

Прологомены к высокому искусственному интеллекту

Конспект лекций

Автор конспекта:

Король Михаил

Содержание

Лекция 1. Вводная	3
Лекция 2. Свойства сложных сетей.	4
Лекция 3. Ассортативность и дисассортативность.	7
Лекция 4. Идентификация степенных распределений.	11
Лекция 5. Метод Прюснера, Mind-Brain problem.	18
Лекция 6.	23

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 асимптотически при $X \rightarrow \infty$ дадут нам одно из степенных распределений при том или ином значении α .

Какими свойствами обладает данное распределение?

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

$$\text{Var}(X) = \mathbb{E}(X - \mathbb{E}(X))^2$$

Существуют такие значения α , при которых $\mathbb{E}(X) \rightarrow \infty$, 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.

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

В случае неориентированных 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 характеристики 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 $N \rightarrow \infty$.

Очевидно, что если граф полносвязный, или близкий к полносвязному (неразрезанный), 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 $\rho \rightarrow 0$. Если мы будем наблюдать промежуточную ситуацию, где $E \rightarrow O(N^\alpha)$, $1 < \alpha < 2$, 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 $(i_0, i_1), (i_1, i_2), \dots, (i_{n-1}, i_n)$ 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

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

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

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

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

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

Оказалось, что если расстояние ведет себя как $O(n^\beta)$, 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.

$$\langle l \rangle \sim O(\ln N)$$

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 потере связности, 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 (resilient). 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.

Дисассортативные 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 (фундаментальная 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:

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

где k – степень просматриваемой вершины, n – число ее соседей, k_1, \dots, 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 робастной (в первом приближении вычислительно устойчивой) (если мы немного изменим выборку, то значение характеристики тоже должно немного измениться)

Почему $P(k \mid k_1, \dots, k_n)$ 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, \dots, x_n ей порожденной и случайную величину Y и выборку y_1, \dots, 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}

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

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

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 дисассортативна, 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

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

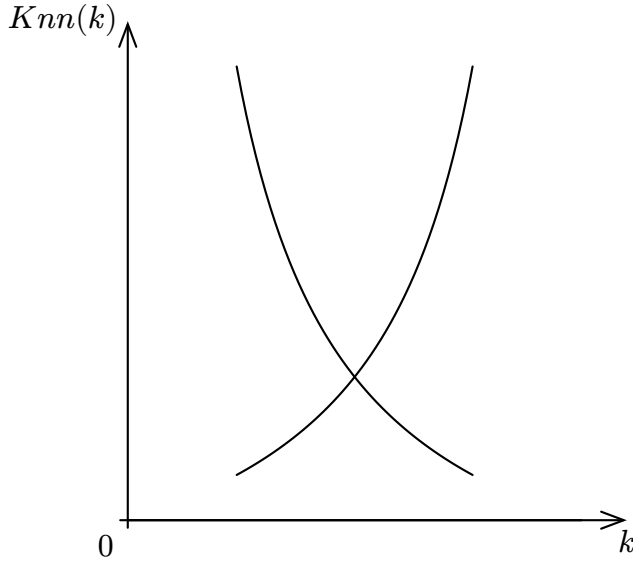
Здесь k_i — степень вершины i , k_j — степени соседних с ней вершин, но суммирование идет не по всем соседним вершинам, но только по тем, которые имеют степень k .

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 $(\xi_1, \dots, \xi_n)_{iid}$. Мы можем ставить в отношении 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?

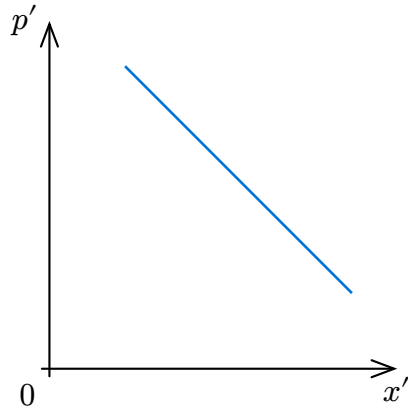
Первый метод носит название Метод Хилла (Hill), он базируется на переходе к двойному логарифмическому масштабу. Если мы прологарифмируем выражение плотности для степенного распределения, мы получим:

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

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

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

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 $p'_1, \dots, p'_n, x'_1, \dots, x'_n$, тогда давайте посчитаем минимум следующей функции:

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

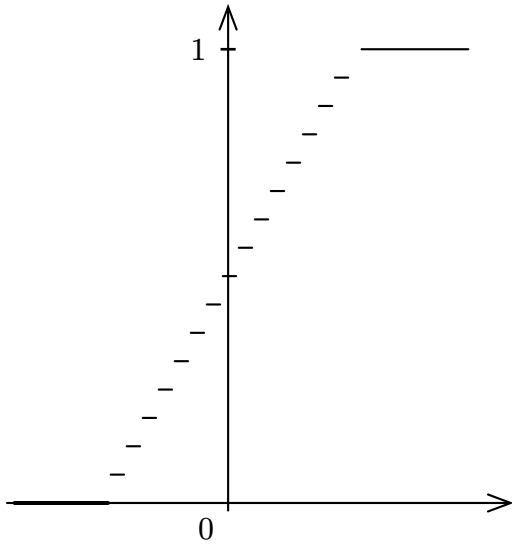
где n – размеры выборки. Отсюда, дифференцируя по C' и α (p'_i и x'_i нам известны) мы находим выражения для оценки C' и α , которые минимизируют это выражение. Если окажется, что полученная таким образом оценка (оценка методом наименьших квадратов) действительно делает этот квадратичный функционал малым, 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 (ξ_1, \dots, ξ_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 (ξ_1, \dots, ξ_n), по возрастанию, и получаем из нее так называемый вариационный ряд (ξ_1^*, \dots, ξ_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 $F(x)$

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

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.

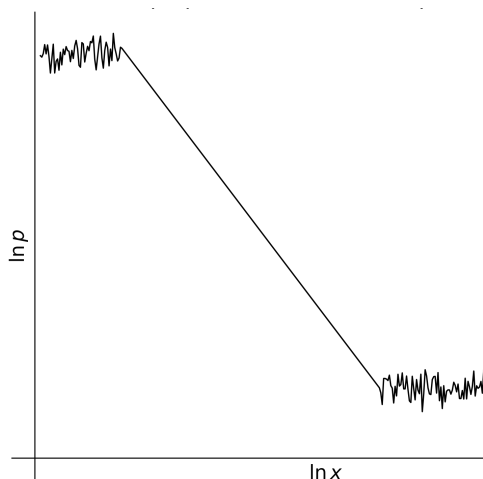
Каковы же недостатки критерия KS ? Для того, чтобы применять критерий KS к выборкам в реальных задачах мы должны знать точное значение параметра α , а для того, чтобы сколько-нибудь адекватно оценить параметр α , мы должны быть уверены, что выборка, с которой мы имеем дело, порождена степенным распределением. 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

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

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 наряду с исходной выборкой рассмотреть еще значительное число синтетических выборок, а именно синтетических выборок, порожденных законом распределения:

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

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 (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 выходит далеко за пределы анализа степенных распределений, но Прюснер рассматривал конкретно Collapse данных степенных распределений.

Более того, Прюснер работал с распределениями вида

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

где C — константа нормализации, $\dots x_c$ — 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 :

рисунок

Степенное распределение обычно предполагает бесконечное число наблюдений, Метод Прюснера лучше прошлых методов, так как может работать с горбами. Метод Прюснера базируется на двух китах:

1 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 сконцентрирована в окрестности максимума, но какой-то объем данных, в силу того, что хвост у него тяжелый, будет сконцентрирован дальше.

гистограмма

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

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 отбрасываемые), 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:

рисунок

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 Плюсер называл Landmark. В силу того, что выборки у нас разные, мы получаем разные положения всех трех участков (они сдвинуты друг на друга). С другой стороны, поскольку α' получаена с помощью какой-то грубой оценки, то мы имеем дело с наклонным участком прямой, но то, что этот участок есть, является первым признаком 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.

рисунок

В этом графике будет участок меньше x_{\min} , далее горизонтальный участок прямой, указывающий на то, что распределение является степенным, и характерный landmark, отвечающий функции горба g . Для осуществления коллапса, перехода от верхнему графику к нижнему, 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 Плюсер 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}, α, 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 развивающейся 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)

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, ее самый нижний уровень – конетком, 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 (например, зрительная кора)

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 манд-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 язык и .., соответственно сильны искусственный интеллект тоже будет таковой. К самоорганизованной критической системе относятся системы удовл. трем след. 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. Размеры лавин подчиняются степенным законам распределения. Если это так, то система является самоорганизованной критичной. 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.

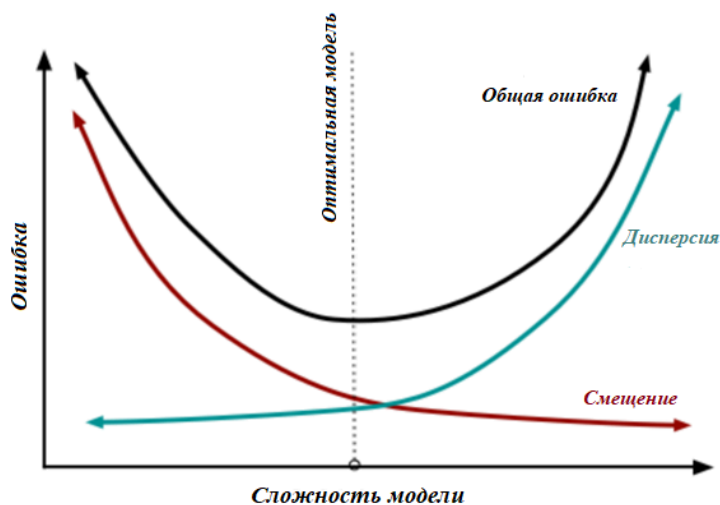
Самый важный тег для нас – сложность. Если мы говорим о сильном искусственном интеллекте, то мы должны говорить о сложности.

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 ... – 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.



TODO: сделать другой график

Если мы продолжаем увеличивать сложность системы, то мы получаем

рисунок 2

Прежде чем переходить к предсказательная сложности, стоит поговорить про информацию и

Шэновская информация. Давайте рассмотрим два вероятностных события

Событие А: Профессор вошел в кабинет.

Событие Б: Профессор убил студента.

Вероятность второго события существенно ниже. После этого Шэнон сказал:

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

Где I - информация, p – вероятность.

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

Функция логарифма удовлетворяет такому свойству. С другой стороны можно доказать, что единственная функция, которая удовлетворяет такому свойству – логарифм.

$$\text{Тогда } I = \ln \frac{1}{p} = -\ln p$$

Предположим, что у нас есть система, которая может находиться в N состояниях.

И в этих состояниях она находится с вероятностями p_1, \dots, p_n . Тогда энтропия по Шэнону есть средняя информация, которую мы знаем о такой системе.

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

Свойства:

Предположим, что мы знаем о системе все, тогда мы можем утверждать, тогда $p_i = 1$, остальные равны нулю. 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.

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

Информационная энтропия – мера неопределенности системы.

$$\text{Как генерировать хаотические ряды? } X_{n+1} = 1 - \lambda X_n^2$$

Введенная таким образом Шэновская энтропия является аддитивной в силу выполнения свойства, более того, можно показать, что она является единственной аддитивной энтропией. Между тем в сложных системах обычно имеет место так называемые не аддитивные энтропии, для которых свойство два не выполняется по банальной причине: считается, что не существует двух подлинно информационно изолированной системы. 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 основные: энтропия Реньи(renyi)

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

Энтропия Цависа(Tsalis), и так далее..

Открытая проблема, к какой сложной сети какую энтропию применять.

Общее методическое замечание: Сложные системы делятся на классы универсальности.

Предсказательная сложность. Предположим, что у нас идет некоторый поток данных в некоем времени, не обязательно метрологическом, и мы можем выделить в нем некоторый момент времени $t = 0$ и два момента: прошлое и будущее. Тогда данные, относящиеся к прошлому мы будем обозначать как X_{past} , а к будущему X_{future} , ... $P(X_{past}), P(X_{future})$, тогда взаимной информацией между будущим и прошлым будет

$$I(X_{future}, X_{past}) = \langle \log_2 \frac{P(X_{future} | X_{past})}{P(X_{future})} \rangle$$

Где усреднение производится по совместному распределению X_{future}, X_{past} .

Эта величина носит название Девергенция Кульбаха Лейблера (Kulback Leibler). Вот математическая мера того, насколько нам сложно предсказать будущее по прошлому. Если мы умножим числитель и знаменатель на $P(X_{past})$, то получим

$$\begin{aligned} - \langle \log_2 P(X_{past}) \rangle - \langle \log_2 P(X_{future}) \rangle &= - [- \langle \log_2 P(X_{future}, X_{past}) \rangle] = \\ &= H(T') + H(T) - H(T + T') \end{aligned}$$

.....

T = для отрезка future, T' длина отрезка past

Замечание 1:

Выражение, введенное для данной информацией является симметричным, тогда равносильными являются задачи как предсказания, так и подсказания, когда мы прошлое определяем по будущему.

Энтропия состоит из двух составляющих:

первое – экстенсивная энтропия, и связана с течением времени. То есть, любая система подчиняется второму началу термодинамики. В этом смысле энтропия растет всегда, причем растет пропорционально протекающему времени. Это имеет место абсолютно для любой системы. И это в общем то неинтересно.

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

второе – Суб-экстенсивная энтропия, которая ведет себя как $o(T)$ и определяет поведение системы, которая отличает ее от всех остальных

Мы хотим показать, что разные системы дают нам принципиально разные 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 экстенсивные составляющие энтропии сокращаются. Соответственно, совместная информация определяется только Суб-экстенсивной энтропией.

$$H_s(T') + H_s(T) - H_s(T + T')$$

Предположим, что мы знаем все прошлое, тогда взаимная информация в пределе

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

.....

Здесь можно сказать, что, когда мы говорим о такой хорошей, фундаментальной величине, она должна удовлетворять двум, противоречащим

- 1) она следует из базовых принципов
- 2) ее можно наблюдать в эксперименте

Метрика хорошо измеряется.