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 вопрос от хаотического? Это базовый вопрос теории прогнозируемых рядов. Регулярные ряды – простые, хаотические ряды – сложные. Простые системы, обычные системы искусственного интеллекта порождают регулярные ряды, сильный искуственный 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.

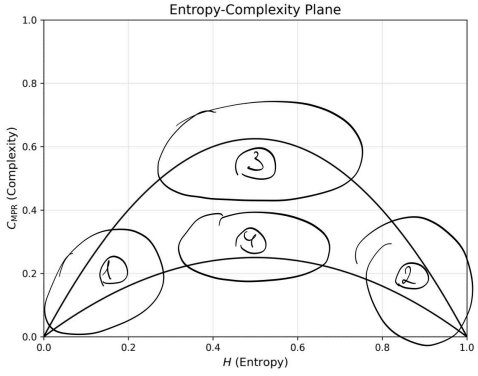
$$0 \leq H \leq 1$$

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 детерменированный 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, $\varepsilon_i \sim N(0, \sigma^2)$