С. А. ТЕРЕХОВ
ООО «Нейрок Техсофт»,
г. Троицк, Московская обл.
E-mail: alife@narod.ru

СЛУЧАЙНЫЕ ГАУССОВСКИЕ ПРОЦЕССЫ В ЗАДАЧАХ АППРОКСИМАЦИИ ДАННЫХ

Аннотация

Тематика лекции связана с актуальной задачей аппроксимации данных. Особое внимание уделено гауссовым процессам, моделирующим условную плотность вероятности. Методы описания данных гауссовыми процессами позволяют надежно оценивать неопределенность (риск) прогноза на новых данных. Основное внимание уделено методам приближения данных гауссовскими процессами и тесно связанным с ними оценкам величины риска прогноза, определяемого такими приближениями. Алгоритмы ориентированы на коллекции данных малых и средних размеров (до 10000 примеров) и могут эффективно применяться при планировании экспериментов.

S. A. TEREKHOFF
Neurok Techsoft, LLC,
Troitsk, the Moscow Region
E-mail: alife@narod.ru

RANDOM GAUSSIAN PROCESSES IN DATA APPROXIMATION PROBLEMS

Abstract

This Lecture discusses Gaussian process approximations of conditional probability densities. Methods are oriented to data regression task for mid-size sample collections. The possibility of estimate both expectation and uncertainty is well suited for design of experiments (DoE) and planning of data acquisition process.

УДК 001(06)+004.032.26(06) Нейронные сети

О степени зрелости отрасли знаний можно судить по тому, насколько в ней апостериорные распределения отличаются от априорных. При совпадении — знание заканчивается.

Введение: Практика байесова обучения машин

Обучение машины (алгоритма, программы, робота) содержательно основывается на классических понятиях и методах теории вероятности. Машина формирует представление о распределении данных и способна принимать самостоятельные решения, пользуясь оценкой условной вероятности оптимальности этих решений, при заданных или наблюдаемых внешних условиях.

Байесово исчисление вероятностей является надежным формальным способом оперирования вероятностями сложных событий. В байесовом подходе естественным образом сочетаются *оперативные навыки* ситуационного обучения (в форме функции правдоподобия) и *долговременные «знания»* (в форме априорных вероятностей).

Современная практика разработки и использования обучающихся алгоритмов, как правило, основывается на оценивании распределений дискретных наборов параметров моделей. Таковы [3] искусственные нейронные сети, машины опорных векторов (SVM), вероятностные деревья и комитеты логических или ассоциативных (в том числе, нечетких) решающих правил.

Крайне привлекательной при оценивании вероятностей была бы возможность *проводить вычисления непосредственно с функциями* многих переменных, минуя их дискретное параметрическое представление. Таким математическим аппаратом является теория стохастических процессов и полей [1]. Стохастический (или случайный) процесс [2] обобщает понятие одной случайной переменной на совокупность (конечную или бесконечную) зависимых случайных переменных, упорядоченных дискретным или непрерывным «индексным» множеством. Множество может представлять собой, например, действительную прямую («время»), область в многомерном числовом пространстве («объем»), а также иной запас элементов (при выполнении определенных требований на его измеримость).

Отдельной реализацией случайной переменной является *число* (если соответствующая вероятность определена над числовым множеством). Реализацией же случайного процесса является *траектория*, которая являет-

ся *функцией*, определенной на элементах индексного множества. В дальнейшем изложении мы ограничимся случаем стохастических процессов и случайных функций, определенных в некоторой области пространства $\mathbb{R}^N$ содержащей векторы обучающих данных.

Стохастический процесс полностью определяется совокупностью совместных распределений всех конечных наборов его случайных переменных. Для частного случая *гауссовских процессов*, все одномерные распределения которых являются нормальными, программу байесового оценивания условных вероятностей удается провести последовательно до конца. Необходимые вычисления сводятся к стандартным задачам линейной алгебры, для которых хорошо приспособлены современные компьютеры.

Особенностью гауссовских процессов является возможность их однозначного задания указанием действительной функции математического ожидания и симметричной неотрицательно определенной действительной функции, описывающей ковариацию пары переменных. Моменты распределений более высоких порядков восстанавливаются по этой паре функций.

Наиболее полно теоретически проработана задача аппроксимации (регрессии) набора данных. В терминах условных гауссовских процессов эта задача понимается, как поточечное оценивание математического ожидания для вероятностного распределения функций — траекторий процесса. При заданной ковариационной функции, оценка математического ожидания однозначно восстанавливается по конечному набору обучающих данных в произвольной тестовой точке. Рассмотрением задачи регрессии мы и ограничимся в этой Лекции.

Дальнейшее изложение построено по следующему плану. Вначале рассматривается постановка задачи регрессии и предлагается базовый алгоритм ее решения. Далее обсуждаются вопросы, связанные с выбором оптимальных ковариационных функций. В завершении рассматривается содержательная интерпретация моделей условных гауссовских процессов — оценивание неопределенности и риска, информационная значимость переменных, а также возможные применения моделей в задачах планирования экспериментов.

Интерес автора к тематике условных гауссовских процессов во многом вызван появлением книги *Расмуссена* и *Вильямса* [4], в которой суммированы основные теоретические результаты и концептуальные вычислительные алгоритмы. По существу, данная лекция является вводной интерпретацией материалов [4] и других публикаций, с общим, традиционным для лекций автора, уклоном в прикладные аспекты.

Задача аппроксимации (регрессии) данных

Исходным материалом для задачи аппроксимации (и для более широкого класса задач обучения) является набор числовых данных

$$D = \{(\vec{x}_i, y_i);\, 1, \ldots, n\} = (X, \vec{y}), \tag{1}$$

которые являются примерами некоторой (неизвестной) функциональной зависимости $z(\vec{x})$. На практике нет возможности наблюдать точные значения $z(\vec{x})$, поскольку данные всегда содержат некоторый информационный шум. Строго говоря, скрытый от наблюдателя процесс порождения данных D может и не предполагать наличие зависимости $z(\vec{x})$. Например, ставки игрока в покер могут не являться функцией от набора имеющихся на руке карт, а отражать иные факторы (настроение, размер свободной суммы, характер поведения соперников и т. д.). Мы, тем не менее, будем в данном контексте считать, что искомая зависимость $z(\vec{x})$ носит причинно-следственный характер и существует. Целью задачи аппроксимации является нахождение *модели* $f(\vec{x})$ этой зависимости, мало уклоняющейся от имеющихся данных:

$$|y(\vec{x}) - f(\vec{x})| < \varepsilon. \tag{2}$$

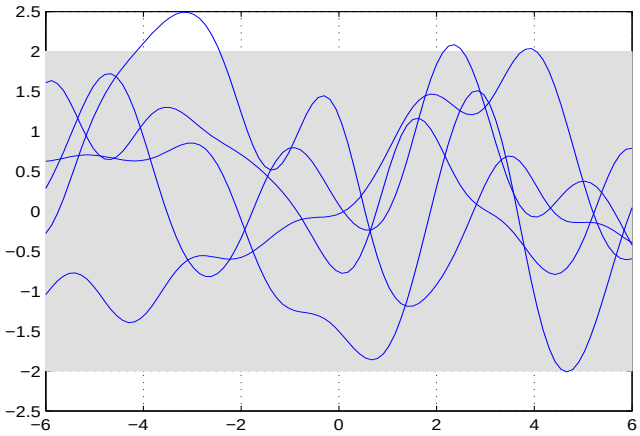
Нас интересует такая, в некотором смысле, наилучшая модель, для которой уклонение от *будущих* данных также мало. Количественное требование состоит в необходимости минимизации выбранного функционала *риска* ошибок (или функции потерь) $\mathrm{Err}(y; f)$. Отметим, что функциональный характер риска никак не связан с проблемой моделирования данных — он всецело определяется особенностями конкретного приложения.

Ожидаемый риск связан с функцией потерь обычным соотношением:

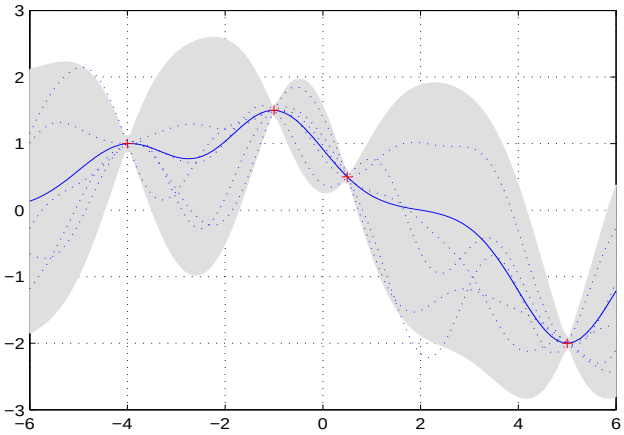
$$R(f(\vec{x}_0); \vec{x}_0))| = \int \mathrm{Err}(y; f) p(y|\vec{x}_0, D) dy. \tag{3}$$

Оптимальным решением в условиях, когда задан новый вектор переменных $\vec{x}_0$, является выбор такого значения $f(\vec{x}_0)$, которое минимизирует риск R в (3).

К прерогативе моделирования данных относится нахождение $p(y|\vec{x}, D)$ — распределения условной вероятности значений выхода y при известных входах $\vec{x}$ и факте наличия набора ранее наблюдавшихся данных D. Это распределение вероятностей объясняет уклонения ε регрессионной модели от данных в формуле (2).



(a)



(b)

Рис. 1. Априорное (*вверху*) и апостериорное (*внизу*) распределение траекторий одномерного гауссовского процесса для четырех обучающих примеров

Различают два подхода к построению регрессионных моделей, направленных на уменьшение риска, связанного с *переобучением* — запоминанием обучающих данных при слабой способности к обобщению. Первый, традиционно используемый в обучении машин, состоит в априорном ограничении класса функций, представимых моделью, и дальнейшем поиске лучшей модели из класса. При этом подходе неизбежна дилемма выбора между слишком гибким, богатым классом функций, либо излишне упрощенными функциями, не способными представить данные.

При втором подходе потенциально допустимыми являются *любые* функции, но с различными априорными вероятностями. Более «гладкие» функции получают более высокую исходную вероятность участия в модели. В дальнейшем, оценивается функция правдоподобия объяснения моделью обучающих данных, которая вместе с априорным распределением приводит к апостериорному распределению, максимум которого и указывает на искомую модель. Это достигается в методологии гауссовских процессов.

Упрощенно процесс отбора функций приведен на рис. 1. Априорное распределение допускает широкий класс функций, при этом распределение вероятности в каждой точке нормально. Если имеются обучающие данные, то в апостериорном распределении участвуют только функции, им не противоречащие. Искомая регрессионная зависимость, на качественном языке, соответствует математическому ожиданию траекторий, распределенных в соответствии с условной апостериорной вероятностью.

Возможность выбора из бесконечного числа возможных функций путем указания конечного числа обучающих точек, на первый взгляд, представляется крайне удивительной. В действительности, распределение траекторий гауссовского процесса оценивается лишь в конечном наборе тестовых (новых) точек, таким образом, континуум оказывается вне поля рассмотрения.

Конечно, «наивная» процедура отбора траекторий из окрестности обучающих данных далека и от практики, и, собственно, от теории байесовых вероятностей, однако она на «карикатурном» уровне поясняет суть вероятностного вывода.

Последовательно усложняя задачу, начнем рассмотрения со случая идеальных данных.

Данные без шума

Перейдем теперь к более формальному рассмотрению проблемы оценивания регрессионной аппроксимации данных. Гауссовский процесс f_x есть

запас случайных переменных, любое конечное подмножество которых имеет совместное гауссовское распределение. Можно показать [2], что действительный гауссовский процесс полностью определяется указанием его математического ожидания (среднего) $m(x)$ (функции, сопоставляющей каждой из случайных переменных ее математическое ожидание) и ковариационной функции $K(x, x')$:

$$m(x) = E[f_x], \tag{4a}$$

$$K(x, x') = \mathrm{cov}(f_x, f_{x'}) = E[(f_x - m(x))(f_{x'} - m(x'))], \tag{4b}$$

где через E обозначена операция взятия среднего, а x и x' — «номера» случайных переменных. В зависимости от выбранного множества случайных переменных, это могут быть, например, моменты времени (для одномерного процесса), либо пространственные координаты для процессов[1] большей размерности, или какой-то иной способ перечисления переменных. В вычислениях мы будем предполагать, что переменные образуют некоторый дискретный набор (узлы сетки), хотя все формальные построения справедливы и для континуумов. В дальнейшем будем понимать под x просто пространственные координаты случайной переменной.

Реализации процесса f_x это функции $f(x)$, называемые траекториями процесса. Значение функции $f(x)$ (f в точке x) есть реализация случайной переменной с координатой x. Для дискретного набора из n переменных m и f есть векторы размерности n, а K — матрица размерности $n \times n$.

Для i-го примера из набора обучающих данных (1) экспериментальное наблюдение есть y_i, ему соответствует реализация случайной переменной с координатами x_i, равная $f(x_i)$.

Выражение $f_x \sim GP(m(x), K(x, x'))$ означает, что функция $f(x)$ является реализацией гауссовского процесса GP, с параметрами $m(x)$ и $K(x, x')$.

Ограничимся в этом разделе конкретным законом ковариаций в виде гауссиана:

$$K(x, x') = \exp\bigl(-\frac{1}{2}|x - x'|^2\bigr). \tag{5}$$

Ковариационная функция определена для пары случайных переменных, которым в гауссовском процессе приписаны пространственные координаты. Поэтому окончательное выражение записано в виде функции координат. Выбор ковариационной функции определяет характер распределения функций, являющихся реализациями гауссовского процесса.

[1] Многомерные случайные процессы также называются *случайными полями*.

Рассмотрим для наглядности случай одной переменной. Вычисления могут быть проведены для конечного (но произвольного) набора точек, в качестве которого выберем дискретную сетку

$$X = \{x_i = h(i - n/2),\ i = 1, \ldots, n\}$$

на отрезке фиксированной длины. Траектория центрированного процесса на этой выборке точек есть набор реализаций нормально распределенных переменных:

$$\vec{f} = [f_{x_1}, f_{x_2}, \ldots, f_{x_n}] \sim N(\vec{0}, K(X, X)).$$

Здесь (и далее в контексте конкретных вычислений) использованы матричные обозначения: X и $\vec{f}$ — наборы размера n, $K(X, X) = K$ — матрица размерности $n \times n$.

Для получения реализации вектора нормальных случайных чисел с заданной ковариационной матрицей используется следующий алгоритм:

Алгоритм 1. Реализация нормального случайного вектора с заданным набором средних $\vec{m}$ и ковариационной матрицей K.

1. Вычислить разложение Холесского (симметричной, положительно определенной) матрицы $K = LL^T$, где L — нижняя треугольная матрица ([5], с. 134).
2. При помощи датчика случайных чисел получить вектор *независимых* нормально распределенных чисел $\vec{u}$, с нулевым средним и единичной дисперсией.
3. Вычислить искомый вектор коррелированных случайных чисел $\vec{f} = \vec{m} + L\vec{u}$.

Примеры функций на дискретной сетке, вычисленных по этому алгоритму, приведены на рис. 1а. Ковариационная функция является бесконечно дифференцируемой, что обеспечивает бесконечную дифференцируемость траекторий процесса (с вероятностью единица). Одно из важных наблюдаемых свойств выборочных функции состоит в наличии характерного масштаба изменений («длины волны»), равного в данном примере единице.

Перейдем теперь к задаче об условных вероятностях. Рассмотрим два набора точек $X = \{x_i,\ i = 1, \ldots, n\}$ и $X^* = \{x^*_i,\ i = 1, \ldots, n^*\}$, и соответствующие им наборы случайных переменных $\vec{f} = \{f_i = f_{x_i},\ i = 1, \ldots, n\}$ и $\vec{f}^* = \{f^*_i = f_{x^*_i},\ i = 1, \ldots, n^*\}$ из числа включенных в гауссовский процесс. Их совместное априорное (безусловное) распределение, по-прежнему,

является нормальным, с блочной ковариационной матрицей:

$$\begin{bmatrix} \vec{f} \\ \vec{f}^* \end{bmatrix} \sim N\left(\vec{0}, \begin{bmatrix} K(X,X) & K(X,X^*) \\ K(X^*,X) & K(X^*,X^*) \end{bmatrix}\right). \tag{6}$$

Наложим теперь условия идеальной, без шума, наблюдаемости значений $\vec{f}$. Другими словами, будем считать, что пары (x_i, f_i) известны и являются *обучающими* данными. Условное распределение неизвестных *тестовых* данных $\vec{f}^*$, после матричных преобразований [5], принимает вид:

$$\left(\vec{f}^* | X^*, X, \vec{f}\right) \sim N(\vec{m}^*, K^*), \tag{7a}$$

$$\vec{m}^* = K(X^*,X)K(X,X)^{-1}\vec{f}, \tag{7b}$$

$$K^* = K(X^*,X^*) - K(X^*,X)K(X,X)^{-1}K(X,X^*). \tag{7c}$$

Вычисление условных вероятностей проводится путем суммирования в базовом соотношении для полной вероятности:

$$p(\vec{f}^*|\vec{f}) = \frac{p(\vec{f}^*, \vec{f})}{\sum_{\vec{f}^*} p(\vec{f}^*, \vec{f})}.$$

Все матрицы и векторы в правых частях выражений (7) известны. Поскольку результирующее условное распределение также гауссово, для получения выборки можно воспользоваться Алгоритмом 1.

Заметим, что гауссовский процесс с ковариационной функцией (5) формально эквивалентен байесовой линейной регрессии по бесконечному набору фиксированных базисных функций. Этот факт устанавливается теоремой Мерсера (Mercer's theorem, строгая формулировка приведена в [2], с. 261), согласно которой для симметричных положительно определенных непрерывных функций

$$K(x,x') = \sum_{i=1}^{\infty} \lambda_i \phi_i(x)\phi_i(x'),$$

где пара (λ_i, ϕ_i) — i-е собственное значение и собственная функция интегрального уравнения с ядром K:

$$\int K(x,x')\phi(x)dx = \lambda\phi(x').$$

Все рассмотренные формулы и алгоритмы автоматически применимы и к случаю многомерных данных: для этого требуется лишь корректно вычислять попарные расстояния между векторами в ковариационной матрице (5).

Приведенное в данном разделе решение задачи регрессии имеет два практических недостатка. Первый заключается в наличии шума в реальных обучающих данных. Второй связан с произвольностью выбора ковариационной функции, и, как следствие, произвольностью длины волны корреляций, которая, при таком выборе, не обязана соответствовать характеру данных. Это, например, наблюдается в характере регрессионной зависимости на рис. 1b. Далее в Лекции обсуждаются обобщения, призванные исправить ситуацию.

Данные с шумом

Реальные данные наблюдаются с шумом (2). Характер шума может существенно зависеть от специфики прикладной задачи. В Лекции мы ограничимся случаем, когда значения входов (координат случайных переменных процесса) $\vec{x}$ известны *точно*, а вся неопределенность отнесена к выходам, которые нужно аппроксимировать распределением траекторий f. Будем также считать, что систематические факторы устранены, и нормально распределенные ошибки в наблюдениях не коррелированы друг с другом.

В этих условиях ковариационная функция наблюдений из набора $\vec{y} = \{y_i,\ i = 1, \ldots, n\}$ есть:

$$\mathrm{cov}(y_i, y_j) = \mathrm{cov}(f_{x_i}, f_{x_j}) + \sigma^2 \delta_{ij} = K(x_i, x_j) + \sigma^2 \delta_{ij}, \tag{8}$$

где δ_{ij} — символ Кронекера. Правая часть выражения (8) зависит от индексов только наблюдаемых случайных переменных, поэтому совместное распределение тестовых и обучающих переменных дается формулой (6) с заменой:

$$K = K(X, X) \to K(X, X) + \sigma^2 I = K + \sigma^2 I, \tag{9}$$

где I — единичная матрица размера $n \times n$. Ясно, что и апостериорное распределение (7) также сохраняет свой вид с заменой (9). Приведем это распределение для практически важного случая одной тестовой точки x^*:

$$m^* = \vec{k}^{*T} (K + \sigma^2 I)^{-1} \vec{y}, \tag{10a}$$

$$D^* = K(x^*, x^*) - \vec{k}^{*T} (K + \sigma^2 I)^{-1} \vec{k}^*. \tag{10b}$$

Здесь использовано обозначение $\vec{k}^*$ — вектор ковариаций x^* со всеми обучающими примерами X. Заметим, что уровень шума σ пока является свободным параметром.

Дисперсия (variance) прогноза (10b) не зависит от значений наблюдаемых переменных, а определяется только их ковариациями с тестовой точкой (т. е. для пространственного случая, имеется зависимость только от относительных координат обучающих и тестовых примеров).

Интересен характер зависимости математического ожидания (10a). Здесь зависимость от выходов обучающих примеров линейна — т. е. прогноз есть линейная комбинация известных наблюдений. С другой стороны, он может рассматриваться, как аппроксимация ядрами:

$$f(x^*) = \sum_{i=1}^{n} \alpha_i K(x_i, x^*), \tag{11}$$

где $\vec{\alpha} = (K+\sigma^2 I)^{-1}\vec{y}$. Таким образом, параметры регрессии на классе гауссовских процессов выражаются либо прямыми коэффициентами линейной модели, либо дуальными коэффициентами аппроксимации, аналогичной машинам опорных векторов (SVM). В отличие от SVM здесь в разложении участвуют все обучающие вектора, а не только опорные.

Отметим следующее важное обстоятельство. При одновременном анализе *нескольких* тестовых точек оцененная плотность распределения отражает не только дисперсию (как в (10b)), но и взаимную ковариацию тестовых данных, что характеризует их взаимную информационную зависимость.

Для завершения вопроса об условном распределении траекторий в тестовой точке при зашумленных обучающих наблюдениях, получим выражение для вероятности объяснения обучающих данных (evidence) гауссовским процессом. Эта вероятность, участвующая в знаменателе формулы Байеса, играет решающую роль в следующем разделе Лекции, при ответе на второй из поставленных проблемных вопросов — какая модель наиболее адекватна имеющимся обучающим данным?

Вероятность объяснения данных $p(\vec{y}\,|X)$ по определению есть сумма по всем реализациям процесса от произведения априорного распределения (prior) и функции правдоподобия (likelihood):

$$p(\vec{y}\,|X) = \int p(\vec{y}\,|\vec{f}, X)p(\vec{f}\,|X)d\vec{f}. \tag{12}$$

Здесь, как и ранее, векторы означают перечень случайных переменных, отвечающих обучающим примерам. Априорная вероятность $p(\vec{y}\,|X)$ — гауссовская:

$$\ln\bigl[p(\vec{f}\,|X)\bigr] = -\frac{1}{2}\Bigl(\vec{f}^{\,T}K^{-1}\vec{f} + \ln[\det(K)] + n\ln[2\pi]\Bigr), \tag{13}$$

где $\det(\cdot)$ — операция взятия определителя матрицы. Функция правдоподобия описывает, в нашем случае, независимые компоненты шума:

$$p(\vec{y}\,|\vec{f}, X) \sim N(\vec{f}, \sigma^2 I). \tag{14}$$

Интеграл от произведения гауссовских функций вычисляется точно, и результат имеет ожидаемый вид гауссовского распределения:

$$\ln\bigl[p(\vec{y}\,|X)\bigr] = -\frac{1}{2}\Bigl(\vec{y}^{\,T}(K+\sigma^2 I)^{-1}\vec{y} + \ln[\det(K+\sigma^2 I)] + n\ln[2\pi]\Bigr). \tag{15}$$

При практических вычислениях по формулам (10) и (15) вначале находится разложение Холесского для матрицы $(K + \sigma^2 I)$ (см. Алгоритм 1 и комментарии к нему), которое затем используется как для оценивания среднего, так и для вычисления корня из определителя матрицы (9). Последний для нижней треугольной матрицы есть произведение ее диагональных элементов.

Итак, для *выбранной априорно модели* ковариационной функции и *заданного шума* в данных задача регрессии решена полностью.

Выбор модели гауссовского процесса

Принципиальный вопрос остался не раскрытым в предыдущем разделе — какая из моделей является наилучшей для имеющихся данных? Это вопрос постоянно возникает в различных вариациях при обучении машин. Одна из его типичных форм: «Сколько нейронов скрытого слоя нейронной сети следует выбрать для конкретного набора обучающих данных?»

Хороший ответ на такой вопрос[2] займет не один десяток страниц текста [3]. При этом следует иметь в виду, что теоретические оценки (включающие и информационную значимость входов, и согласованное оценивание

[2]Автор склонен считать, что на вопрос *в такой прямой форме* полностью удовлетворительного ответа нет.

шума и пропусков в данных, и критерии останова, и регуляризацию при обучении, и ...) вполне могут привести к выбору такой нейронной сети, которую на практике *не удается* обучить имеющимися алгоритмами оптимизации.

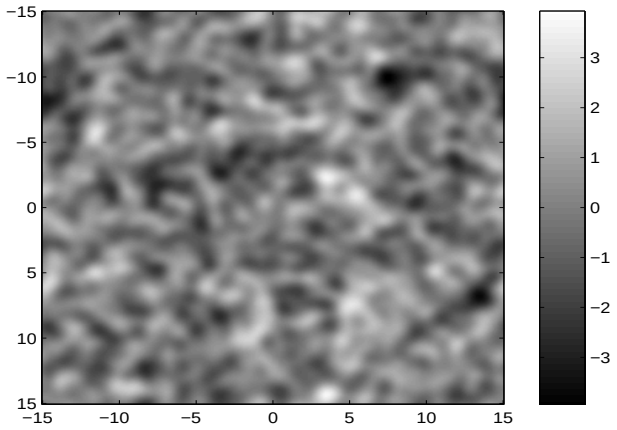
Хороший пример подобного рода — обучение теоретически достаточной нейронной сети с двумя нейронами задаче «исключающее ИЛИ». Градиентные алгоритмы обучения просто не могут достичь обученного состояния! Между тем, сеть с тремя нейронами обучается легко. Нетрудно распространить эту ситуацию на реалистичный случай 2000 обучающих примеров и размерности входов 10, для которого байесовы оценки окажутся оптимальными для, скажем, 30 нейронов. Удастся ли на практике достигнуть для такой сети теоретически оптимального обучения?

Условные гауссовские процессы отличаются тем, что их «обучение» сводится к хорошо известным задачам линейной алгебры с полными (не разреженными) матрицами. Для размерности 5000–10000 строк (равной числу обучающих примеров) прямые алгоритмы с хранимыми в памяти компьютера матрицами гарантированно приводят к устойчивому результату. Разумеется, нужно учесть, что число операций в таких вычислениях пропорционально n^3, и все необходимо проделать для *каждого* тестового примера (в формулах предыдущего раздела участвуют ковариации тестового примера со всеми обучающими примерами). Поэтому обучаемая машина на основе гауссовского процесса является, скорее, «концепт-авто», нежели повседневным инструментом инженера, анализирующего данные.

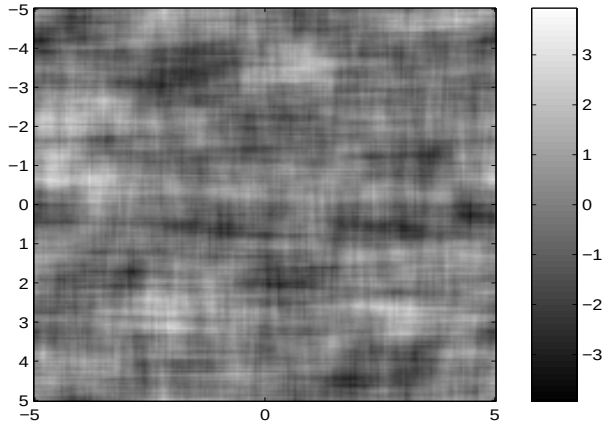
В этой Лекции гауссовские процессы предлагается использовать для предварительных исследований и оценок качества данных. Здесь встает ряд вопросов — какая точность моделирования в данной задаче *принципиально* достижима? Где находится неустранимый уровень неопределенности? Практически полезное решение задачи затем может быть воспроизведено нейронной сетью, комитетом, или каким-то иным эффективным алгоритмом.

Вероятностные модели условных гауссовских процессов и данных включают свободные параметры ковариационной функции и информационного шума. Вид ковариационной функции существенно влияет на характер выборочных траекторий. На рис. 2. приведены примеры траекторий двухпараметрических гауссовских полей для различных ковариационных функций.

Выбор ковариационной функции не произволен. Допустимые функции от пары случайных переменных должны обладать свойством симметрич-



(a)



(b)

Рис. 2. Траектории гауссовских полей для различных ковариационных функций. Изотропный (*вверху*) и анизотропный (*внизу*) случай. Расчеты выполнены при помощи библиотеки программ моделирования стохастических процессов и полей в ООО «Нейрок Техсофт».

ности и неотрицательной определенности. Как правило, предполагают стационарность гауссовского процесса, что приводит к зависимости в ковариационной функции только от разности пар координат (трансляционная инвариантность). В литературе по гауссовским процессам [2,4] обсуждается широкий круг ковариационных функций, моделирующих различные аспекты поведения траекторий. Имеются также специальные ковариационные функции, содержащие вместо симметричной разности скалярное произведение аргументов, аналогично базовой функции искусственного нейрона.

Для ковариационных функций произвольного (допустимого) вида развиты приближенные методы оценивания параметров и траекторий, в частности, основанные на Монте-Карло.

Общая задача построения модели включает как дискретный выбор между типами гауссовских процессов, так и оценивание числовых параметров. Под параметрами модели данных мы будем понимать все имеющиеся в задаче свободные параметры. В данной Лекции мы ограничимся только одним функциональным видом — гауссовскими ковариациями, поскольку для них вычисления интегралов с функциями правдоподобия проводятся аналитически.

Однако, даже при таких ограничениях, модель данных содержит свободные переменные — масштабы изменения функций (например, ширина гауссиана), дисперсия сигнала и дисперсия шума. В общем случае число параметров может расти, но предлагаемые ниже подходы остаются справедливыми и в этих случаях. Оптимальный выбор параметров гауссовского процесса аналогичен процедуре *обучения*.

Параметрическое представление гауссовской корреляционной функции для данных с шумом имеет вид:

$$K(\vec{x}_i, \vec{x}_j) = \sigma_1^2 \exp\Big(-\frac{1}{2}(\vec{x}_i - \vec{x}_j)^T M(\vec{x}_i - \vec{x}_j)\Big) + \sigma_2^2 \delta_{ij}. \qquad (16)$$

Множество параметров включает матрицу и два скаляра: M, σ_1, σ_2. Дальнейшие упрощения связываются с выбором M. Часто используемые варианты — изотропия, эллипсоид и декомпозиция на произведение матриц с неполным рангом:

$$\begin{aligned} M_1 &= l^{-2} I, \\ M_2 &= \big[\mathrm{diag}(\vec{l})\big]^{-2}, \\ M_3 &= \Lambda\Lambda^T + \big[\mathrm{diag}(\vec{l})\big]^{-2}. \end{aligned} \qquad (17)$$

Здесь l — скаляр или вектор с компонентами, имеющими «физическую» размерность длины (для переменных $\vec{x}$, заданных в пространственных координатах). Оптимальные значения этих параметров, таким образом, будут отражать характерные масштабы изменения функций модели. Это в дальнейшем может служить полезным инструментом анализа данных, и, в частности, связываться с относительной информативностью переменных.

Масштаб ковариационной функции характеризует то, насколько требуется сместиться по данной координате, чтобы соответствующие случайные переменные стали некоррелированными. Если масштаб изменений велик, это означает, что данная координата (данный параметр) не несет информации о функции и, эффективно, не дает вклада в модель. Модель M_3 в (17) является аналогом факторного анализа данных: в многомерном пространстве выделяется малое число наиболее информативных направлений — линейных комбинаций координат (факторов в терминах математической статистики).

Байесова индукция

Байесовы методы анализа данных уже неоднократно рассматривались в лекционных циклах [6]. Общая идея байесова подхода к выбору модели состоит в следующем. При заданных параметрах модель объясняет данные с вероятностью, которая называется функцией правдоподобия. Вследствие конечности имеющегося набора данных, установить достоверно идеальные значения параметров не представляется возможным. В байесовом формализме они рассматриваются как случайные величины. Формула Байеса, получаемая из принципа полной вероятности, связывает апостериорные (после наблюдения данных) распределения вероятности параметров с их априорными (предполагаемыми до эксперимента) распределениями:

$$p(\vec{w}\,|\,\vec{y},X)=\frac{p(\vec{y}\,|X,\vec{w})p(\vec{w})}{p(\vec{y}\,|X)=\int d\vec{w}\cdot p(\vec{y}\,|X,\vec{w})p(\vec{w})}. \tag{18}$$

Априорные распределения параметров различны для различных моделей (и, вообще говоря, зависят от дополнительной полезной информации относительно моделируемой системы или процесса). Произвольность в априорных распределениях может быть частично устранена выбором их параметрического представления, определяемого набором *гипер*параметров

θ. Формула Байеса принимает вид:

$$p(\vec{w}\,|\,\vec{y}, X; \theta) = \frac{p(\vec{y}\,|X, \vec{w})p(\vec{w}|\theta)}{p(\vec{y}\,|X; \theta)}. \tag{19}$$

К распределению вероятности данных $p(\vec{y}\,|X; \theta)$ индуктивно применяется тот же байесов принцип — апостериорные распределения θ связываются с априорными. Внешний вид формулы такой же, как и (18), но с заменой $\vec{w}$ на θ. Формально, такая рекурсия может быть бесконечной, однако ее принято замыкать на некотором уровне, заменяя выбор параметров выбором среди фиксированного набора типов моделей. До определенной степени, такое завершение индукции на конечном шаге выглядит искусственным (можно мыслить гипер-модели моделей и т. д.). Здесь нужно иметь в виду, что у байесовой индукции есть естественный уровень останова — когда имеющиеся данные уже не изменяют апострериорное распределение в сравнении с априорным. Другими словами, конечность информации в имеющихся данных не позволяет формально бесконечно улучшать модели — для новых уровней байесова анализа нужна новая информация (дополнительные данные), что выводит задачу за рамки ее исходной постановки.

Часто на практике прибегают к приближениям. Одно из них состоит в игнорировании априорного распределения гипер-параметров и максимизации их функции правдоподобия (которая и есть evidence в модели нижнего уровня). Таким образом, оптимальные значения параметров максимизируют знаменатель[3] в формуле Байеса (19).

Приближения другого рода связываются со статистическими экспериментами, в частности, с оцениванием наилучших параметров путем перекрестной проверки (кросс-валидации) [3]. Методы перекрестной проверки широко обсуждаются в стандартной литературе по математической статистике. Несмотря на существенную критику кросс-валидационных экспериментов, они имеют одно важное достоинство — при оценивании параметров и сравнении моделей может использоваться любая функция потерь. Это весьма актуально, например, в приложениях, связанных с вопросами безопасности, контроля доступа и пр.

Здесь мы остановимся на первом уровне байесовой индукции и покажем, как оцениваются параметры модели гауссовского процесса путем оп-

[3]В зарубежной литературе для нормировочного множителя в формуле Байеса используются термины *evidence* и *marginal likelihood* (интегральная функция правдоподобия). Смысл этого знаменателя близок к понятию *полной статсуммы*, используемому в статистической физике.

тимизации evidence. Заметим, что знаменатель (19) является нормализованной плотностью распределения выходов модели $\vec{y}$, поэтому при максимизации вероятности наблюдаемых данных *автоматически* отдается предпочтение более простым моделям. Таким образом не требуется введение дополнительных регуляризирующих слагаемых (типа длины описания, штрафов за большие значения параметров и пр.)

Выражение (15) для evidence было получено в предыдущем разделе. Оно представляет собой сумму трех слагаемых, описывающих, соответственно, качество аппроксимации данных, сложность модели и (постоянную) нормировку, учитывающую число обучающих примеров. Обозначив $\widetilde{K} = K + \sigma^2 I$, получим общее выражение для градиента evidence относительно параметров:

$$\frac{\partial}{\partial \theta_j} \ln\bigl[p(\vec{y}\,|X;\theta)\bigr] = \frac{1}{2}\Bigl(\vec{y}^T \widetilde{K}^{-1} \frac{\partial \widetilde{K}}{\partial \theta_j} \widetilde{K}^{-1} \vec{y} - \mathrm{tr}\Bigl(\widetilde{K}^{-1} \frac{\partial \widetilde{K}}{\partial \theta_j}\Bigr)\Bigr), \qquad (20)$$

где $\mathrm{tr}(\cdot)$ — операция взятия следа матрицы.

Стоимость вычисления градиента пропорциональна n^3, при этом обращение матрицы выполняется один раз для всех компонентов. Оптимизация может проводиться с применением традиционных методов поиска минимума гладкой функции без ограничений на переменные. К таким алгоритмам относятся, например, сопряженные градиенты или RProp, широко используемые при обучении нейронных сетей.

В простейшем случае изотропной модели M_1 в (17) оптимизация проводится в пространстве трех скалярных параметров. Нужно отметить, что функция (15) может не являться выпуклой, что приводит к проблеме локальных минимумов. На практике большинства вычислительных трудностей можно избежать путем выбора начального приближения, учитывающего специфику данных (так, характерный уровень шума может быть известен из анализа погрешностей измерений, характерный масштаб оценивается по статистике расстояний до ближайших соседей или на основе выборочной константы Липшица).

Моделирование

В завершающем разделе лекции рассматриваются примеры реалистичного моделирования данных траекториями гауссовских процессов.

Что необходимо для проведения вычислений?

Оценивание параметров условного гауссовского процесса проводится путем анализа «ядерных» матриц (4) для парных комбинаций из обучающих и тестовых (новых) данных. Эффективность таких вычислений базируется на удачном выборе контейнера для хранения полных симметричных матриц большого размера и доступа к матричным элементам. Тип доступа определяется требованиями вычислительных алгоритмов.

Основной вычислительный алгоритм здесь — **разложение Холесского** для симметричной положительно определенной матрицы. При выборе алгоритма нужно учесть, что накопление ошибок округления может нарушать свойство положительной определенности. На практике для обеспечения устойчивости может использоваться регуляризирующее предобуславливание матрицы (например, искусственное добавление малых положительных поправок к диагональным элементам).

Второе важное обстоятельство — наличие малых матричных элементов для взаимно удаленных пар точек и быстро убывающей ядерной функции, описывающей корреляции гауссовского процесса. Эта проблема обостряется при поиске оптимальных значения параметров, когда пробные масштабы затухания корреляций не соответствуют обучаемым данным. Отчасти, исчезающие элементы матрицы можно использовать для эффективной разреженной организации данных, однако при этом нужно гарантировать сохранение «связности» данных (включая новые тестовые примеры). В противном случае, функция правдоподобия и условная плотность распределения могут вырождаться до отдельных пиков исчезающей ширины.

Библиотека обращения с полными матрицами и алгоритм разложения Холесского дают решение задачи оценивания регрессии при заданных параметрах модели гауссовского процесса. Для автоматического поиска наиболее вероятных значений этих параметров необходимо использование алгоритма безусловной **оптимизации функции** многих переменных. Целевой функцией выступает интегральная функция правдоподобия данных (evidence).

Задача оптимизации не является в строгом смысле выпуклой, поэтому алгоритм должен учитывать наличие локальных минимумов. Для задач с невысокой размерностью пространства входов (10–30) могут использоваться алгоритмы гладкой локальной оптимизации с рестартами. Схема принятия решения об останове поиска может учитывать фактически уровень достигнутого правдоподобия, а также некоторые эвристические оценки.

Например, если уровень шума в данных известен априори или может быть оценен независимо, то практическим критерием останова может служить близость оцениваемого моделью и априорного уровня шума.

В целом, задача с 2000–3000 обучающими примерами при размерности данных $\sim$10 до конца решается в рамках вычислительных возможностей системы вычислений Matlab. С ростом размерности и объема данных необходимо привлечение специализированных разработок.

Искусственные данные

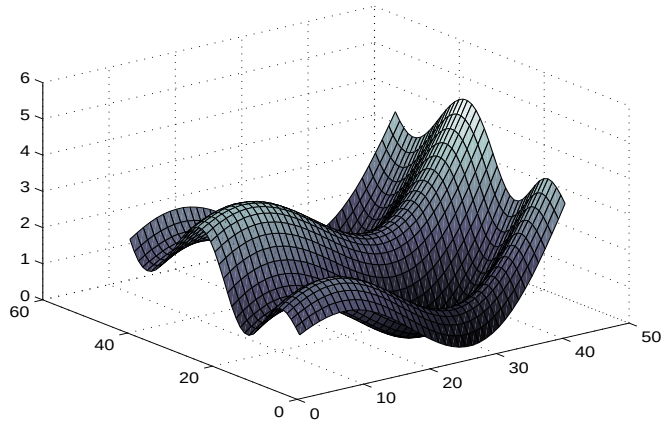
Иллюстрации свойств вычислительных моделей традиционно начинаются с рассмотрения модельных примеров с искусственными данными. Для задачи нелинейной регрессии хорошей начальной точкой могут служить модельные функции, предложенные *Hwang* [10]. Фактические данные приведены в коллекции DELVE [11].

Набор данных содержит значения 5 сложных для аппроксимации функций двух переменных. Для каждой функции имеется две выборки данных — идеальная (без шума) и зашумленная нормальными случайными числами с дисперсией (std) 0.25. Для моделирования была выбрана функция № 4, результат аппроксимации которой по 2000 случайным примерам приведен на рис. 3а.

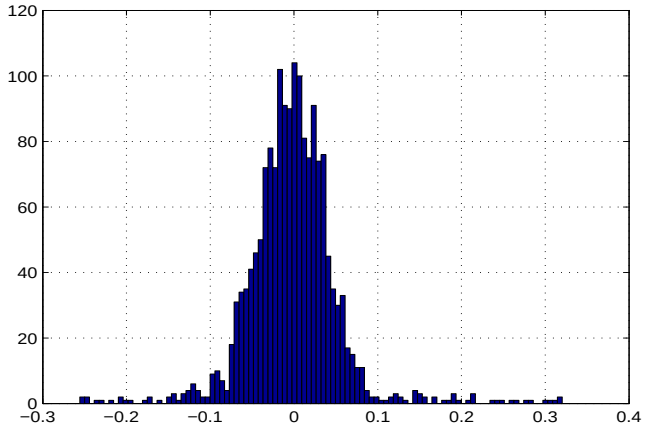
Среднеквадратичная ошибка аппроксимации при оптимальных значениях параметров модели составляет 0.018, что меньше уровня шума в данных. Оцениваемый моделью шум данных (один из свободных параметров модели) 0.252, что хорошо согласуется с фактическим уровнем шума. Модель также предсказывает характерный масштаб 0.46 для изменения одной из переменных и 0.21 для другой, что оценивает их сравнительную информативность. Эта асимметрия хорошо наблюдается на рис. 3а.

Разумеется, такое наблюдение возможно лишь в пространстве двух измерений, выводы же модели отражают информативность и для случая большего числа переменных.

При вычислениях также оценивается интегральный уровень правдоподобия полученной регрессионной зависимости, который может использоваться при сравнении различных моделей.



(a)



(b)

Рис. 3. Результаты аппроксимации модельной функции: **(a)** аппроксимация по зашумленным данным; **(b)** распределение ошибок аппроксимации

Регрессия вложения финансового временного ряда

Задача прогнозирования временных рядов является одной из наиболее часто встречающихся на практике. Отвлекаясь от целей прогнозирования (а также от вопросов адекватности регрессионных моделей, основывающихся на гауссовском распределении уклонений [12]), рассмотрим типичный временной ряд индекса РТС. Ежедневные данные к концу торговой сессии доступны на Интернет-узле РосБизнесКонсалтинг (`http://www.rbc.ru/`). Приводимое ниже «наивное» исследование направлено не на решение бизнес-задачи принятия решения на основе прогноза ряда, а на обсуждение характерных проблем, встречающихся при обработке реальных данных.

Для нормализации данных и устранения систематических трендов можно перейти от значения индекса к относительной величине — доходности. Однодневные относительные доходности ряда $T(t)$ вычисляются по формуле:

$$r_d(t) = \frac{T(t+d) - T(t)}{T(t)}, \quad d = 1. \tag{21}$$

Величина доходности имеет распределение с «тяжелыми» крыльями (более 1,5% данных выходит за 3 стандартных отклонения), как и отмечалось в [12]. Важен, однако, не сам закон распределения данных, а характер их уклонения[4] от функциональной модели. Зададимся вопросом о степени прогнозируемости однодневной доходности индекса на сегодня, как функции набора доходности в предыдущие дни:

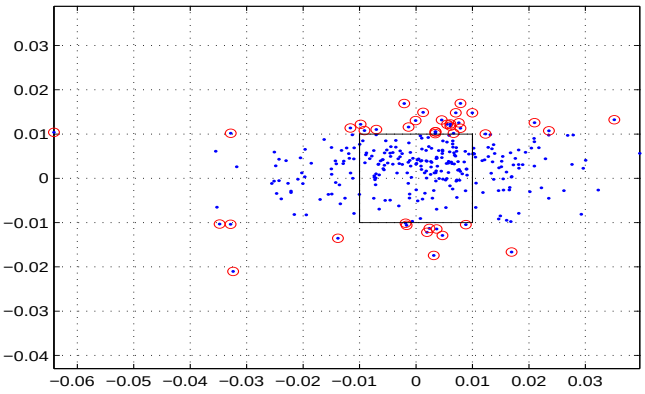
$$r_1(t) = f\big(r_1(t-1), r_2(t-2), \ldots, r_k(t-k)\big). \tag{22}$$

Полученное вложение ряда в k-мерное числовое пространство позволяет свести задачу прогнозирования к задаче восстановления регрессии по многомерным данным, которой и посвящена данная Лекция.

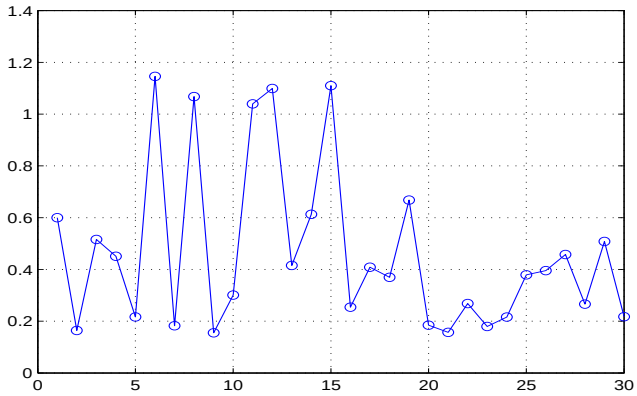
Доходности за разные временные промежутки имеют разный масштаб. Если предположить броуновский характер дрейфа, то амплитуда доходности на k дней, в среднем, в $\sqrt{k}$ выше однодневного показателя. Поэтому столбцы таблицы данных вложенного ряда обычно подвергаются нормализации.

Регрессионная модель зависимости (22) строится аналогично рассмотренному выше примеру с искусственными данными. Число свободных па-

[4]На практике целесообразна проверка нормальности окончательного распределения ошибок.



(a)



(b)

Рис. 4. Результаты экспериментов с регрессионной моделью: **(a)** Зависимость прогнозируемой доходности от фактической. Квадрат в центре отделяет данные с низкими значениями (1% и менее). Кружки соответствуют 40 из 300 дней, когда прогнозируемая доходность по модулю превышала 1%. **(b)** Оценки относительной значимости предыдущих доходностей к сегодняшнему дню для оценивания однодневной доходности на завтра.

раметров модели (17) равно $k + 2$. Оптимальные значения параметров выбираются путем минимизации интегрального правдоподобия.

Для обучения использовался доступный отрезок ряда с момента начала его истории (01.09.1995). Последние 300 значений оставлены для тестирования.

Результаты тестирования (зависимость прогнозируемого значения доходности на завтра от ее фактического значения) приведены на рис. 4a.

Правильный знак изменения доходности прогнозируется в 24 из 40 случаев. Коэффициент линейной корреляции равен 0.12.

Некоторый интерес представляют результаты анализа относительной значимости факторов (предыдущих доходностей). На рис. 4b приведены значения обратных характерных масштабов изменений регрессионной зависимости по каждой из переменных. Напомним, что эти масштабы есть, суть, оптимальные значения параметров ковариационной функции гауссовского процесса.

Наиболее значимы 6, 8, 11, 12 и 15-дневные доходности. Напротив, однодневная вчерашняя доходность не является лидером по значимости.

Эта информация может использоваться при разработке прогностических моделей, учитывающих множество экономических факторов. В таких моделях сам ряд достаточно представить пятерками чисел, отвечающих наиболее информативным масштабам. Исследования экономических особенностей задачи прогнозирования выходит за рамки Лекции, заинтересованный читатель может обратиться к литературе [12].

О прикладном значении условных гауссовских процессов

Рассмотренные в лекции алгоритмы моделирования условных вероятностей в задачах регрессии относятся к разряду вычислительно трудоемких. Базовые этапы вычислений требуют затрат, пропорциональных кубу числа обучающих примеров. Традиционно, вычисления такой сложности редко рассматриваются, как имеющие серьезную перспективу для прямого использования в приложениях.

Однако имеется широкий класс прикладных проблем, характеризующихся, с одной стороны, высокой стоимостью экспериментов, а с другой стороны сложным нелинейным поведением измеряемой функции (отклика системы). К таким задачам относятся:

- планирование вычислительных и натурных экспериментов;
- оптимизация параметров сложных систем на основе экспериментальных данных.

Обе задачи оперируют относительно небольшим числом имеющихся к данному моменту измерений, и основная проблема состоит в указании оптимальной точки или плана (набора точек) для последующих измерений или вычислений. При этом крайне важной является оценка не только ожидаемого значения функции, даваемого регрессионной моделью, но и оценка неопределенности будущего измерения. Традиционные методы планирования экспериментов оперируют с простейшими регрессионными моделями, которые применимы только в локальной области пространства параметров. Гауссовские процессы свободны от этого недостатка, и предоставляют новые возможности.

Так, в задаче планирования технологических мероприятий на нефтяных и газовых месторождениях критическим является отбор скважин-кандидатов, для которых минимальна неопределенность эффекта при максимальной, либо положительно достаточной, ожидаемой величины эффекта.

При решении проблемы адаптации математических моделей месторождений к наблюдаемым выходам продукта требуется, наоборот, указывать такие наборы пробных параметров среды (нефтеносного пласта), при которых максимально понижается совокупная неопределенность регрессионной модели.

Алгоритмы гауссовских процессов дают надежные оценки, как самой поверхности регрессии, так и ее неопределенности. При этом модели не содержат произвольных параметров, их оптимальные значения находятся автоматически, в процессе построения модели. Это позволяет рассматривать такие многоцелевые постановки задач оптимизации, как нахождение оптимума при ограничениях на уровень риска, и , наоборот, минимизация риска при ограничениях на уровень полезного эффекта.

Гауссовские процессы также занимают важное место в арсенале методов [7–9] предобработки и предварительного анализа данных при разработке нейросетевых информационных моделей. Гауссовский процесс дает близкую к наилучшей возможной оценку достижимого уровня точности модели, а также оценку уровня шума и значимости факторов. На последующих этапах строятся нейросетевые модели данных со сравнимой точностью, причем поиск нейронных сетей происходит как бы в условиях «с известным ответом», что значительно повышает технологичность и доказательность моделирования. В приложениях используются уже нейросетевые модели, поскольку они обеспечивают необходимую эффективность, быстродействие и компактность решений.

В заключение отметим, что одним из путей преодоления ограничений на число обучающих примеров в гауссовский процессах является использование комитетов и ансамблей моделей, рассмотренных в одной из предыдущих лекций автора [14].

Благодарности

Автор благодарит *А. Воронцова*, инициировавшего тематику случайных процессов и полей в компании «Нейрок Техсофт», за многочисленные комментарии и обсуждения. *А. Воронцовым* и *А. Черкасовым* разработана промышленная библиотека C-программ моделирования многомерных гауссовских полей, примеры расчетов по которой использованы в Лекции.

Литература

1. *Тутубалин В. Н.* Теория вероятностей и случайных процессов. – М.: МГУ, 1992.
2. *Булинский А. В., Ширяев А. Н.* Теория случайных процессов. – М.: Физматлит, 2005.
3. *Bishop C. M.* Neural networks for pattern recognition. – Oxford University Press, 1995.
4. *Rasmussen C. E., Williams C. K. I.* Gaussian processes for machine learning. – MIT Press, 2006.
   URL: http://www.gaussianprocess.org/gpml/
5. *Голуб Дж., Ван Лоун Ч.* Матричные вычисления. – М.: Мир, 1999.
6. *Шумский С. А.* Байесова регуляризация обучения // Лекция для IV Школы-семинара *«Современные проблемы нейроинформатики»*. – М.: МИФИ, январь 2002. – с. 30–93.
7. *Vapnik V. N.* The nature of statistical learning theory. – Springer-Verlag, 1995.
8. *Hastie T., Tibshirani R., Friedman J.* The Elements of statistical learning. – Springer, 2001.
9. *Терехов С. А.* Технологические аспекты обучения нейросетевых машин // Лекция для VIII Школы-семинара *«Современные проблемы нейроинформатики»*. – М.: МИФИ, январь 2006. – с. 13–73.
10. *Hwang J.-N., Lay S.-R., Maechler M., Martin R. D., Schimert J.* Regression modeling in back-propagation and projection pursuit learning // *IEEE Transactions on Neural Networks*. – 1994. – vol. 5, No. 3. – pp. 342–353.

11. DELVE project www page:
URL: http://www.cs.toronto.edu/∼delve/
12. *Мандельброт Б., Хадсон Р. Л.* (Не)послушные рынки. – М.: Вильямс, 2006.
13. Экономико-математическое моделирование / Под ред. *И. Н. Дрогобыцкого*. – М., 2004.
14. *Терехов С. А.* Гениальные комитеты умных машин // Лекция для IX Школы-семинара *«Современные проблемы нейроинформатики»*. – М.: МИФИ, январь 2007. – с. 11–42.

Задачи

Задача 1. Матричные вычисления

Требуется вывести формулы (6) для условных распределений траекторий гауссовского процесса. Подсказка: воспользоваться леммой об обращении матриц (см. [5], а также URL:
`http://en.wikipedia.org/wiki/Matrix_inversion_lemma`).

Задача 2. Уклонение траекторий от нуля

Траектории (реализации) гауссовских процессов таковы, что плотность распределения случайной переменной в каждой точке является гауссовской. Рассмотрим поведение траекторий одномерного гауссовского процесса с гауссовской ковариационной функцией на отрезке $[0, h]$. Максимальное достигаемое на этом отрезке значение (верхняя грань) функции, являющейся траекторией процесса, есть, суть, случайная величина. Каков закон распределения этой случайной величины (он будет включать в качестве параметров ширину ковариационной функции и длину отрезка)? Принципиальным является ответ на вопрос: имеет ли данное распределение «тяжелые крылья» при больших значениях уклонения траектории от нуля?

Сергей Александрович ТЕРЕХОВ, кандидат физико-математических наук, заместитель Генерального директора ООО «Нейрок Техсофт» (г. Троицк, Московская обл.). Область научных интересов — анализ данных при помощи искусственных нейронных сетей, генетические алгоритмы, марковские модели, байесовы сети, методы оптимизации, моделирование сложных систем. Автор 1 монографии и более 50 научных публикаций.