СЕРГЕЙ А. ТЕРЕХОВ
Снежинский Физико-Технический Институт, г. Снежинск.
ООО НейрОК, г. Москва.
E-mail: alife@narod.ru

НЕЙРОСЕТЕВЫЕ АППРОКСИМАЦИИ ПЛОТНОСТИ В ЗАДАЧАХ ИНФОРМАЦИОННОГО МОДЕЛИРОВАНИЯ

Аннотация

С прикладных позиций рассматривается задача аппроксимации плотности распределения, описывающего множество многомерных экспериментальных данных. Предложены эффективные нейросетевые методики аппроксимации формы плотности. Приведены примеры постановок задач анализа данных на основе аппроксимации плотности. Обсуждаются приложения подхода.

SERGE A. TEREKHOFF
Snezhinsk Institute of Physics and Technology (SFTI), Snezhinsk
NeurOK LLC, Moscow.
E-mail: alife@narod.ru

NEURAL APPROXIMATIONS OF PROBABILITY DENSITY IN INFORMATIONAL MODELING

Abstract

The problem of probability density approximation, based on set of multivariate experimental data is considered from the point of view of practical informatics. Effective neural techniques of approximation of the density form are proposed. Statements of several data analysis problems are presented using density approximation approach. Applications of the method are discussed.

Содержание

Содержание

1 Плотность распределения и ее роль в информационном моделировании **1**

2 Подходы к аппроксимации плотности распределения **6**
2.1 Пример 1. Аппроксимация плотности на отрезке 8

3 Бутстреп-выборки **11**

4 Численные эксперименты **13**
4.1 Задача Banana 13
4.2 Задача прогноза загрузки процессора ЭВМ (CompAct) . . 14

5 Обсуждение **16**

6 Благодарности **17**

7 Литература **17**

8 Приложение А. **19**

1 Плотность распределения и ее роль в информационном моделировании

Рассмотрим проблему построения эмпирических моделей на основе числовых данных. Речь пойдет об обучении без учителя на примерах в условиях неопределенности о характере модели. Пусть данные составляют совокупность имеющихся результатов экспериментов или наблюдений над некоторой сложной[1] системой или устройством. В центре рассмотрения лежит матрица наблюдений D_{jk}, в которой $j = 1..N$ - номер наблюдения (одна строка в таблице или запись в базе данных), а $k = 1..M$ - номер наблюдаемой переменной x_k (признака, фактора, свойства и т.д.). Матричным элементом является действительное число - результат наблюдения. Здесь мы ограничимся случаем непрерывных переменных и оставим на время дискретные (ординальные и категориальные) наблюдения.

На этом этапе относительно механизма порождения данных будем предполагать следующее:

- Наблюдаемое значение является реализацией некоторой случайной величины;
- Наблюдаемые данные порождены стационарным процессом (системой), т.е. рассматриваемые случайные величины не зависят от времени;
- Различные наблюдения не зависят друг от друга;

[1] Подчеркнем разницу между экспериментом и наблюдением. При проведении эксперимента условия внешнего воздействия и параметры исследуемой системы управляются экспериментатором. При простом наблюдении система изучается при тех параметрах и условиях, в которых она (случайно) оказалась в момент наблюдения. Различие весьма существенно при изучении, например, социальных, рыночных систем, или живых организмов. В данных случаях, эксперимент, как правило, не возможен.

- Факт наблюдения не влияет на свойства исследуемой системы (процесса).

Таким образом, наблюдаемые данные, суть, порождены некоторой вероятностной [1] средой. При этом можно выделить несколько основных причин, по которым предлагается согласиться с вероятностной[2] трактовкой данных:

- Процесс измерения сопряжен с экспериментальными погрешностями;
- Изучаемая система является сложной [2], т.е. несводимой к сумме свойств отдельных компонент, и наблюдаемое многообразие данных может быть равновероятно объяснено великим множеством структурных описаний, при этом нельзя достоверно предпочесть ни одно из них;
- Объем измерений конечен и не может считаться исчерпывающим описанием системы.

В этих условиях мы будем говорить о статистической модели, описывающей совокупность данных, как об *информационной* модели изучаемой сложной системы.

Наиболее полным статистическим описанием наблюдаемых данных является совместная плотность распределения вероятности точек в векторном пространстве признаков $P(x_1, x_2..x_M)$. Рассмотрим особенности задачи ее аппроксимации. В отличие от традиционных постановок задач сглаживания данных [3], когда в распоряжении у исследователя имеются пары значений "аргумент - функция", при аппроксимации плотности даны только координаты точек в многомерном пространстве. Поэтому будем считать, что аппроксиматором плотности P множества точек D в параллепипеде V является всякая функция A, такая, что:

- A равна нулю вне V;
- нормирована на $V : \int A(X)dV = 1$;
- Отношение интегралов от A по двум объемам V_1 и V_2 из V, содержащим точки из D, "стремится"к отношению числа точек из D в этих объемах.

[2]Содержательными являются также и другие (не вероятностные) трактовки. В частности, система может рассматриваться, как описываемая некоторым числом детерминированных скрытых факторов, а стохастичность наблюдаемых переменных (которых обычно больше, чем факторов) является внешней по отношению к самой системе. Другой подход - нелинейные динамические системы, порождающие наблюдаемый динамический хаос.

Намеренно остановившись на столь нестрогом определении, подчеркнем важные с практической точки зрения свойства задачи:

- Всякая попытка восстановления плотности вдали за границами объема V, содержащего точки наблюдений, потребует дополнительных предположений и ограничений.
- Внутри исследуемого объема задача восстановления плотности также является некорректно поставленной [4], хотя бы уже потому, что решение не единственно.
- У задачи, в некотором смысле, нет "наилучшего"решения, имея в виду использование оцененной плотности для генерации и объяснения новых данных.

Заметим, что некоторые простые, кажущиеся естественными, попытки построить функционалы, оптимизация которых ведет к устранению некорректности задачи, часто приводят к решениям с весьма скромной практической пользой. Так например, если в качестве такого функционала выбрать популярный принцип максимального правдоподобия, т.е.

$$\max \sum_{j \in D} \log(A(D_j))$$

и не предпринять никаких дополнительных мер по регуляризации, то легко получим, что функция

$$A(x) = \frac{1}{\|D\|} \sum_{j \in D} \delta(x - D_j)$$

является неограниченно "правдоподобной", удовлетворяя наилучшим образом всем условиям, наложенным ранее на аппроксиматор. К сожалению, универсальных автоматических рецептов для регуляризации не существует.

В следующих разделах мы вернемся к практическим методам регуляризации задачи аппроксимации, а сейчас зададимся вопросом, что дает исследователю знание совместной плотности распределения признаков изучаемого объекта?

В некотором смысле, описание на языке плотности распределения соответствует описанию на языке волновой функции в квантовой механике. Полнота статистического описания на основе совместной плотности вероятности подразумевает, что в плотности P содержится практически вся[3] информация об исследуемой системе, которую можно почерпнуть

[3] В действительности, всегда происходит некоторая потеря информации вследствие ошибок аппроксимации. Заметим, также, что ограничение информации при аппроксимации плотности объемом информации, содержащимся в исходных данных - является, с точки зрения автора, ответом на дискуссию с А.А. Ежовым [6].

из имеющихся данных. Это, в свою очередь, означает, что все наблюдаемые следствия о значениях и распределениях различных величин при различных условиях могут быть получены путем вычислений функционалов от P. Искомые функционалы в общем случае могут быть оценены методом Монте-Карло с различными вариациями [5].

Рассмотрим семейство важных функциональных запросов, представляющих прикладной интерес.

1. **Однофакторные условные распределения.** Как распределен какой-то их признаков, если значения остальных признаков известны достоверно? В случае, если выделенный признак является выходным (зависимым от остальных) и кодирующим некоторый отклик, эта задача соответствует задаче распознавания образов. В качестве искомого кода образа, задаваемого остальными признаками, принимается наиболее вероятное или среднее значение в распределении признака-кода. Такая плотность вероятности называется условной:

$$P(x_k|x_1^*,\ldots,x_{k-1}^*,x_{k+1}^*,\ldots,x_M^*) =$$
$$= \frac{P(x_1^*,\ldots,x_{k-1}^*,x_k,x_{k+1}^*,\ldots,x_M^*)}{\int dx_k P(x_1^*,\ldots,x_{k-1}^*,x_k,x_{k+1}^*,\ldots,x_M^*)}$$

   Таким образом,имеет место пропорциональность:

$$p(x_k) \sim P(x_1^* \ldots x_k \ldots x_M^*)$$

   В общем случае, в отличие от традиционной задачи аппроксимации поверхности отклика или распознавания образов, при использовании для оценки условной плотности (как функции одной переменной x_k) некоторой аппроксимации совместной плотности P получается *истинное распределение возможных значений результата,* а не только его наиболее вероятное значение.

   Аналогично рассматривается и задача распределения пары признаков или любого другого их числа:

$$P(x_1,\ldots,x_q|x_{q+1}^*,\ldots,x_M^*) \sim P(x_1,\ldots,x_q,x_{q+1}^*,\ldots,x_M^*)$$

2. **Пропуски в данных.** Какие значения может принимать некоторый признак, если значения части других признаков известны достоверно, а оставшихся - неизвестны вовсе? К этой задаче сводится известная проблема заполнения пропусков в таблицах данных, в контексте нейронных сетей подробно рассмотренная в работе [7]. Соответствующие распределения даются маргинальными интегралами от совместной плотности:

$$P(x_k|x_q^* \ldots x_M^*) \sim \int dx_1..dx_h P(x_1..x_h,\ldots,x_k,\ldots,x_q^* \ldots x_M^*)$$

3. **Вероятностный прогноз.** Какова форма плотности распределения условных вероятностей выделенных признаков, если известны однофакторные плотности распределения остальных переменных. Результат:

$$P(x_1 \dots x_q | p(x_{q+1}) \dots p(x_M)) \sim$$

$$\sim \int dx_{q+1} p(x_{q+1}) \dots dx_M p(x_M) P(x_1 .. x_M)$$

4. **Обратные задачи.** Каковы должны быть величины переменных $x_1 ... x_{k-1}, x_{k+1}, .. x_M$, чтобы наиболее вероятным значением для переменной x_k было число $x_k^\star$? Эта задача сводится к (численному) поиску решений нелинейного уравнения на максимум величины условной вероятности

$$\max_{x_k} P(x_k | x_1 ... x_{k-1}, x_{k+1}, .. x_M) = x_k^\star$$

как функции $M-1$ переменной. Эффективность решения такой задачи во многом определяется функциональной простотой выбранной аппроксимации плотности.

Большой интерес представляют также различные комбинации перечисленных задач (например, случай, когда часть переменных известны точно, для других известны лишь оценки распределений, а прочие - неизвестны).

Важно отметить, что все выражения, в которых используется совместная плотность, требуют ее определения с точностью до некоторого постоянного множителя - везде встречаются только отношения плотностей. Эта деталь существенно облегчает нашу задачу!

2 Подходы к аппроксимации плотности распределения

Переходим теперь собственно к вопросу построения аппроксимаций. Основной объем литературы по вопросам аппроксимации плотности вероятности можно условно разбить на такие крупные разделы, как:

- *Параметрические методы аппроксимации.* Подходы этого раздела основаны на предположении о конкретном функциональном виде распределения, который зависит от некоторых параметров. Эти параметры далее выбираются на основе статистических критериев максимального правдоподобия или максимума апостериорной вероятности описания данных моделью. Теория и методики этого раздела подробно обсуждаются в текстах по математической статистике и обработке результатов экспериментов.

- *Методы непараметрической статистики.* Сюда относятся выборочные частотные гистограммы (мало применимые в многомерном случае, см. ниже) и широкий класс методов, основанных на аппроксимации плотности смесью базисных функций [8-10]. Частным случаем такой аппроксимации являются Гауссовы смеси [11], радиальные базисные функции, а также вейвлет-методы [12].

Плотность распределения является функцией многих переменных - по числу исследуемых признаков. В условиях, когда имеется дополнительная информация о степени зависимости или независимости признаков, эта плотность может быть факторизована на функции меньшего числа переменных. При дополнительных сведениях об обуславливании одних признаков другими удобнее говорить об условных[4] плотностях распределений, которые также зависят от меньшего числа переменных (т.е., только от обуславливающих признаков). Такое факторизованное описание дается байесовыми сетями событий [5].

Байесовы сети являются, по-видимому, самой революционной технологией последнего десятилетия[5] в области искусственного интеллекта. Основное их достоинство - универсальное и интуитивно понятное представление моделей, основанных на данных. В противоположность нейронным сетям, графические байесовы модели допускают прямую интерпретацию. Байесова сеть состоит из узлов, соответствующих переменным задачи, и ребер, отражающих зависимости между переменными. Отсутствие ребра между двумя узлами означает независимость между их переменными.

Баейсовы сети - столь обширная и крайне интересная научная область, что, не в силах более подробно обсуждать здесь этот вопрос, автор предлагает ознакомиться с ним по работе [14].

Важно, что в результате построения байесовой сети происходит редукция плотности. Так, например [15], для байесовой сети, отражающей зависимости между 9 переменными (Рис. 1), плотность факторизуется следующим образом:

$$P(A,B,C,D,E,F,G,H,I) = P(A)P(B)P(C|A,B)P(D|C)\times$$

$$\times P(E|D)P(F|C)P(G|E,F)P(H|G)P(I|G)$$

Каждая из полученных функций зависит от 1 или 2 переменных. Далее, если ограничиться вместо описания распределений вероятности

[4]Существует точка зрения (см. например, [13]), что вероятность, как мера наших ожиданий (belief), всегда является условной, т.е. обусловленной текущим объемом информации, имеющимся у исследователя.

[5]Бил Гейтс заявил в интервью газете Лос-Анжелос Таймс (28 октября 1996), что байесовы сети более перспективны и значимы, чем DOS (http://www.cs.berkeley.edu/ murphyk/Bayes/la.times.html)

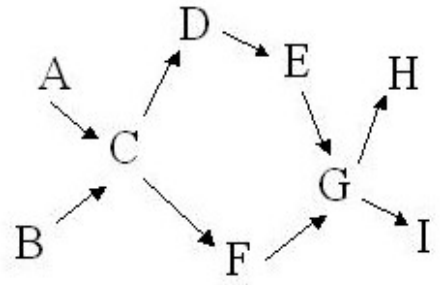

Рис. 1: Байесова сеть для 9 переменных

только их математическими ожиданиями, тогда формально связь между обуславливающими (входными) и обуславливаемыми (выходными) признаками может быть эффективно представлена многослойной нейронной сетью [16], известной своими хорошими аппроксимационными свойствами (см. например, [17]). И, наконец, если в распоряжении исследователя имеются формальные логические соотношения между входными и выходными переменными, то их можно отразить в форме экспертных систем или функциональных уравнений.

В данной работе предлагается подход на основе замены задачи аппроксимации эквивалентной задачей классификации, при этом рассмотрение ведется на исходном уровне совместной многомерной плотности распределения вероятности. Суть метода состоит в построении наилучшего решающего правила, позволяющего отличить наблюдаемую совокупность данных от некоторой искусственной выборки данных с известной плотностью распределения. Этот метод является, по-видимому, относительно новым [18] и мало распространен[6]. Рассмотрим его в самой простой форме - на примере.

2.1 Пример 1. Аппроксимация плотности на отрезке

Пусть в нашем распоряжении имеется выборка из суммы двух различных Гауссовых распределений, априорные вероятности которых равны:

$$P_{exact} = \frac{0.5}{\sqrt{2\pi \cdot 1^2}} \exp\Big(- \frac{|x-1|^2}{2 \cdot 1}\Big) + \frac{0.5}{\sqrt{2\pi \cdot 0.5^2}} \exp\Big(- \frac{|x-3|^2}{2 \cdot 0.5}\Big)$$

График точного значения плотности представлен сплошной линией на Рис.3. Объем выборки - $2N$ точек. Для аппроксимации такой плотности распределения на отрезке $[-3..5]$ дополнительно (искусственно)

[6] Автору посчастливилось предложить этот метод независимо от других публикаций

разыграем еще $2N$ равномерно распределенных случайных точек. Их плотность распределения равна константе:

$$P_0 = \frac{1}{x_{right} - x_{left}} = \frac{1}{8}$$

Объединим оба множества в одно - их $4N$ наблюдений. Первым $2N$ точкам сопоставим значение "класс A", равное 1. Равномерно распределенным $2N$ точкам припишем "класс B"со значением 0. Класс B выполняет эффективную функцию шума. Наша задача состоит в построении классификатора, отличающего точки сигнала (класс A) от этого шума. Выход классификатора будет принимать значения в интервале [0..1]. В качестве классификатора предлагается использовать многослойную нейронную сеть с сигмоидальными нейронами [16], обучаемую методом *RProp* (Приложение A).

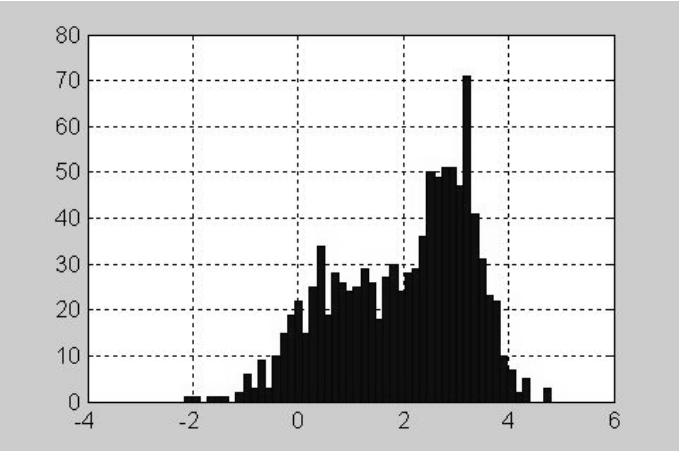

Рис. 2: Гистограмма обучающих данных класса A (сигнал)

Итак, на вход нейросетевого классификатора подается координата одной их точек совокупной выборки, и при обучении требуем, чтобы выход нейросети был равен 0 либо 1 в зависимости от класса, к которому принадлежит данная точка. Класс каждой точки достоверно известен (по построению).

Без сомнений, у такого классификатора нет шансов безошибочно обучиться, так как предъявляемые ему точки обоих классов эффективно перемешаны на отрезке. Для каждой точки в отдельности нет способа достоверно сказать, порождена она сигналом или шумом. К чему же будет сходиться обученный классификатор?

Для ответа на этот вопрос выделим окрестность dx некоторой точки x на исследуемом отрезке. В этой окрестности число точек класса близко к $P(x)dx$, а число голосующих за класс B стремится к $P_0(x)dx$. Ясно,

что минимизирующее квадрат уклонения (т.е. оптимальное) значение на выходе классификатора стремится к величине

$$y(x) = \frac{P(x)}{P(x) + P_0(x)}$$

равной отношению числу голосов, поданных за "1"к общему числу голосов. Отсюда для формы плотности легко находим

$$P(x) = P_0 \frac{y(x)}{1 - y(x)}$$

Простой анализ чувствительности дает

$$\delta P \sim P \frac{1}{y(1-y)} \delta y$$

т.е. ошибка аппроксимации сосредоточена в основном в областях насыщения сигмоиды выходного нейрона классификатора.

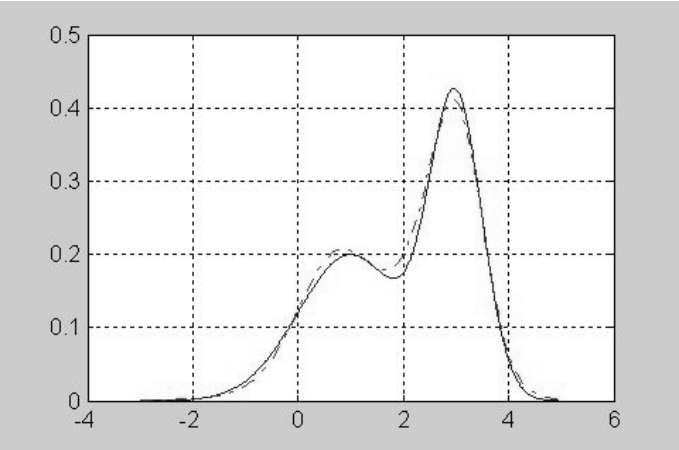

Рис. 3: Результат нейросетевой аппроксимации плотности распределения (пунктир) в сравнении с точной формой плотности (сплошная линия)

Результаты аппроксимации этим методом приведены на Рис.3, где также приведен график точной формулы двух Гауссианов. Достаточно легко видеть, что такой аппроксимации весьма трудно достичь, сглаживая гистограмму.

В приведенном примере, в действительности, нигде не использовался факт одномерности распределения. В точности такой же метод используется для аппроксимации многомерных данных. Остановимся теперь на особенностях и видимых проблемах предложенного метода аппроксимации плотности.

- *"Проклятие"размерности.* Простые оценки показывают, что при росте числа переменных надежды на надежное статистическое описание данных тают на глазах. Так, для построения достоверной гистограммы в пространстве 10 измерений даже с двумя интервалами по каждому из них требуется масштаба $30 \times 2^{10} = 30,000$ точек. Такие объемы данных имеются только в приложениях, связанных или относящихся к реальному времени (поступление финансовых тиков на биржах, или онлайн-диагностика). У экспертов, изучающих некоторое устойчивое явление или систему, обычно имеются базы данных лишь с 100...5000 записями. На такой взрывной характер зависимости объема требуемых данных от размерности обратили внимание еще пионеры раскопки данных *(data mining)*, такие как *John Tukey* [19]. В нашем подходе эта фундаментальная проблема, разумеется, не устраняется, но частично ослабляется тем фактом, что очень сложные, существенно многомерные, формы поверхности плотности встречаются *редко*, поэтому на практике *почти всегда* нейросетевой классификатор эффективно находит главные особенности и направления вариации функции.

- *Качество обобщения и регуляризация.* Суть проблемы состоит в том, что при улучшении качества аппроксимации обучающих данных возникает переход к их прямому запоминанию, при этом теряются обобщающие свойства модели. К настоящему моменту разработан широкий круг статистических алгоритмов оценки качества обобщения (см. [16], с. 376). Обычно оценка ошибки представляется в общей форме

$$\textit{Ошибка обобщения} = \textit{Ошибка обучения} + \\ + \textit{Штраф за сложность модели}$$

Соотношение между двумя составляющими ошибки наиболее последовательно оценивается при Байесовом обучении. Строго говоря, в нашем подходе используются два типа регуляризации - выбор аппроксиматора контролируемой сложности и регуляризация шумом в данных. При использовании нескольких выборок из шумового распределения каждая точка сигнала, образно говоря, окружается облаком шума, что препятствует прямому запоминанию. · Соотношение между двумя составляющими ошибки наиболее последовательно оценивается при Байесовом обучении. Строго говоря, в нашем подходе используются два типа регуляризации - выбор аппроксиматора контролируемой сложности и регуляризация шумом в данных. При использовании нескольких выборок из шумового распределения каждая точка сигнала, образно говоря, окружается облаком шума, что препятствует прямому запоминанию.

- *Эффективные обучающие выборки.*Здесь проблема заключается в том, что число примеров, генерируемых аппроксимируемой плотностью конечно и задано заранее, в то время как имеется полная свобода в генерации данных "шумового"распределения. Какой выбор данных предпочесть? Ответ[7] на этот вопрос рассмотрен в следующем разделе.

3 Бутстреп-выборки

Для построения семейства равномощных обучающих выборок их экспериментальных точек и базового однородного распределения необходимо иметь возможность порождать новые наборы экспериментальных данных. Генерация новых выборок из базового равномерного распределения p_0 не представляет проблем, а что делать с исследуемым экспериментальным распределением? Один из возможных ответов на этот вопрос - метод бутстреп[8] - предложен, по-видимому, Брэдом Эфроном [20] в начале 70-х годов.

Метод, в целом, основан на следующем наблюдении. Для множества точек в многомерном пространстве плотность в форме суммы дельта-функций обладает максимальным правдоподобием (см. начало лекции). Следовательно, если для порождения новых выборок пользоваться этим распределением, то это эквивалентно выборкам из этого множества некоторого нового множества точек *с возвратом*. В выборке будут повторения, но, в некотором смысле, такая выборка статистически распределена так же, как и исходное множество точек. Вводные лекции по бутстреп-методам имеются в Интернет [21], поэтому мы не будем излишне подробно останавливаться на теоретических вопросах. Отметим лишь, что число различных бутстреп-выборок длины N из совокупности N наблюдений равно C_{2N-1}^{N-1} . Это значение легко получить, если рассмотреть задачу размещения $N-1$ перегородок в цепочке из N шариков. Важно, что число вариантов для практических интересных объемов данных заведомо велико (см. Таб 1.).

[7]По-прежнему, уровень строгости и обоснованности даваемых рекомендаций далек от уровня формулировок теорем. Проблема здесь - методологическая. За редким исключением невозможно высказать абсолютно верное (и при этом практически полезное!) утверждение относительно некоторого конкретного набора данных. 10-летний опыт автора показывает, что каждый раз в новой задаче данные устроены совершенно по-разному, и часто исследование приходится начинать практически с нуля. Собственно, это и имел в виду John Tukey, еще в 60-х годах отделяя область анализа данных от матстатистики [19].

[8]Бутстреп - транслитерация английского Bootstrap (дословно "тянуть за застежки ботинок"), означающая "использовать существующий вариант системы или процесса для создания нового варианта"(Lingvo 5.0). В компьютерной литературе этим термином называются самостартующие программы.

Таблица 1. Зависимость числа различных бутстреп-выборок от размера исходной совокупности данных

N	5	10	20	30
Число выборок	125	93539	6.93×10^{10}	$5.9\times.10^{16}$

Имея в своем распоряжении неограниченное число выборок из шумового распределения и соответствующие бутстреп-выборки для полезного сигнала, можно применять постраничное обучение. Это обучение снижает риск систематического смещения результата, вызванного фиксированностью данных, так как на разных эпохах обучения нейросети могут использоваться различные обучающие выборки.

Поскольку многомерные задачи аппроксимации могут приводить к необходимости обучения нейросетей с большим числом неизвестных весовых коэффициентов, в этой работе мы отказались от быстросходящихся методов на основе обращения Гессиана (или его приближений) и использовали адаптивный метод *RProp* (см. Приложение А).

4 Численные эксперименты

В этом разделе мы попытаемся проиллюстрировать особенности предложенного метода аппроксимации на реалистичных задачах. Соответствующие базы данных доступны в Интернет [22], поэтому возможна свободная независимая апробация.

4.1 Задача Banana

В этой игрушечной задаче о двух измерениях предлагается построить аппроксимацию плотности по сложно распределенным на плоскости выборкам точек из нескольких пятен сложной формы (см. Рис 4). Данные для экспериментирования доступны в Интернет[9]

В исходной постановке точки считаются принадлежащими двум классам (точки и плюсы на Рис. 4), и требуется решить задачу классификации. В расширенной постановке переменная признака класса, принимающая значения 0 или 1, добавляется к исходным двум координатам точек, и предлагается построить аппроксимацию плотности в совокупном пространстве 3-х измерений. После успешного обучения нейросетевого классификатора, с полученным распределением плотности можно решить несколько различных задач. При этом мы будем придерживаться последовательного Байесового подхода к вычислениям, а именно:

- Построенная аппроксимация плотности является "объективной"в том смысле, что в ней отражена лишь та информация, которая имелась в данных.

[9]URL: http://ida.first.gmd.de/ raetsch/data/banana/banana.data.tar.gz

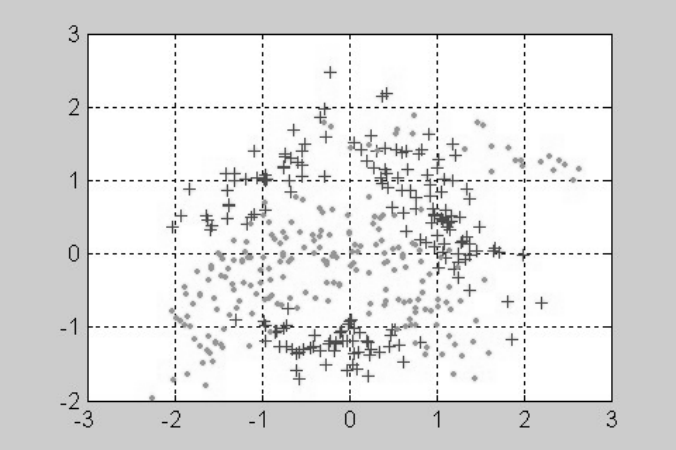

Рис. 4: Исходные данные в задаче Banana

- При решении конкретной задачи исследователь *добавляет* новую априорную информацию в ее условии и интересуется тем, как знание этой дополнительной информации в сочетании с "объективной"плотностью отразится на апостериорных условных распределениях.

Выясним, например, как распределены примеры класса 0 (точки на Рис.4), имеющие координату y, равную 0.5? В постановке задачи содержатся дополнительные сведения о распределении двух из трех переменных (а именно, две переменные известны достоверно). Для ответа на вопрос задачи формально требуется вычислить свертку 3-х мерной плотности распределения с двумя дельта-функциями - или, что то же самое, нормированную зависимость плотности от одной переменной при фиксированных значениях двух оставшихся. Типичные результаты приведены на Рис. 5.

Другой вопрос: каковы относительные вероятности встретить примеры разных классов в окрестности координаты $(1.5, 0.5)$? Ответ - появление плюсика в этой точке в $1,5$ раза вероятнее, чем точки.

4.2 Задача прогноза загрузки процессора ЭВМ (CompAct)

Рассмотрим более реалистичную задачу прогноза доступности процессора ЭВМ для пользователя в реальных условиях эксплуатации (этой ЭВМ). База данных для выбранной задачи (DELVE repository) содержит измерения параметров системной и аппаратной активности в компьютере $SunSparcstation20/712$ с $128M$ памяти, работающем в многопользовательском режиме в условиях университета. Пользователи обычно выполняют большое количество разнообразных задач от доступа в

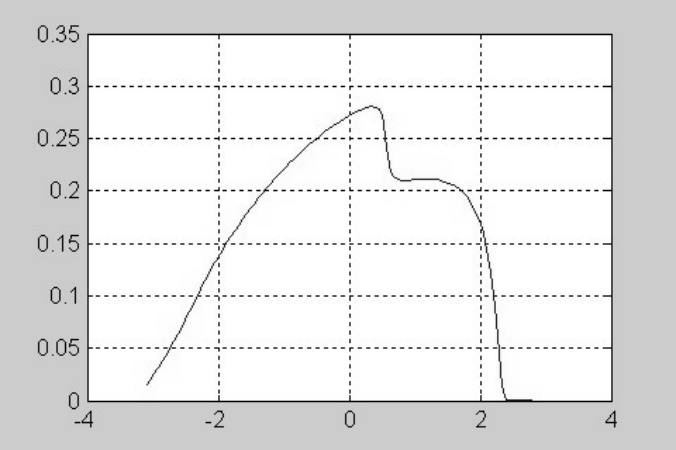

Рис. 5: Апостериорное распределение точек выделенного класса с известным значением одной из координат

Интернет и редактирования файлов до проведения ресурсоемких расчетов. Данные о состоянии системы записывались 1 раз в 5 секунд. В результате получена 8192 запись с 13 параметрами (см. Таб. 2).

Таблица 2. Параметры задачи о загрузке ЭВМ.

Обозначение	Описание параметра	Мин.	Макс.
lread	Число операций чтения-передач в сек. между системной и пользовательской памятью	0	1845
Lwrite	Число операций записи-передач в сек. между системной и пользовательской памятью	0	575
scall	Число системных вызовов всех типов в сек.	109	12493
sread	Число системных вызовов на чтение в сек.	6	5318
Swrite	Число системных вызовов на запись в сек.	7	5456
fork	Число системных вызовов "fork"в сек.	0	20
exec	Число системных вызовов "exec"в сек.	0	60
rchar	Число символов в сек., передаваемых посредством системных вызовов на чтение	278	2526649
wchar	Число символов в сек., передаваемых посредством системных вызовов на запись	1498	1801623
runqsz	Размер очереди процессов	1	2823
freemen	Число системных страниц, доступных для пользовательских процессов	55	12027
freeswap	Число блоков диска, доступных для своппинга страниц	2	2243187
usr	Доля времени (в процентах) в течении которого процессоры работают в режиме непосредственного обслуживания пользователя (user mode)	0	99

Суть традиционной постановки задачи сводится к прогнозу последней переменной - пользовательского КПД системы - по значениям остальных параметров. Исчерпывающее решение этой задачи дано в проекте *Delve*[10]. В обсуждаемой в контексте этой лекции методике для реше-

[10]http://www.cs.toronto.edu/ delve/data/datasets.html

ния требуется[11] построить 13-мерную совместную плотность распределения параметров. Для аппроксимации плотности была использована нейронная сеть с 26 нейронами на скрытом слое. Однако, имея в распоряжении такую аппроксимацию, можно рассмотреть и более интересные аналитико-информационные задачи, нежели просто прогноз переменной *usr*. Например, пусть все параметры кроме размера очереди процессов принимают свои средние значения. Как в этих условиях доля пользовательского процессорного времени зависит от размера очереди процессов? Результирующая зависимость приведена на Рис. 6.

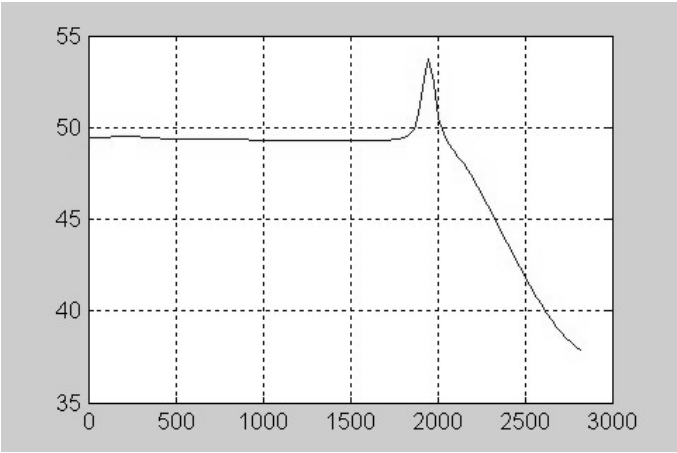

Рис. 6: Зависимость доли пользовательского времени процессора от размера очереди процессов при типичных значениях остальных параметров

Видно, что зависимость носит не очевидный нелинейный характер. Такого типа результат не может быть получен путем простой нейросетевой регрессии.

5 Обсуждение

Статистическое описание на основе совместной плотности распределения вероятности приобретает "второе дыхание"вследствие бурного развития вычислительной техники. Так, если совсем забыть об оптимизации кода, то $1,000,000$ вычислений значения нейросетевой аппроксимации функции плотности от 10 переменных занимает 15 секунд в системе *MatLab* на обычном персональном компьютере. Это позволяет приме-

[11] Следует отметить, что если пользователя интересует лишь прогноз 13-й переменной по значениям остальных 12-ти, то такая частная задача может быть решена путем применения обычной регрессии (например, с использованием нейронной сети). Нашей же целью является построение решения общей задачи прогноза характера распределения части признаков по известной информации об остальных переменных.

нять для оценок условных вероятностей прямые методы Монте-Карло, традиционно считающиеся самыми вычислительно трудоемкими.

Одна из основных проблем, препятствующих созданию полностью автоматизированных методов на основе аппроксимации плотности, как всегда, уходит корнями "в проклятие размерности в многомерном пространстве при типичных объемах данных и типичной сложности задачи плотность распределения сконцентрирована в крошечных областях, объем которых ничтожно мал в сравнении с априорным исследуемым объемом. Тем самым крайне затрудняется получение "полезной"статистики.

При решении практических задач обсуждаемыми методами роль эксперта - специалиста в предметной области - состоит в формулировке априорного знания о параметрах задачи в терминах распределений вероятности и интерпретации полученных апостериорных распределений, а роль специалиста в вычислительной математике и информатике сводится к построению эффективных аппроксимаций и получению достоверных оценок при вычислении интегралов методами Монте-Карло. Роль компьютера - выполнять операции над числами.

6 Благодарности

При подготовке этой лекции автор общался со многими специалистами в фирме "НейрОК"и за ее пределами, всем им большое спасибо. Ю.В. Тюменцев взял на себя нелегкий труд приведения всех лекций в единую систему. Особо хочется поблагодарить А.Н. Горбаня и Н.Г. Макаренко за полезные обсуждения и советы, а также Livermore Software Technology Corporation за финансовую поддержку. Сотрудники Н.Г. Макаренко оказали неоценимую помощь в использовании системы TEX.

7 Литература

1. *Венцель Е.С.* Теория вероятностей. – М.: Высшая Школа,2001.
2. *Терехов С.А.* // Сб. Нейроинформатика. – (А.Н.Горбань, В.Л. Дунин-Барковский, А.Н.Кирдин, Е.М.Миркес, А.Ю.Новоходько, Д.А.Россиев, С.А.Терехов, М.Ю.Сенашова, В.Г.Царегородцев.) Новосибирск, Наука, 1998, С.101-136.
3. *Бердышев В.И.,Петрак Л.В.* Аппроксимация функций. Сжатие численной информации. Приложения. –Екатеринбург, 1999.
4. *Тихонов А.Н., Арсенин В.Я.* Методы решения некорректных задач.– М.: Наука, 1974
5. *Radford M.Neal* Probabilistic Inference Using Markov Chain Monte Carlo Methods.–Technical Report CRG-TR-93-1, 25 Sep 1993, Dept. of Computer Science, University of Toronto.

6. *Ежов А.А.* // Дискуссия о нейрокомпьютерах - 10 лет спустя. Материалы круглого стола "Нейроинформатика-99". М. МИФИ, 2000, С.58
7. *Ghahramani Z.,Jordan M.I.* Learning from Incomplete data. //MIT AI Memo-1509, 1994.
8. *Jordan M. I., Jacobs R.A.* Hierarchical Mixtures of Experts and the EM Algorithm.//MIT AI Memo 1440, 1993, URL:ftp://publications.ai.mit.edu/ai-publications/1000-1499/AIM-1440.ps.Z
9. *Zeevi A.J., Meir R.* Density Estimation Through Convex Combinations of Densities: Approximation and Estimation Bounds.// Neural Networks. – 1996. – v.10. – pp.99–109.
10. *Jonathan Q. Li Andrew R. Barron.* Mixture Density Estimation. // NIPS'99
11. *Perry Moerland.* Mixtures of latent variable models for density estimation and classification.//IDIAP-RR 00-25, 2000 URL: ftp://ftp.idiap.ch/pub/reports /2000 /rr00-25.ps.gz
12. *Терехов С.А.* Вейвлеты и нейронные сети.// Лекция по нейроинформатики., М. МИФИ, 2001, С.142-182
13. *Giulio D'Agostini.* Bayesian Reasoning in High Energy Physics - Principles and Applications.// CERN Yellow Report 99-03, July 1999.
14. *Heckerman D.* A Tutorial on Learning with Bayesian Networks.//Microsoft Tech Rep MSR-TR-95-6, 1995
15. *Minka Th.* Independence Diagrams.//Tech Rep. MIT, 1998. URL: http:// www-white.media.mit.edu/ tpminka/papers/diagrams.html
16. *Bishop C.M.* Neural Networks for Pattern Recognition. Oxford University Press, 1995.
17. *Горбань А.Н.*Обобщенная аппроксимационная теорема и вычислительные возможности нейронных сетей.// Сибирский Журнал Вычислительной Математики. – 1998. – Т.1 – С. 11-24.
18. *Likas, A.* Probability Density Estimation Using Artificial Neural Neural Networks.// Computer Physics Communications. – 2001. – vol. 135, no. 2, pp. 167-175.
19. *Donoho D.L.* Aide-Memoire. High-Dimensional Data Analysis: The Curses and Blessings of Dimensionality.// AMS "Math Challenges of the 21st Century", Stanford, August 8, 2000.
20. *Efron B., Tibshirani R.* Cross-Validation and the Bootstrap: Estimating the Error Rate of a Prediction Rule.//Stanford University Technical report, May 1995.
21. *Holmes S.*Course 208 Lectures. Introduction to the Bootstarp.//tanford University, 1999 URL: http://www-stat.stanford.edu/ susan/courses/s208/web1.html
22. *Blake C.L., Merz C.J.*UCI Repository of Machine Learning Databases.//1998. URL: http://www.ics.uci.edu/ mlearn/MLRepository.html
23. *Гилл Ф.,Мюррей У.,Райт М.* Практическая оптимизация.– М.: Мир, 1985.

24. *Riedmiller M., Braun H.* A direct adaptive method for faster backpropagation learning: The RPROP algorithm.//Proceedings of the IEEE International Conference on Neural Networks (ICNN),H. Ruspini, editor, P. 586 - 591, San Francisco, 1993.
25. *Radford M. Neal* Learning Stochastic Neural Networks. Technical Report CRG-TR-90-7, 1990, Dept. of Computer Science, University of Toronto.
26. *Serge A. Terekhoff* Direct, Inverse and Combined Problems in Complex Engineered System Modeling by Artificial Neural Networks. Proc. SPIE AeroSense Conference, Orlando, Florida, 21-24 April 1997. Vol.3077, Paper 71.
27. *Терехов С.А.* Лекции по теории и приложениям искусственных нейронных сетей. Снежинск, 1994-1998. URL: http://alife.narod.ru/lectures/neural/Neu_index.htm

8 Приложение А.

Эффективное обучение больших нейронных сетей.

Задачи аппроксимации плотности при больших размерностях пространства признаков и, соответственно, больших объемах данных могут приводить к необходимости обучения нейросетей с большим числом неизвестных параметров (синаптических весов нейронов). Это обстоятельство может ограничивать применимость мощных методов обучения типа Левенберга-Маркара или BFGS, т.к. в них требуется решение плохо обусловленных систем линейных уравнений высокой размерности [23]. С другой стороны, традиционные схемы градиентного спуска с дифференцированием нейросети методом обратного распространения крайне медленно сходятся.

Напомним, что обучение нейросети обычно связывается с движением против градиента ошибки с целью ее минимизации:

$$\triangle W_{ij}(n) = -\epsilon \frac{\partial E(n)}{\partial W_{ij}}$$

Более удачное направление поиска для сложных поверхностей ошибки может быть получено путем использования "момента", т.е. памяти о направлении движения на предыдущем шаге:

$$\triangle W_{ij}(n) = -\epsilon \frac{\partial E(n)}{\partial W_{ij}} + \mu \triangle W_{ij}(n-1)$$

В методе *RProp* (*R*esilient *Prop*agation - "эластичное"распространение), предложенном в [24], коррекция каждого синаптического веса зависит только от знака производной и от поправки на предыдущем шаге. При этом поправка для каждого веса индивидуальна и адаптивна.

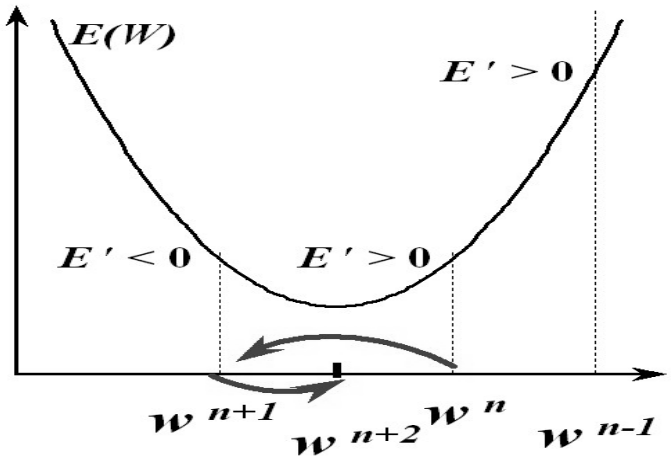

Рис. 7: Идея алгоритма оптимизации методом *RProp*

В случае, если компонента градиента не изменила знак по сравнению с предыдущей итерацией, оптимизация проходит "гладко", и можно увеличить шаг *для этой компоненты*. В противном случае шаг уменьшается (см. Рис 7):

$$\sigma = sign\Big(\frac{\partial E(k-1)}{\partial w_{ij}}\frac{\partial E(k)}{\partial w_{ij}}\Big)$$

$$\Delta_{ij}^{(k)} = \begin{cases} min\{\eta^{+}\Delta_{ij}^{(k-1)},\ \Delta_{max}\},\ \sigma = 1 \\ max\{\eta^{-}\Delta_{ij}^{(k-1)},\ \Delta_{min}\},\ \sigma = -1 \\ \Delta_{ij}^{(k-1)}, \quad \sigma = 0 \end{cases}$$

Поправка к значению веса синапса вычисляется по формуле:

$$\triangle W_{ij}^{(k)} = \begin{cases} -sign\Big(\frac{\partial E(k)}{\partial w_{ij}}\Big)\Delta_{ij}^{(k)},\ \sigma \geq 0 \\ -\triangle W_{ij}(k),\ \sigma < 0 \end{cases}$$

На практике используются следующие типичные значения для параметров, вариация которых мало меняет картину:

$$h^{+} = 1.2,\ h^{-} = 0.5,\ D_{min} = 0.0001,\ D_{max} = 50.$$

Как видно из изложения, метод *RProp* является не чем иным, как алгоритмом оптимизации функции многих переменных. Никакая специфика оптимизации именно ошибки нейросети не использована. Именно в таком виде в завершении лекции хотелось бы привести этот алгоритм оптимизации в виде программы для *MatLab*.

```
% RPROP          Поиск безусловного минимума функции многих
%                переменных методом RProp
%
% (c) 2002, Сергей А. Терехов
%
function [x, fx, epoch, dnorm] = ...
    rprop( fgrad, x, max_epochs, min_dnorm )

    % -- Константы метода RProp
    delta_0   = 0.01;
    delta_max = 50;
    delta_min = 1e-6;
    delta_inc = 1.2;
    delta_dec = 0.5;

    % -- Инициализация
    delta_x = delta_0*ones(size(x));
    grad_x = zeros(size(x));
    grad_sign = zeros(size(grad_x));

    alldone = 0;
    epoch = 0;

    % -- Основной цикл поиска минимума
    while ~alldone
        % -- Вычислить градиент оптимизируемой
        %    функции в точке с координатами x
        [grad_x, fx] = feval( fgrad, x );

        % -- Применить RProp алгоритм для
        %    вычисления индивидуального шага по
        %    каждой координате
        wrk = grad_x.*grad_sign;
        delta_x = ((wrk>0)*delta_inc + ...
            (wrk<0)*delta_dec + ...
            (wrk==0)).*delta_x;
        grad_sign = zeros( size(grad_sign) );
        grad_sign = grad_sign + ...
            ( (wrk>=0)&(grad_x>0) ) - ...
            ( (wrk>=0)&(grad_x<0) );
        delta_x = min(delta_x, delta_max);
        delta_x = max(delta_x, delta_min);
        dx = - delta_x.*sign(grad_x);
```

```
        % -- Сделать шаг по координатам
        x = x + dx;

        % -- Контроль останова итераций
        dnorm = norm( dx ) / ...
            ( sqrt(prod(size(dx))) + eps );
        epoch = epoch + 1;
        alldone = (epoch >= max_epochs) | ...
            (dnorm <= min_dnorm);
    end
return;

% TEST           Тестовая функция для алгоритма RProp
%
function [grad, fun] = test( x )
    ndim = length( x );
    fun = 0;
    grad = zeros( size(x) );
    for j = 1:ndim
        fun = fun + 0.5*( x(j) - j )^2;
        grad(j) = x(j) - j;
    end
return
```

Вызов программы из командной строки MatLab весьма прост:

```
[x, fx, epoch, dnorm] = ...
    rprop( 'test', [11 3 19 124], 100, 1e-6 );
```

Если программа не работает, то попробуйте удалить комментарии на русском языке (или замените их английскими). Для обучения нейросети или решения других задач оптимизации функций большого числа переменных от пользователя требуется запрограммировать вид функции, названной здесь *test* (и, по возможности, ознакомиться с книгой [23]).

Сергей Александрович ТЕРЕХОВ, кандидат физико-математических наук, зав. лабораторией искусственных нейронных сетей СФТИ, зам. генерального директора ООО "НейрОК". 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.