Решение регрессионной задачи машинного обучения на основе теории случайных функций

Бахвалов Ю.Н., к.т.н., Малленом Системс, г. Череповец

Аннотация.

В статье исследуется регрессионная задача машинного обучения как задача многомерной аппроксимации с использованием математического аппарата теории случайных функций. Показано, что если допустить существование в бесконечномерном функциональном пространстве некоторых симметрий плотности вероятности, то этого достаточно, чтобы получить точный метод решения задачи.

Ключевые слова: машинное обучение, регрессия, случайная функция, симметрия, корреляционная функция, спектральная плотность, полигармонический сплайн.

Solving a machine learning regression problem based on random function theory

Bakhvalov Y.N., Ph.D. Mallenom Systems, Cherepovets

Annotation.

The article examines the regression problem of machine learning as a problem of multidimensional approximation using the mathematical apparatus of the theory of random functions. It is shown that if we assume the existence of certain probability density symmetries in an infinite-dimensional function space, then this is enough to obtain an exact method for solving the problem.

Keywords: machine learning, regression, random function, symmetry, correlation function, spectral density, polyharmonic spline.

В отличие от прототипа данного исследования (размещенного автором на [4], которое не публиковалось в научных рецензируемых журналах) в данной статье задача ставится и решается в другом виде, исходно как задача аппроксимации. Дополнены положения о симметриях в функциональном пространстве, исправлены многие неточности, изменено представление случайной функции на более правильное. Многие преобразования изменены и добавлены новые.

Пусть есть обучающая выборка в виде набора векторов на входе $x_1, x_2, \dots, x_k (x_i \in R^n)$ размерностью n и набор значений на выходе $y_1, y_2, \dots, y_k$ $(y_i \in R)$.

Решение регрессионной задачи машинного обучения в этом случае можно представить как решение задачи аппроксимации в виде:

$$y_i = f(x_i) + u_i \quad (1)$$

где $f(x)$ – функция, связывающая входные и выходные значения x_i и y_i, искомая модель, которую нужно определить, основываясь на обучающей выборке;

$u_1, u_2, \dots, u_k \in R$ – независимые случайные величины, с нормальным законом распределения и нулевым математическим ожиданием, которые, можно считать, что были добавлены к $f(x_i)$, и поэтому ее значения могут не совпадать точно с y_i (тем самым моделируем погрешности или неоднозначности, которые могут присутствовать в обучающей выборке)

Задачу также можно интерпретировать следующим образом:

Пусть изначально существовала некоторая неизвестная функция $f(x)$. Затем над ней были выполнены эксперименты. Для некоторой выборки $x_1, x_2, \dots, x_k$ были найдены значения $f(x_i)$ (но которые нам неизвестны), к которым были прибавлены случайные, неизвестные нам величины $u_1, u_2, \dots, u_k$. В результате к уже известным $x_1, x_2, \dots, x_k$ дополнительно были получены $y_1, y_2, \dots, y_k$.

Теперь же задача состоит в том, чтобы по последовательностям x_i и y_i предположить какая могла быть $f(x)$.

Первоначально может показаться (даже если известны характеристики величин u_i), что такая формулировка ничего не дает для решения, поскольку ничего неизвестно о природе $f(x)$ и то, какая она может быть в (1) никак не раскрывается.

Но рассмотрим сначала сильно упрощенную версию задачи.

Допустим, что у нас есть “подсказка”. Предположим, что кто-то заранее сообщил нам варианты правильных ответов, какая может быть функция $f(x)$. Пусть подсказка была в виде последовательности вариантов функций $f_j(x)$ и соответствующих им вероятностей p_j, что это правильный ответ. Но вероятности этот кто-то в подсказке дал нам априорные, не зная, что затем над $f(x)$ были проведены эксперименты и у нас еще есть независимая дополнительная информация в виде последовательностей x_i и y_i.

Предположим, что нам известны дисперсии случайных величин $u_1, u_2, \dots, u_k$. Допустим, что они все одинаковы и равны σ^2. Тогда вероятности того, что j-тый вариант функции в подсказке $f_j(x)$ окажется правильным ответом, будут пропорциональны $\widehat{p_j}$:

$$\widehat{p_j} = p_j \frac{1}{(2\pi\sigma^2)^{\frac{k}{2}}} e^{-\frac{1}{2\sigma^2}\sum_{i=1}^{k} u_i^2} \quad (2)$$

где u_i можем определить как:

$$u_i = y_i - f_j(x_i)$$

Перебрав последовательно все $\widehat{p_j}$ и найдя максимальное из них, тем самым выберем наилучший вариант $f_j(x)$ как наилучшее решение задачи аппроксимации в (1).

Но, что представляет собой множество вариантов функции $f(x)$ с заданными на нем вероятностями (или заданной на нем функции распределения вероятностей)? Это ни что иное как описание случайной функции.

Как видно выше, при наличии такой подсказки (заданной случайной функции с конечным множеством реализаций), сразу же было получено решение.

Но аналогичные рассуждения можно провести и без “подсказки” в явном виде. Возьмем в качестве возможных вариантов функции $f(x)$ множество всех непрерывных вещественных функций, существующих в R^n. Это множество можно представить в бесконечномерном гильбертовом пространстве, каждое измерение которого это возможные значения функции для каждого некоторого конкретного значения $x \in R^n$.

На этом множестве в свою очередь может быть задана функция плотности вероятности $P(f)$ (что в совокупности дает описание случайной функции, которую можно использовать для решения задачи).

Эту функцию $P(f)$ можно считать неизвестной, однако, рассматривая функцию плотности вероятности в этом пространстве можно допустить существование некоторых симметрий (которые вполне можно было предположить перед тем, как над $f(x)$ были выполнены эксперименты).

1. Допустим, что если для некоторых двух функций $f_1(x)$ и $f_2(x)$ существуют такие A и t, что

$$f_2(x) = f_1(\mathrm{A}x + t),\ x, t \in R^n \quad (3)$$

A- некоторая матрица поворота

тогда плотность вероятности для $f_1(x)$ и $f_2(x)$ должна быть одинаковой $P(f_1) = P(f_2)$.

Т.е. для любого подмножества функций, которые преобразуются друг в друга путем поворота или параллельного переноса должна быть задана одинаковая плотность вероятности.

2. Допустим, что если для некоторых двух функций $f_1(x)$ и $f_2(x)$ выполняется:

$$f_2(x) = kf_1(x/k), x \in R^n \quad (4)$$

где $k \in R$ – некоторый коэффициент

тогда плотность вероятности для $f_1(x)$ и $f_2(x)$ должна быть одинаковой $P(f_1) = P(f_2)$.

Т.е. для любого подмножества функций, которые преобразуются друг в друга изменением масштаба должна быть задана одинаковая плотность вероятности.

3. Допустим, что если предположить, что решением является одна из функций на подмножестве функций, которые описываются как:

$$f_k(x) = kf_1(x), k \in R \quad (5)$$

где k – некоторое вещественное число (назовем его в данном контексте амплитудой $f_1(x)$)

$f_1(x)$ – некоторая произвольная непрерывная функция, взятая чтобы породить множество в (5)

Тогда условная плотность вероятности на этом подмножестве функций будет определяться нормальным законом от амплитуды k, с математическим ожиданием при $k = 0$.

Т.е. если мы берем любую непрерывную функцию (как одну из реализаций) и сравним ее с другой, которая повторяет первую, но отличается от нее только амплитудой, то функция с меньшей амплитудой будет более вероятна, чем с большей. А закон распределения вероятностей будет иметь характер нормального закона, зависящего от амплитуды (в (5) обозначенной через k) с математическим ожиданием равным нулю. В данном случае мы не уточняем, какая может быть дисперсия (подразумевая, что не нулевая, иначе говорить о нормальном законе распределения не имело бы смысла), а говорим лишь о характере закона распределения вероятностей.

4. Прибавим ко всем функциям в множестве, что описано в п.3. одну и ту же функцию $f_2(x)$ и получим новое множество функций, элементы которого можно описать как:

$$f_k^*(x) = kf_1(x) + f_2(x), k \in R \quad (6)$$

где k – некоторое вещественное число (амплитуда $f_1(x)$)

$f_1(x)$ и $f_2(x)$ – любые непрерывные функции

Предположим, что какую бы мы функцию $f_2(x)$ не взяли, плотность вероятности на этом подмножестве функций также как и в п.3. сохранит нормальный закон распределения от амплитуды k (но математическое ожидание уже может не равняться нулю).

То есть, если мы возьмем произвольную непрерывную функцию $f_2(x)$, а затем также возьмем произвольную непрерывную функцию $f_1(x)$ и умножая на различные варианты амплитуды k будем прибавлять ее к $f_2(x)$, порождая таким образом множество функций $f_k^*(x)$ (как бы совершающие колебания вокруг $f_2(x)$), то плотность вероятности $P(f)$ на этом множестве функций будет соответствовать нормальному закону распределения от амплитуды k (но математическое ожидание не обязательно будет соответствовать $k = 0$).

Рассмотрим следствия из п.1-4, какие будут свойства у случайной функции, для которой они выполняются.

Поскольку будем считать п.3-4. справедливыми для любой непрерывной функции, тогда из этого следует, что $P(f)$ будет бесконечномерным нормальным законом распределения в гильбертовом пространстве. Фактически, пунктами 3 и 4 мы и предполагаем, что имеем дело бесконечномерным нормальным законом распределения, впрочем, не уточняя его характеристик (кроме предположения о математическом ожидании в точке ноль).

Однако, любое невырожденное (что выполняется в нашем случае) многомерное нормальное распределение можно свести к вектору независимых нормальных случайных величин. Это означает, что обязательно найдется такая последовательность функций (образующих базис, по которому может быть разложена любая другая непрерывная функция), с помощью которой можно выразить рассматриваемую нами случайную функцию как линейную комбинацию этой последовательности функций и независимых случайных нормальных величин. А это не что иное как каноническое разложение случайной функции. Т.е. из 3 и 4 пунктов следует, что у такой случайной функции обязательно должно существовать каноническое разложение.

Каноническое разложение случайной функции (В.С. Пугачев [1], стр.248-249):

$$F(x) = m_f(x) + \sum_{j=1}^{\infty} V_j \varphi_j(x), \quad (7)$$

где $m_f(x)$ – функция математического ожидания,

V_j – некоррелированные случайные величины, математические ожидания которых равны нулю (коэффициенты канонического разложения),

$\varphi_j(x)$ – координатные функции канонического разложения

Из п.3. следует, что $m_f(x) = 0$. А также, поскольку количество функций $\varphi_j(x)$ в нашем случае будет несчетным, (7) обобщается до интегрального канонического представления случайной функции:

$$F(x) = \int_{\Lambda} V(\lambda)\varphi(x,\lambda)d\lambda, \tag{8}$$

где $V(\lambda)$ – белый шум параметра λ,

$\varphi(x,\lambda)$ – некоторая (неслучайная) функция аргумента x и параметра λ

Таким образом получается разложение всех возможных реализаций (непрерывных функций) по некоррелированным бесконечно малым элементарным случайным функциям $V(\lambda)\varphi(x,\lambda)d\lambda$. Но пока мы лишь сделали вывод что из п.3 и п.4 следует существование такого разложения.

Рассмотрим теперь п.1.

Из него следует, что рассматриваемая случайная функция (8) при выполнении п.1 должна быть стационарной. А значит, она должна иметь спектральное разложение.

Интегральное каноническое представление стационарной случайной функции представлено В.С. Пугачевым в [1], стр.333. Но используемый там белый шум $V(\omega)$, интенсивность которого может быть различной для разных частот, мы можем представить напрямую как произведение составляющих

$$V(\omega) \to \frac{1}{2}V(\omega)\sqrt{\frac{S(\omega)}{d\omega}}, \tag{9}$$

где $V(\omega)$ будет уже иметь одинаковую интенсивность на всех частотах (почему использование (9) будет корректным, будет показано ниже).

Тогда мы можем представить нашу случайную функцию в следующем виде:

$$F(x) = \frac{1}{2}\int_{R^n} V(\omega)\sqrt{\frac{S(\omega)}{d\omega}}e^{i\omega x}d\omega, \quad x,\omega \in R^n \tag{10}$$

где $S(\omega)$ – спектральная плотность, неотрицательная вещественная симметричная функция,

$V(\omega)$ – комплексная случайная функция, каждое значение которой для промежутка, соответствующего $d\omega$, является независимой случайной величиной (кроме дополнительного условия, о котором ниже), действительная и мнимая части которой распределены по нормальному закону с математическим ожиданием равным нулю и дисперсией равной единице. Т.е. $V(\omega)$ это белый шум с одинаковой дисперсией на всех частотах.

Дополнительное условие: поскольку все возможные варианты реализаций $f(x)$ должны быть вещественными, значит для любых частот ω значения $V(\omega)$ и $V(-\omega)$ должны быть комплексно сопряженными.

Чтобы отразить это условие поделим пространство частот R^n на две части R_1^n и R_2^n какой-либо гиперплоскостью размерности $n-1$, проходящей через начало координат.

Тогда $V(\omega)$ можно записать:

$$V(\omega) = \begin{cases} V_R(\omega) + iV_I(\omega), \text{если } \omega \in R_1^n \\ V_R(\omega) - iV_I(\omega), \text{если } \omega \in R_2^n \end{cases}, \qquad (11)$$

где $V_R(\omega)$ и $V_I(\omega)$ – вещественные симметричные случайные функции, каждое значение которых для промежутка, соответствующего $d\omega$ из (9), одинаково для частот ω и $-\omega$, является независимой случайной величиной с математическим ожиданием равным нулю и дисперсией равной единице (симметричный вещественный белый шум).

Тогда (10) можно записать как:

$$F(x) = \frac{1}{2}\int_{R_1^n} \left(V_R(\omega) + iV_I(\omega)\right)\sqrt{\frac{S(\omega)}{d\omega}}\, e^{i\omega x} d\omega + \\ + \frac{1}{2}\int_{R_2^n} \left(V_R(\omega) - iV_I(\omega)\right)\sqrt{\frac{S(\omega)}{d\omega}}\, e^{i\omega x} d\omega \qquad (12)$$

Поскольку для любой ω, если $\omega \in R_1^n$ то $-\omega \in R_2^n$ и наоборот, то во втором интеграле можно перейти к области интегрирования R_1^n и одновременно заменить везде ω на $-\omega$ в подынтегральном выражении.

$$F(x) = \frac{1}{2}\int_{R_1^n} \left(V_R(\omega) + iV_I(\omega)\right)\sqrt{\frac{S(\omega)}{d\omega}}\, e^{i\omega x} d\omega \\ + \frac{1}{2}\int_{R_1^n} \left(V_R(-\omega) - iV_I(-\omega)\right)\sqrt{\frac{S(-\omega)}{d\omega}}\, e^{-i\omega x} d\omega =$$

$$= \frac{1}{2}\int_{R_1^n} \sqrt{\frac{S(\omega)}{d\omega}}\left(V_R(\omega)\left(e^{i\omega x} + e^{-i\omega x}\right) + iV_I(\omega)\left(e^{i\omega x} - e^{-i\omega x}\right)\right) d\omega =$$

$$= \int_{R_1^n} \sqrt{\frac{S(\omega)}{d\omega}} \left(V_R(\omega)\cos(\omega x) - V_I(\omega)\sin(\omega x)\right) d\omega \qquad (13)$$

Таким образом мы получили интегральное каноническое представление нашей случайной функции (соответствующее записи (8)), выраженное через действительные функции и независимые действительные нормальные случайные величины с дисперсией равной единице.

Как показали преобразования, выражение (13) полностью соответствует комплексной форме (10), в которой белый шум был представлен как

$$\frac{1}{2}V(\omega)\sqrt{\frac{S(\omega)}{d\omega}}$$

Для более удобного наглядного представления дальнейших преобразований заменим интеграл (13) бесконечной суммой. Обозначим через v_j^R бесконечную последовательность значений $V_R(\omega)$ каждое из которых соответствует своему промежутку области интегрирования $\mathrm{d}\omega$ в R_1^n (ему будет соответствовать частота ω_j). Введем аналогичную последовательности v_j^I для $V_I(\omega)$ и s_j для $S(\omega)$.

Тогда (13) можно записать:

$$F(x) = \sum_{j=0}^{\infty} \sqrt{\frac{s_j}{d\omega}} \left(v_j^R \cos\left(\omega_j x\right) - v_j^I \sin\left(\omega_j x\right)\right) d\omega \qquad (14)$$

Подразумевая, что $d\omega$ это некоторая очень малая область (в пределе бесконечно малая, тогда (14) превращается в (13)).

Запишем выражение (14) чуть иначе:

$$F(x) = \sum_{j=0}^{\infty} v_j^R \sqrt{s_j d\omega}\cos\left(\omega_j x\right) - \sum_{j=0}^{\infty} v_j^I \sqrt{s_j d\omega}\sin\left(\omega_j x\right) \qquad (15)$$

Но выражение (15), это то же самое каноническое разложение случайной функции вида (7), где функция математического ожидания равна нулю, v_j^R и v_j^I это независимые случайные величины с единичной дисперсией а роль координатных функций выполняют $\sqrt{s_j d\omega}\cos\left(\omega_j x\right)$ и $-\sqrt{s_j d\omega}\sin\left(\omega_j x\right)$.

Тогда мы можем выразить корреляционную функцию (каноническое разложение через координатные функции):

$$K_f(x_1, x_2) = \sum_{j=0}^{\infty} s_j\big(\cos(\omega_j x_1)\cos(\omega_j x_2) + \sin(\omega_j x_1)\sin(\omega_j x_2)\big)d\omega$$

$$= \sum_{j=0}^{\infty} s_j \cos\Big(\omega_j(x_1 - x_2)\Big)\, d\omega \qquad (16)$$

Т.е. мы получили автокорреляционную функцию (что и следовало ожидать для стационарной случайной функции):

$$k_f(\tau) = \sum_{j=0}^{\infty} s_j \cos(\omega_j \tau)\, d\omega \qquad (17)$$

Если теперь в (17) обратно вернуться к интегралу, то получим:

$$k_f(\tau) = \int_{R_1^n} S(\omega)\cos(\omega\tau)\, d\omega \qquad (18)$$

Что является известным каноническим представлением корреляционной функции, где $S(\omega)$ – спектральная плотность.

Таким образом мы показали, что представление белого шума из интегрального канонического разложения стационарной случайной функции через $V(\omega)$ (описываемого (11)) и спектральной плотности $S(\omega)$ возможно через выражение (9), а саму случайную функцию будет корректно представить через (10) в комплексной форме или через (13) в действительной.

Теперь рассмотрим, как можно выразить наиболее вероятную реализацию случайной функции (10) (или ее же в представлении (13) или (14)), которая лучше всего соответствует обучающей выборке (исходная задача).

Поскольку все значения v_j^R и v_j^I являются независимыми случайными нормальными величинами с математическими ожиданиями равными нулю и единичной дисперсией, то функция плотности вероятности реализаций случайной функции (которую ранее обозначили как $P(f)$), выраженная через v_j^R и v_j^I будет равна перемножению всех нормальных распределений (бесконечномерный нормальный закон) каждого из них и будет пропорциональна:

$$exp\left(-\frac{1}{2}\sum_{j=0}^{\infty}\Big(\big(v_j^R\big)^2 + \big(v_j^I\big)^2\Big)\right) \qquad (19)$$

Но (15) учитывает только априорную плотность вероятности без учета u_i.

Допустим, что все случайные величины u_i в (1) имеют некоторую дисперсию σ^2. Умножим (19) на совместный нормальный закон распределения u_i, тогда получившаяся экспоненциальная часть будет:

$$exp\left(-\frac{1}{2}\sum_{j=0}^{\infty}\left(\left(v_j^R\right)^2+\left(v_j^I\right)^2\right)-\frac{1}{2\sigma^2}\sum_{i=1}^{k}u_i^2\right) \quad (20)$$

Тогда (аналогично рассуждениям с решением задачи с использованием выражения (2)) поиск наилучшей реализации в (14) или (15) будет поиском таких последовательностей v_j^R и v_j^I которые максимизируют значение (20).

Но поиск максимума (20) равносилен поиску минимума:

$$\frac{1}{2}\sum_{j=0}^{\infty}\left(\left(v_j^R\right)^2+\left(v_j^I\right)^2\right)+\frac{1}{2\sigma^2}\sum_{i=1}^{k}u_i^2\rightarrow \min \quad (21)$$

Обучающая выборка на основании (1) и (15) породит систему уравнений:

$$\begin{cases}\sum_{j=0}^{\infty}v_j^R\sqrt{s_j d\omega}\cos(\omega_j x_1)-\sum_{j=0}^{\infty}v_j^I\sqrt{s_j d\omega}\sin(\omega_j x_1)+u_1=y_1\\ \sum_{j=0}^{\infty}v_j^R\sqrt{s_j d\omega}\cos(\omega_j x_2)-\sum_{j=0}^{\infty}v_j^I\sqrt{s_j d\omega}\sin(\omega_j x_2)+u_2=y_2\\ \ldots\\ \sum_{j=0}^{\infty}v_j^R\sqrt{s_j d\omega}\cos(\omega_j x_k)-\sum_{j=0}^{\infty}v_j^I\sqrt{s_j d\omega}\sin(\omega_j x_k)+u_k=y_k\end{cases} \quad (22)$$

Получается задача минимизации (21) при системе ограничений в виде равенств (22). Ее можно решить методом множителей Лагранжа.

Функция Лагранжа получится следующая:

$$L(v_1^R,v_2^R,..,v_1^I,v_2^I,..,u_1,..,u_k,\lambda_1,..,\lambda_k)=\frac{1}{2}\sum_{j=0}^{\infty}\left(\left(v_j^R\right)^2+\left(v_j^I\right)^2\right)+\frac{1}{2\sigma^2}\sum_{i=1}^{k}u_i^2+$$

$$+\sum_{i=1}^{k}\lambda_i\left(y_i-u_i-\sum_{j=0}^{\infty}v_j^R\sqrt{s_j d\omega}\cos(\omega_j x_i)+\sum_{j=0}^{\infty}v_j^I\sqrt{s_j d\omega}\sin(\omega_j x_i)\right), \quad (23)$$

где λ_i – множители Лагранжа, количество которых равно размеру обучающей выборки.

Из условий $\frac{dL}{d\lambda_i} = 0$ получим снова систему уравнений (22).

Из условий $\frac{dL}{dv_j^R} = 0$ получим:

$$v_j^R = \sqrt{s_j d\omega} \sum_{i=1}^{k} \lambda_i \cos(\omega_j x_i) \tag{24}$$

Из условий $\frac{dL}{dv_j^I} = 0$ получим:

$$v_j^I = -\sqrt{s_j d\omega} \sum_{i=1}^{k} \lambda_i \sin(\omega_j x_i) \tag{25}$$

Из условий $\frac{dL}{du_i} = 0$ получим:

$$u_i = \sigma^2 \lambda_i \tag{26}$$

Т.е. разница, которая обозначена через случайные величины u_i, между наилучшим вариантом $f(x)$ и обучающей выборкой в (1) должна быть пропорциональна множителям Лагранжа (при решении задачи (21) и (22)). И связаны они будут через коэффициент σ^2, представляющий собой дисперсию величин u_i.

Выразим наиболее вероятную реализацию $f(x)$, подставив (24) и (25) в (15):

$$f(x) = \sum_{j=0}^{\infty} s_j \sum_{i=1}^{k} \lambda_i \cos(\omega_j x_i) \cos(\omega_j x)\, d\omega + \sum_{j=0}^{\infty} s_j \sum_{i=1}^{k} \lambda_i \sin(\omega_j x_i) \sin(\omega_j x)\, d\omega =$$

$$= \sum_{i=1}^{k} \lambda_i \sum_{j=0}^{\infty} s_j \cos\left(\omega_j (x_i - x)\right) \tag{27}$$

В (27) получилось не что иное как сумма канонических разложений корреляционной функции (которую уже получили в (16)) с множителями λ_i.

В итоге:

$$f(x) = \sum_{i=1}^{k} \lambda_i k_f(x_i - x) \tag{28}$$

Таким образом получаем, что если допустить выполнение п.1,3,4, то решением (1) должна быть линейная комбинация корреляционных функций (28).

Чтобы найти коэффициенты Лагранжа в (28) подставим (24), (25) и (26) в систему (22).

$$\begin{cases} \sum_{i=1}^{k} \lambda_i k_f(x_i - x_1) + \sigma^2 \lambda_1 = y_1 \\ \sum_{i=1}^{k} \lambda_i k_f(x_i - x_2) + \sigma^2 \lambda_2 = y_2 \\ \dots \\ \sum_{i=1}^{k} \lambda_i k_f(x_i - x_k) + \sigma^2 \lambda_k = y_k \end{cases} \quad (29)$$

Запишем (29) в матричной форме. В этом случае σ^2 прибавляется к главной диагонали:

$$(K + \sigma^2 E)\lambda = Y, \quad (30)$$

где K – квадратная матрица элементов $k_{ij} = k_f(x_i - x_j)$

E – единичная матрица

λ – вектор столбец $(\lambda_1, \lambda_2, \dots, \lambda_k)$

Y – вектор столбец $(y_1, y_2, \dots, y_k)$

Соответственно λ из (30) выражается как:

$$\lambda = (K + \sigma^2 E)^{-1} Y \quad (31)$$

Если в (31) взять $\sigma^2 = 0$, то очевидно, что задача аппроксимации (1) превращается в задачу интерполяции. Регулируя значение σ^2 можно определять степень точности соответствия $f(x)$ и обучающей выборки в (1), чтобы избежать проблему “переобучения”.

В итоге, через выражения (28) и (31) было получено решение задачи (1), однако в преобразованиях была использована спектральная плотность $S(\omega)$ (через которую выражается корреляционная функция в (18)), которую пока никак не определили.

Обозначим как $S_{f_1}(\omega)$ спектральное представление функции $f_1(x)$, которая является некоторой конкретной реализацией случайной функции (10), в виде:

$$S_{f_1}(\omega) = \frac{1}{2} V_1(\omega) \sqrt{\frac{S(\omega)}{d\omega}}, \tag{32}$$

где $V_1(\omega)$ – некоторая конкретная реализация $V(\omega)$, т.е. совокупность значений, которые $V(\omega)$ приняла, в результате чего (10) обратилась в одну из своих реализаций $f_1(x)$.

$$f_1(x) = \int_{R^n} S_{f_1}(\omega) e^{i\omega x} d\omega \tag{33}$$

Аналогично обозначим как $S_{f_2}(\omega)$ спектральное представление некоторой другой функции $f_2(x)$.

Предположим, что $f_1(x)$ и $f_2(x)$ такие, что для них выполняется п.2. (4).

$$f_2(x) = \int_{R^n} S_{f_2}(\omega) e^{i\omega x} d\omega = k f_1\left(\frac{x}{k}\right) = k \int_{R^n} S_{f_1}(\omega) e^{i\omega \frac{x}{k}} d\omega =$$

$$= \int_{R^n} k^{n+1} S_{f_1}(\omega) e^{i\omega \frac{x}{k}} d\frac{\omega}{k} = \int_{R^n} k^{n+1} S_{f_1}(\omega k) e^{i\omega x} dx \tag{34}$$

Сравнив начало и конец в (34) получим соотношение между спектральными представлениями для $f_1(x)$ и $f_2(x)$ для которых выполняется п.2:

$$S_{f_2}(\omega) = k^{n+1} S_{f_1}(\omega k) \tag{35}$$

Вернемся теперь к (32), но запишем его иначе, выразив $V_1(\omega)$:

$$V_1(\omega) = 2 S_{f_1}(\omega) \sqrt{\frac{d\omega}{S(\omega)}} \tag{36}$$

Поскольку в п.3 мы условились, что априорная плотность вероятностей таких функций $f_1(x)$ и $f_2(x)$ должна быть одинаковой, это значит, что выражение (19) для них обеих должно давать одно и то же значение.

Но выражение под суммой в (19) мы можем заменить как:

$$\left(v_j^R\right)^2 + \left(v_j^I\right)^2 = \left|V\left(\omega_j\right)\right|^2 \tag{37}$$

Квадрат модуля (36) для частоты ω_j будет:

$$\left|V_1(\omega_j)\right|^2 = 4\frac{\left|S_{f_1}(\omega_j)\right|^2}{S(\omega_j)}d\omega \tag{38}$$

Подставив (38) в (19), получим:

$$exp\left(-2\sum_{j=0}^{\infty}\frac{\left|S_{f_1}(\omega_j)\right|^2}{S(\omega_j)}d\omega\right) \tag{39}$$

Но эта сумма в (39) в пределе даст интеграл:

$$exp\left(-2\int\limits_{R_1^n}\frac{\left|S_{f_1}(\omega)\right|^2}{S(\omega)}d\omega\right) = \ exp\left(-\int\limits_{R^n}\frac{\left|S_{f_1}(\omega)\right|^2}{S(\omega)}d\omega\right) \tag{40}$$

Значит, если для $f_1(x)$ и $f_2(x)$ плотность вероятности будет одинаковой, тогда одинаковым должно быть и значение выражения (40), что выполняется, когда равны между собой интегральные его части:

$$\int\limits_{R^n}\frac{\left|S_{f_1}(\omega)\right|^2}{S(\omega)}d\omega = \int\limits_{R^n}\frac{\left|S_{f_2}(\omega)\right|^2}{S(\omega)}d\omega \tag{41}$$

Подставим в (41) найденное ранее соотношение между спектральными представлениями (35):

$$\int\limits_{R^b}\frac{\left|S_{f_1}(\omega)\right|^2}{S(\omega)}d\omega = \int\limits_{R^n}\frac{\left|S_{f_2}(\omega)\right|^2}{S(\omega)}d\omega = \int\limits_{R^n}\frac{k^{2n+2}\left|S_{f_1}(\omega k)\right|^2}{S(\omega)}d\omega =$$

$$= \int\limits_{R^n}\frac{k^{n+2}\left|S_{f_1}(\omega k)\right|^2}{S(\omega)}d(\omega k) = \int\limits_{R^n}\frac{k^{n+2}\left|S_{f_1}(\omega)\right|^2}{S\left(\frac{\omega}{k}\right)}d\omega \tag{42}$$

Сравнив начало и конец в (42) получим соотношение для спектральной плотности:

$$\frac{S(\omega/k)}{S(\omega)} = k^{n+2} \tag{43}$$

Поскольку рассматриваемая случайная функция стационарна, то автокорреляционная функция (18) должна обладать радиальной симметрией. Тогда радиальной симметрией должна обладать и ее спектральная плотность.

Такой функцией, удовлетворяющей (43) будет:

$$S(\omega) = a\|\omega\|^{-(n+2)}, \tag{44}$$

где a – некоторый коэффициент

Так как $\omega \in R^n$ и является многомерной величиной $(\omega_1, \omega_2, \dots, \omega_n)$, то (44) можно записать и так:

$$S(\omega_1, \omega_2, \dots, \omega_n) = a(\omega_1^2 + \omega_2^2 + \cdots + \omega_n^2)^{-\frac{n+2}{2}}, \omega_1, \omega_2, \dots, \omega_n \in R \tag{45}$$

Как видно, выражение (43) будет справедливо при любом коэффициенте a в (44) или (45), который можно выбрать исходя из представления, что $k_f(0)$ будет являться дисперсией случайной функции. Соотношение $k_f(0)$ и σ^2 будет определять, насколько точно $f(x)$ должна соответствовать обучающей выборке. Как видно по (31) и (28), если одновременно умножить k_f и σ^2 на некоторый множитель, то функция (28) останется неизменной.

Рассмотрим более подробно результат, получившийся в (44) и (45).

Обозначим τ в $k_f(\tau)$ в (18) как вектор $(\tau_1, \tau_2, \dots, \tau_n)$ и используя спектральную плотность (45) запишем (заодно перейдя из области интегрирования R_1^n в R^n):

$$k_f(\tau_1, \tau_2, \dots, \tau_n) = \frac{a}{2} \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \frac{\cos(\omega_1\tau_1 + \omega_2\tau_2 + \cdots + \omega_n\tau_n)}{(\omega_1^2 + \omega_2^2 + \cdots + \omega_n^2)^{\frac{n+2}{2}}} d\omega_n d\omega_{n-1} \dots d\omega_1 \tag{46}$$

Чем будет являться сечение корреляционной функции (46), например, при условии τ_n=0? Рассмотрим самый внутренний интеграл в (46) при τ_n=0.

$$\int_{-\infty}^{\infty} \frac{\cos(\omega_1\tau_1 + \omega_2\tau_2 + \cdots + \omega_{n-1}\tau_{n-1})}{(\omega_1^2 + \omega_2^2 + \cdots + \omega_n^2)^{\frac{n+2}{2}}} d\omega_n =$$

$$= 2\cos(\omega_1\tau_1 + \omega_2\tau_2 + \cdots + \omega_{n-1}\tau_{n-1}) \int_0^{\infty} \frac{d\omega_n}{(\omega_1^2 + \omega_2^2 + \cdots + \omega_n^2)^{\frac{n+2}{2}}} \tag{47}$$

Сделаем замену:

$$\frac{\omega_1^2 + \omega_2^2 + \cdots + \omega_{n-1}^2}{\omega_1^2 + \omega_2^2 + \cdots + \omega_{n-1}^2 + \omega_n^2} = 1 - t \tag{48}$$

Если выполнить преобразования, тогда (47) преобразуется в:

$$\frac{\cos(\omega_1\tau_1 + \omega_2\tau_2 + \cdots + \omega_{n-1}\tau_{n-1})}{(\omega_1^2 + \omega_2^2 + \cdots + \omega_{n-1}^2)^{\frac{(n-1)+2}{2}}} \int_0^1 t^{-\frac{1}{2}}(1-t)^{\frac{n-1}{2}} dt =$$

$$= \frac{\cos(\omega_1\tau_1 + \omega_2\tau_2 + \cdots + \omega_{n-1}\tau_{n-1})}{(\omega_1^2 + \omega_2^2 + \cdots + \omega_{n-1}^2)^{\frac{(n-1)+2}{2}}} B\left(\frac{1}{2}, \frac{n+1}{2}\right), \quad (49)$$

где B – бета-функция Эйлера

Тогда сечение (46) при τ_n=0 будет:

$$k_f(\tau_1, \dots, \tau_{n-1}) =$$
$$= \frac{a}{2} B\left(\frac{1}{2}, \frac{n+1}{2}\right) \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \frac{\cos(\omega_1\tau_1 + \cdots + \omega_n\tau_{n-1})}{(\omega_1^2 + \omega_2^2 + \cdots + \omega_{n-1}^2)^{\frac{(n-1)+2}{2}}} d\omega_{n-1} \dots d\omega_1 \quad (50)$$

Поскольку $\frac{a}{2} B\left(\frac{1}{2}, \frac{n+1}{2}\right)$ играет роль некоторого коэффициента, сравнивая (50) и (46) можно сделать вывод, что если для корреляционной функции в пространстве R^n провести через начало координат сечение размерностью R^{n-1}, то оно будет эквивалентно корреляционной функции в пространстве R^{n-1} выраженное снова через (46) (отличаясь лишь коэффициентом a в (44)-(45)).

Отсюда можно сделать вывод, чтобы найти корреляционную функцию, достаточно будет определить её для одномерного случая на промежутке $\tau \in [0, +\infty)$, а затем пользуясь свойством ее радиальной симметрии, определить ее для пространства любой размерности.

Спектральная плотность (44) (если взять $a = 1$) для одномерного случая будет:

$$S(\omega) = |\omega|^{-3}, \omega \in R \quad (51)$$

Случайная функция, выраженная в виде (13) в одномерном случае примет вид:

$$F(x) = \int_0^{\infty} \frac{V_R(\omega)\cos(\omega x) - V_I(\omega)\sin(\omega x)}{\omega\sqrt{\omega d\omega}} d\omega, \qquad x, \omega \in R \quad (52)$$

Корреляционная функция (18) в одномерном случае:

$$k_f(\tau) = \int_0^{\infty} \frac{\cos(\omega\tau)}{\omega^3} d\omega \quad (53)$$

Модуль от частоты в (52) и (53) можно убрать, поскольку интегрирование в будет только в положительной области.

Однако в (52) и (53) видно, что при приближении частоты ω к нулю выражение под интегралом стремится к бесконечности.

Возьмем некоторое очень малое значение $\omega_0 > 0$ и рассмотрим случайную функцию, аналогичную (52), но у которой (у всех ее реализаций) частоты меньше ω_0 отсутствуют.

Ее выражение будет отличаться от (52) только нижним пределом интегрирования:

$$F(x) = \int_{\omega_0}^{\infty} \frac{V_R(\omega)\cos(\omega x) - V_I(\omega)\sin(\omega x)}{\omega\sqrt{\omega d\omega}} d\omega \tag{54}$$

Корреляционная функция для нее будет:

$$k_f(\tau) = \int_{\omega_0}^{\infty} \frac{\cos(\omega\tau)}{\omega^3} d\omega \tag{55}$$

Гармоническому колебанию с частотой ω_0 будет соответствовать период:

$$T_0 = \frac{2\pi}{\omega_0} \tag{56}$$

Пусть ω_0 взято такое, что период T_0 многократно больше, чем диапазон значений x из обучающей выборки. Тогда все колебания частот, меньших чем ω_0, внутри этой области будут вырождаться в линейные слагаемые и мало отличаться друг от друга и от частот несколько больших, чем ω_0.

Поэтому можно ожидать, что если значения обучающей выборки x_i ограничены некоторым конечным диапазоном, то всегда можно взять такое малое ω_0, что (54) и (55) будет допустимой с практической точки зрения заменой (52) и (53), а при устремлении ω_0 к нулю они становятся эквивалентными.

Выполним дважды интегрирование (55) по частям:

$$k_f(\tau) = \int_{\omega_0}^{\infty} \frac{\cos(\omega\tau)}{\omega^3} d\omega = -\frac{\cos(\omega\tau)}{2\omega^2}\Bigg|_{\omega_0}^{\infty} - \tau \int_{\omega_0}^{\infty} \frac{\sin(\omega\tau)}{2\omega^2} d\omega =$$

$$= \left(-\frac{\cos(\omega\tau)}{2\omega^2} + \tau\frac{\sin(\omega\tau)}{2\omega}\right)\Bigg|_{\omega_0}^{\infty} - \frac{\tau^2}{2}\int_{\omega_0}^{\infty} \frac{\cos(\omega\tau)}{\omega} d\omega \tag{57}$$

Одним из множителей в последнем слагаемом в (57) получился интегральный косинус.

Рассмотрим его более подробно (принимая во внимание, что рассматриваем лишь $\tau \geq 0$).

$$-\int_{\omega_0}^{\infty} \frac{\cos(\omega\tau)}{\omega} d\omega = -\int_{\omega_0\tau}^{\infty} \frac{\cos(\omega\tau)}{\omega\tau} d(\omega\tau) =$$

$$= \gamma + \ln(\omega_0\tau) + \int_0^{\omega_0\tau} \frac{\cos(\omega) - 1}{\omega} d\omega \,, \tag{58}$$

где γ – постоянная Эйлера-Маскерони

Тогда (57) преобразуется в:

$$\frac{\cos(\omega_0\tau)}{2\omega_0^2} - \tau^2 \frac{\sin(\omega_0\tau)}{2\omega_0\tau} + \frac{1}{2}\tau^2 \left(\gamma + \ln(\tau) + \ln(\omega_0) + \int_0^{\omega_0\tau} \frac{\cos(\omega) - 1}{\omega} d\omega \right) =$$

$$= \frac{1}{2}\tau^2 \left(\ln(\tau) + \left(\ln(\omega_0) + \gamma + \int_0^{\omega_0\tau} \frac{\cos(\omega) - 1}{\omega} d\omega - \frac{\sin(\omega_0\tau)}{2\omega_0\tau} \right) \right) +$$

$$+ \frac{\cos(\omega_0\tau)}{2\omega_0^2} \tag{59}$$

Поскольку, как уже рассмотрели выше, $k_f(\tau)$ мы можем умножить на произвольный коэффициент, а важно лишь соотношение $k_f(0)$ и σ^2, то в (59) мы можем убрать $\frac{1}{2}$, а (59) записать как:

$$k_f(\tau) = \tau^2\big(\ln(\tau) - b(\tau)\big) + c(\tau), \tag{60}$$

где

$$b(\tau) = \frac{\sin(\omega_0\tau)}{\omega_0\tau} - \ln(\omega_0) - \gamma - \int_0^{\omega_0\tau} \frac{\cos(\omega) - 1}{\omega} d\omega \tag{61}$$

$$c(\tau) = \frac{\cos(\omega_0\tau)}{\omega_0^2} \tag{62}$$

Если устремить ω_0 к нулю, то $b(\tau)$ и $c(\tau)$ устремятся к бесконечности.

$$\lim_{\omega_0 \to +0} b(\tau) \to +\infty,$$

$$\lim_{\omega_0 \to +0} c(\tau) \to +\infty$$

А случайная функция, описываемая (54) и (55) будет превращаться в случайную функцию, описываемую (52) и (53). Хотя видно, что $c(\tau)$ будет устремляться к бесконечности гораздо быстрее, чем $b(\tau)$, значение которого будет в основном определяться логарифмом $(-\ln(\omega_0))$.

Но если взять ω_0 достаточно малым, чтобы период (56) был значительно больше, чем диапазон изменения τ, то $b(\tau)$ и $c(\tau)$ фактически превращаются в константы, а в (60) мы получим корреляционную функцию, которую удобно использовать для вычислений.

Таким образом мы получили автокорреляционную функцию (60) для одномерного случая и $\tau \in [0, +\infty)$. Ее легко обобщить на многомерный случай, поскольку она должна обладать радиальной симметрией.

Получаем итоговый результат:

Автокорреляционная функция:

$$k_f(\tau) = \tau^2(\ln(\|\tau\|) - b) + c\ , \tau \in R^n \tag{63}$$

где b и c – константы, которые можно оценить по формулам (61) и (62)

Выпишем еще раз полученные формулы (28) и (31).

Решением задачи аппроксимации (1) будет функция:

$$f(x) = \sum_{i=1}^{k} \lambda_i k_f(x_i - x) \tag{64}$$

Коэффициенты λ_i определяются системой уравнений:

$$\lambda = (K + \sigma^2 E)^{-1} Y \tag{65}$$

где K – квадратная матрица элементов $k_{ij} = k_f(x_i - x_j)$

E – единичная матрица

λ – вектор столбец $(\lambda_1, \lambda_2, \dots, \lambda_k)$

Y – вектор столбец $(y_1, y_2, \dots, y_k)$

σ^2 – дисперсия случайных величин u_i из (1)

Полученная линейная комбинация (64) функций (63) известна также как полигармонический сплайн (или его разновидность, сплайн тонкой пластины (thin plate spline) [2] и [3]). Но в данном случае есть нюансы использования, такие как использование коэффициентов b и c, которые оцениваются через

(61) и (62), прямое введение σ^2 случайных величин u_i в систему уравнений (65), чтобы решить задачу аппроксимации.

Пример:

В качестве наглядного примера рассмотрим одномерный случай аппроксимации.

Возьмем обучающую выборку, у которой большая часть входных значений лежит в интервале от 0 до 10, т.е. для оценки значений b и c примем, что τ изменяется от 0 до 10.

Возьмем $\omega_0 = 0.001$. Таким образом соответствующий этой частоте период (56) будет равен $T_0 = 6283.1853$, что примерно в 600 раз больше, чем изменения между значениями x_i. Вполне можно допустить, что наличие частот меньших ω_0 (с еще большим периодом) для решения задачи в нашей области аппроксимации не имеет существенного значения.

Сравним значения $b(\tau)$ и $c(\tau)$ из (61) и (62) при выбранной $\omega_0 = 0.001$ и значениях $\tau = 0$ и $\tau = 10$.

$\mathrm{b}(0) = 1 + 6.907755 - 0.577216 + 0 = 7.33054$

$b(10) = 0.999983 + 6.907755 - 0.577216 + 0.000025 = 7.330548$

$c(0) = 1000000$

$c(10) = 999950.000416$

Значения $b(\tau)$ и $c(\tau)$ при $\tau = 10$ относительно их же при $\tau = 0$ изменились на 0.0001% и 0.005% соответственно. Поэтому возьмем в качестве них просто константы $\mathrm{b} = 7.33054$ и $c = 1000000$.

Попробуем сначала взять $\sigma^2 = 0$ в (65). В этом случае все u_i в (1) становятся равными нулю, $f(x)$ должна точно без ошибок пройти через все точки из обучающей выборки. Задача аппроксимации превращается в задачу интерполяции. Что и наблюдается на рисунке 1.

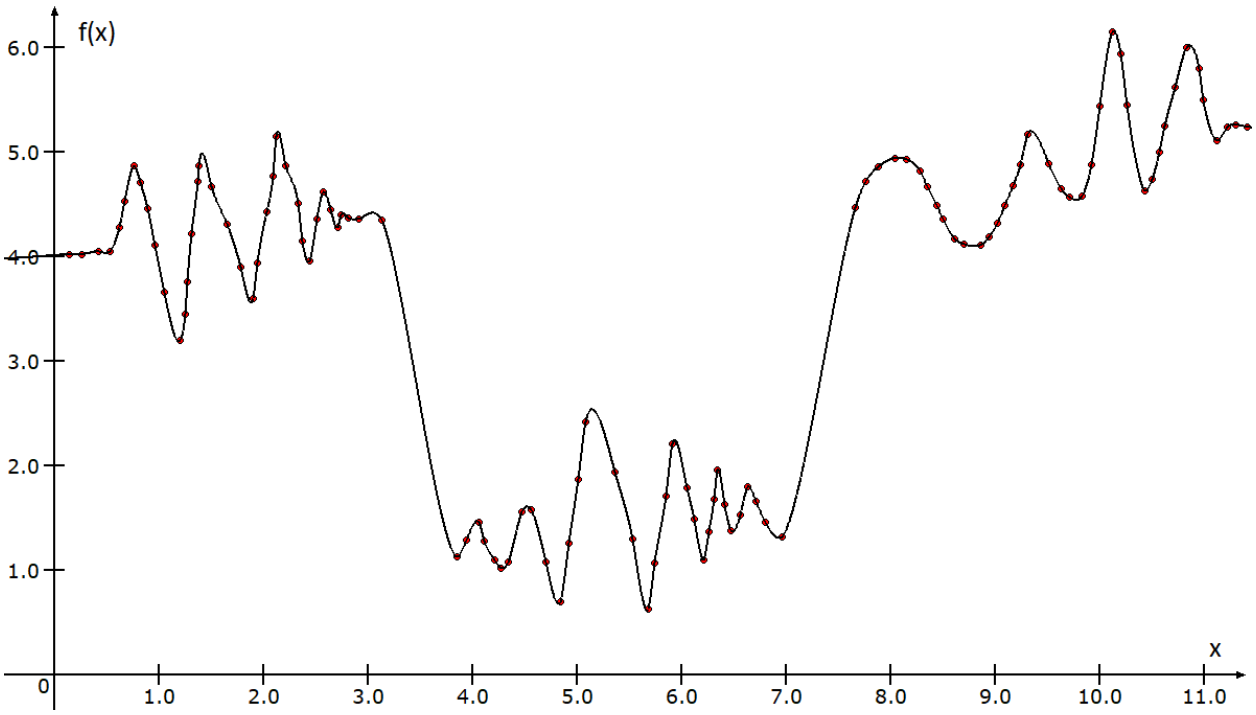

Рисунок 1.

Функция (64) легко воспроизводит любую сложную нелинейность без каких-либо скачков или осцилляций между соседними точками.

Если взять $\sigma^2 > 0$, то переходим к исходной задаче аппроксимации (1).

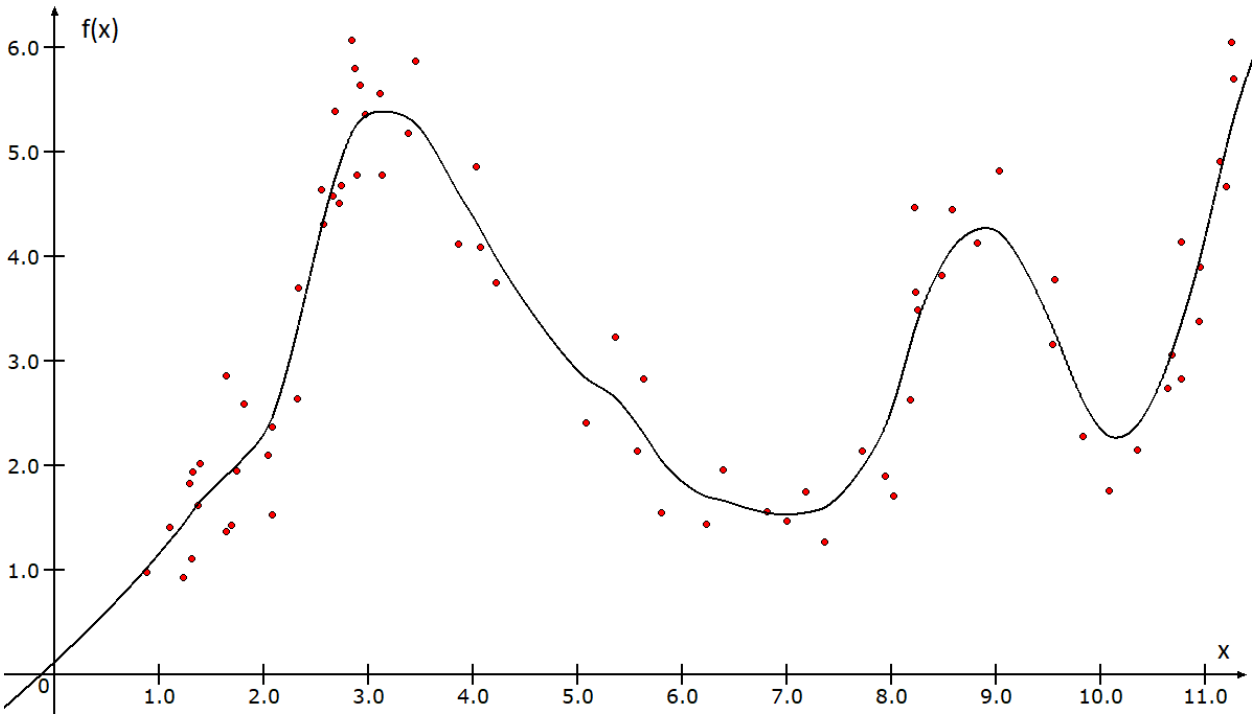

Рисунок 2.

В данном случае, изображенном на рисунке 2, $\sigma^2 = 0.625$.

Заключение.

Таким образом было показано, что регрессионная задача машинного обучения как задача многомерной аппроксимации, выраженная (1), может быть решена на основе теории случайных функций, если предположить существование в функциональном пространстве некоторых симметрий плотности вероятности (3)-(6). Решением будет линейная комбинация корреляционной функции, спектральная плотность которой (44)-(45) также может быть выведена из одной из симметрий (4). Далее показано, что корреляционной функции с такой плотностью с некоторыми нюансами соответствует полигармоническому сплайну в виде (63). И как итог, выражения (61)-(65) для решения задачи (1).

Список литературы:

1. В.С. Пугачев, Теория случайных функций и её применение к задачам автоматического управления. Изд. 2-ое, перераб. и допол. — М.: Физматлит, 1960.
2. R.L. Harder and R.N. Desmarais: Interpolation using surface splines. Journal of Aircraft, 1972, Issue 2, pp. 189−191
3. Bookstein, F. L. (June 1989). "Principal warps: thin plate splines and the decomposition of deformations". IEEE Transactions on Pattern Analysis and Machine Intelligence. 11 (6): 567–585. doi:10.1109/34.24792
4. http://www.machinelearning.ru/wiki/index.php?title=Многомерная_интерполяция_и_аппроксимация_на_основе_теории_случайных_функций