функцию таким же образом. (+ рисунок)

Такая функция является хорошим приближением к истинной функции распределения $F(x)$. Более того, для этой функции доказаны весьма сильные математические свойства:

1) $plim\ \hat{F}_n \underset{n\to\infty}{\to} F(x)$ – Теорема Гливенто-Кантелли. Это означает, что эимперическая функция распределения является тем более сильным приближением к истинной функции, чем больше размер выборки N.

Соответственно, мы можем утверждать, что если мы имеем дело с большой выборкой, то мы можем рассматривать восстановленную по выборке функцию

распределения как хороший аналог, замену истинной функции распределения, которую мы не знаем.

Более того, для эмперической функции распределения Колмогоровым было доказано еще более сильное утверждение. Если мы рассмотрим статистику KS вида

$$KS = \sqrt{n} \sup_{-\infty<x<+\infty} |F(x) - \hat{F}_n(x)|$$

То полученная случайная величина будет иметь одно и то же распределение для всех функций $F(x)$, так называемое распределение Колмогорова. Оно даже выразимо в виде элементарных функций.

Это означает, что мы можем угадывать истинную функцию распределения.

Пусть у нас есть некое нейродифференциальное уравнение:

$$\dot{x} = f(x(t), w(t))$$

Которое представляет собой некий объект, описывающий динамику процесса обучения. Соответствующий дискретный вариант, каким бы способом дискретизации он не был получен, (методом Эйлера или более продвинутым методом численного интегрирования Ранги-Кутта или другими) мы будем считать его аппроксимацию.

Если на фазовом пространстве данного нейродифференциального уравнения (на пространстве, на котором определен $x, x \in \mathbb{R}^n$) задано некое распределение $P(x)$, то оно будет меняться с течением времени t под действием потока этого нейродифференциального уравнения. Как собственно говоря под действием потока любого дифференциального уравнения.

То есть на самом деле имеем $P(x, t)$.

Здесь возникают два вопроса:

1) Что будет происходить с функцией $P(x)$ под действием потока нейродифференциального уравнения? Как она будет меняться?

2) Что мы хотим от этой функции $P(x)$? Как она должна соотноситься с нашими выборками $(\xi_1 ...., \xi_n)$ и $(\eta_1, ..., \eta_n)$? Ответить на второй вопрос легче.

Мы хотим, чтобы в момент времени $t = 0$, то есть, $P(x, 0)$, она как можно меньше уклонялась от эмперической функции $\hat{P}_0(x)$.

По той же логике на выходе сети $t = T$, $P(x, T) \parallel \hat{P}_1(x)$. ($P(x, T)$ как можно меньше уклонялось от эмперической функции $\hat{P}_1(x)$.)

Хотим выборку на входе преобразовать в выборку на выходе так, чтобы преобразование производилось данной нейросетевой архитектурой.

Дивергенция Кульбака-Лейблица (Kullback, Leiblez)

$$D_{KL}(P(x) \parallel q(x)) = \int_{\mathbb{R}^n} q(x) \ln \frac{p(x)}{q(x)}$$

Такая функция не удовлетворяет требованиям метики, но она онастолько эффективна и популярна, что является методом по умолчанию.

Мы должны сказать, что наша функция $P(x,t)$ для того чтобы реализовывать процедуру обучения данного нейродифференциального уравнения должна удовлетворять следующему свойству:

$$\lambda_1 \cdot D_{KL}\Big(P(x,0) \parallel \hat{P}_0(x)\Big) + \lambda_2 \cdot \Big(P(x,T) \parallel \hat{P}_1(x)\Big) \to \min_{P(x,t)}$$

Где $1 \geq \lambda_1, \lambda_2 \geq 0, \lambda_1 + \lambda_2 = 1$, часто берут просто $\lambda_1 = \lambda_2 = \frac{1}{2}$

$q(x)$ и $p(x)$ в данном случае заданы, интеграл дает определенное число. Мы функции $P(x,t)$ дали в соответствие значение функционала, и мы хотим найти такую $P(x,t)$, которая бы минимизировала данный функционал.

Это классическая задача вариационного исчисления.

В отличии от классической постановки безусловного вариационного исчисления данная постановка имеет 3 ограничения:

$$1)\ \forall t \in [0,1], P(x,t) > 0 \hookrightarrow \int P(x,t)dx = 1$$

$$2)\ \begin{cases} \dot{x} = f(x(t), w(t)) \\ x(0) = x_0 \end{cases}$$

3) $P(x,t)$ меняется под действием потока данного нейродифференциального уравнения не произвольно, а некоторым вполне определенным образом. Давайте выясним, как же меняется $P(x)$ под действием потока дифференциального уравнения. Это уравнение Перрона-Фробениуса. УПФ.

Вначале обратимся к дискретному случаю (он проще). Рассмотрим отображение фазового пространства $\mathbb{R}^n$ в себя.

$$x_{i+1} = f(x_i), f : \mathbb{R}^n \to \mathbb{R}^n$$

Рассмотрим один шаг данного отображения. $y = f(x)$. На самом деле это просто замена координат в пространстве $\mathbb{R}^n$. Предполагаем, что $f(x)$ дифференцируема необходимое количество раз. Выбор системы координат произволен. Можем взять любые, которые получаются друг из друга диффиаморфными преобразованиями.

К сути рассматриваемого явления система координат не имеет. Физические законы независимы от системы координат. Вероятности и качественные свойства не зависят от системы координат.

Как меняются вероятности при смене системы координат? Давайте формализуем эту идею, построим математическую модель этого утверждения:

Возьмем произвольную точку y и некую ее бесконечно малую окрестность

$$\left[y - \frac{\Delta y}{2}, y + \frac{\Delta y}{2}\right]$$

Где Δy бесконечно мало. Так же рассмотрим ее прообраз:

$$x_i : \left[x_i - \frac{\Delta x}{2}, x_i + \frac{\Delta x}{2}\right]$$

У нас есть отображение f, оно отображает точки y в точки x_i. $P(x)$ при этой замене координат преобразуется в $q(y)$, при этом мы требуем:

$$q(y)\Delta y = \sum_i P(x_i)\Delta x$$

Эта запись говорит о том, что вероятности сохраняются. При этом очевидно, что длины взятых окрестностей тоже будут меняться (для разных f)

Разделим последнее равенство на Δy, и устремим $\Delta x \to 0$

$$q(y) = \sum_i p(x_i)\frac{\Delta x}{\Delta y}$$

По теореме об обратной функции получаем

$$q(y) = \sum_i p(x_i)/f'(x_i)$$

Нам будет удобно переписать это выражение пользуясь свойствами δ-функции.

В математике кроме обычных привычных нам функций существуют так называемые «обобщенные» функции, которые представляют собой некое обобщение привычного понятия функции, сохраняющего его свойства, но не представимого в терминах отображения. Такие функции получаются естественным путем, причем естественным с точки зрения приложений, прикладной математики, и естественным с точки зрения чистой математики. С точки зрения чистой математики в рамках функционального анализа доказывается полнота пространства функций, интугрируемых в терминах L^2, то есть всех таких функций, для которых

$$\int_a^b f^2(x)dx < \infty$$

Но сюда попадают и другие весьма интересные функции. Одна из них δ-функция (Дирака). Формально она определяется следующим образом:

$$\delta(x) = \begin{cases} 0, & x \neq 0 \\ ?, & x = 0 \end{cases}$$

$$\int_{\mathbb{R}} \delta(x)dx = 1$$

Вся квантовая механика построена на этой функции. Откуда она берется? Давайте рассмотрим последовательность вполне приличных функций: это функции, которые равны n на промежутке $\frac{1}{n}$, и равны нулю на остальной области определения. Эти функции интегрируемы с квадратом, каждая из них принадлежит пространству L^2, интеграл для каждой равен единице.

Мы только что сказали, что пространство L^2 полно, значит предел ряда последовательности таких функций тоже функция, которая принадлежит L^2. Что мы знаем про этот предел? Этот предел существует, он равен нулю во всех точках, не равных нулю, он уйдет к бесконечности в точке 0, но интеграл равен единице. Это и есть наша дельта-функция.