模式识别和机器学习

庞廷海[1]

四川大学计算机学院

2015年9月13日

[1]Email： pthaike@gmail.com

0.1 引言

尽管机器学习发展在计算机科学之外了，但是模式识别仍然根源于工程。然而，这些行为可以被看做一个领域的两面性，在过去的十年里，他们都已经经历了很长的发展。尤其是贝叶斯方法，已经从专业领域发展成为主流，同时图模型已经发展成为一个描述和应用概率模型的通用框架。通过估计推理算法范围的发展，如变分贝叶斯（ variational Bayes）和期望传播，贝叶斯的实践运用已经大大增强。同样的，基于核（kernel）的新方法已经对算法和应用产生重要影响。

这本书反映了这些年的发展，同时提供了模式识别和机器学习领域的全面介绍。其主要正对优越的本科生和博士第一年的学生，如研究者和专业人才。并且假设以前没有模式识别和机器学习概念。需要有多元微积分和线性代数的基础知识，如果熟悉概率将会更有帮助，虽然不是必须的，因为本书也包含了对基本概率理论的介绍。

因为本书视角很广，因为就不能提供完整的参考文献列表，特别是没有精确提供原始的思路。相反的，目的是给出参考，提供更多的细节比希望提供进入点更有可能，在某些例子中，只是一个非常广泛的文献而已。因此，参考文献提供更多最近的教科书和评论文章，而不是原始资源。

这本书需要很多额外的材料支持，包括笔记和本书完整的图片集合。鼓励读者访问网站了解最近的信息

`http://research.microsoft.com/~cmbishop/PRML`

0.1.1 练习

出现在每章节后面的练习是成本书的重要组成部分。每个练习都仔细挑选来加强在书中的概念解析，或者通过显著方式制定和推广他们，并且每个练习都根据其难度，用($\star$) 表示一个需要几分钟完成的简单练习到 ($\star\star\star$)表示明显复杂。

很难知道到什么程度，这些解决方案才能被广泛使用。这些事情需在自学中将会找出有效的答案非常有益，尽管很多课程助教都要求需要提供答案只能通过出版社使得练习可以在课堂中使用。为了满足这些冲突的要求，这些练习帮助强化书中的主要观点，或者填补重要的细节，可以在树的网站上找

到PDF的答案。例如练习WWW表示的。剩下的练习可以通过练习出版社就可以获取（练习方式在书的网站上给出了）。强烈建议读者独立完成练习，在需要的时候再找答案。

尽管这本书专注于概念和原理，在教学过程中，学生最好使用一些合适的数据集来尝试一些主要算法的实验。一个姐妹篇 (Bishop and Nabney,2008) 将解决模式识别和机器学习的实践方面，并且是通过MATLAB软件来实现了在本书中讨论的大部分算法。

0.1.2 致谢

首先我要表达我对Markus Svensen真挚的谢意，他提供了巨大的帮助，准备了图片并且用LATEX对本书进行排版，他的帮助是宝贵的。

我非常感谢微软研究院提供的高度激励的研究环境和给我自由时间来写书（稳重的观点只是来自我的，和微软或其关联的公司不一定相同）。

==========

0.2 数学符号

我尝试保持书的数学内容尽量少来实现该领域的正确理解。然而这种最小水平是，并且强调一个对微积分，线性代数和概率理论对清晰地理解现代模式识别和机器学习技术。不过这本书强调的不是数学困难，而是强调概念。

我已经尝试在整本书中使用连续的符号，尽管这意味着从一些对应的研究文献中分离出来。向量使用小写粗体罗马字符表示，如x，并且所有的向量都假设为列向量。上标T表示向量或者矩阵的转置，所以x^T表示为行向量。大写粗体罗马字母，如M表示矩阵。符号$(w_1, \ldots, w_M)$表示M元素的一个行向量，对应的列向量表示为$w = (w_1, \ldots, w_M)^T$。

符号$[a, b]$表示a到b的闭区间，这个区间包括a和b，而(a, b)对应于开区间，排除a和b。同样地，$[a, b)$表示包括a但不包括b的区间。然而对于大多数情况，不总是需要精确地包括区间的端点。

单位矩阵$M \times M$表示为I_M，也可以简写为I，这里不会对其维度产生歧义。当$i = j$并且$i \neq j$时，它的元素I_{ij}等于0。

一个功能表示为$f[y]$，其中$y(x)$是一些函数，功能得概念在附录D中讨论。

符号$g(x) = O(f(x))$表示$|f(x)/g(x)|$当$x \to \lim\inf$时是有界的。例如当$g(x) = 3x^2 + 2$时,$g(x) = O(x^2)$。

一个函数$f(x, y)$关于随机变量x的期望表示为$\mathbb{E}_x[f(x, y)]$。

在一些情况下，作为变量的均值也没有歧义，这可以通过省略后缀来简化，如$\mathbb{E}[x]$。如果x是以z为条件的分布，那么对应的条件期望写作$\mathbb{E}_x[f(x)|z]$。同样地，变量表示为$var[f(x)]$，并且向量变量的协方差写作$cov[x, y]$，我们也用$cov[x]$来简写$cov[x, x]$。期望和协方差的概念在1.2.2节来介绍。

如果我们有N

目录

第 1 章介绍

在数据中查找模式的问题是一个基础性问题，并且有很长而成功的历史。例如，Tycho Brahe在16世纪广泛的天文观察是的开普勒发现了天体运动规律，这又提供了经典力学发展的跳板。同样的，原子光谱规律的发现也对量子物理的发展和验证起到了关键的作用。模式识别领域关注于使用计算机算法来自动地发现数据中的规律，并且使用这些规律进行如对不同类别进行分类的一些活动。

思考手写数字的识别例子，如图1.1。每个数字对应28×28个像素，可以使用包含784个实数的向量$\mathbf{x}$来表示。目标是去构建一个机器，向量$\mathbf{x}$作为输入，产生数字$0, \ldots, 9$作为输出。可以使用手工规则或者启发式来根据笔画形状来区分数字，但是这种方法在实践中会导致增加例外的规则，并且不约而同地给出糟糕的结果。

更优的结果是采用机器学习方法，这种方法中有一个很大的数字集合N $\{\mathbf{x_1}, \ldots, \mathbf{x_N}\}$ 称为训练集，用来调整得到可适应模型的参数。在训练数据集中的数字分类已经提前给出，通常通过单独手工贴标签来检查他们。我们可以使用目标向量$\mathbf{t}$来表达一个数字的分类，表示对应数字的定义。对于用向量来表示的类别技术会在后面来进行讨论。注意到这里对于每一个数字图像$\mathbf{x}$，使用一个目标向量来表示。

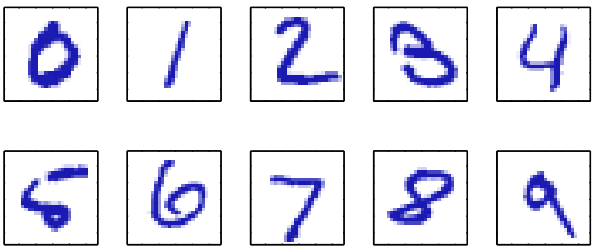

图 1.1: Examples of hand-written digits taken from US zip codes

机器学习算法运行的结果可以表示为一个函数$\mathbf{y}(\mathbf{x})$，函数使用一个新的数字图像$\mathbf{x}$作为属兔，产生一个输出向量$\mathbf{y}$，其和目标向量的编码相同。函数$\mathbf{y}(\mathbf{x})$的精确格式在基于训练数据训练阶段时候确定，也被称作学习阶段。当模型确定后，就可以用来确定一个新的的数字图像，包括一个测试数据集。新样本的分类正确性不同于用来训练数据样本的能力称作一般化（generalization）。在实践应用中，输入的各种各样向量使得训练数据可以包含所有可能输入向量的极小部分，并且这也是模式识别的中心目标。

对于大多数的实践应用，原始输入变量都预处理为新空间的变量，是为了希望可以可以很容易地解决模式识别问题。例如，在数字识别问题中，数字的图像通常被转化或者规范化，从而使得每个数字可以在一个固定大小的框中。这可以减少数字类中的变量，因为数字的位置和规模都相同了。这更容易使用子序列模式识别算法在不同类中区别进行区别。这种预处理阶段有时候称为特征提取（feature extraction）。注意到测试数据必须和训练数据一样使用同样的步骤进行预处理。

预处理也会用来提高计算性能。例如，如果目标是在高速视频流的实时人脸检测，计算机必须每秒处理大量的像素，并且直接展示这给一个复杂的模式识别算法或许计算不可行的。代替地，发现有用特征的目标是能计算更快，并且还保留有用的区别信息来使得可以区别人脸和非人脸。这些特征作为模式识别算法的特征。例如，在矩形子区域上的图像强度均值可以非常有效地进行评估，并且一个特征集可以在快速人脸检测上非常有效。因为这些特征的数量比像素点更少，这种预处理表达了降维的一种形式。必须注意到在预处理过程中，因为信息的丢弃，如果这个信息对于处理问题是重要的，那就可以会遭受系统整体精度的痛苦。

训练数据包含的样本是输入向量对于目标向量的应用是监督学习(supervised learning)问题。例如数字识别样本，其中的目标是将每个输入向量赋值为一个有限的离散数字，称为分类问题(classification)。如果期望的输出包含一个或多个连续的变量，这种任务称为回归(regression)。分类问题的一个例子是预测化工生产过程的产率，其中输入包裹关注的反应物，温度和压强。

在其他的模式识别问题中，训练数据包含一个输入向量集$\mathbf{x}$，没有对应任何的目标值。在这种非监督学习(unsupervised learning)问题可能会在数据中发现一组相同样本，称为聚类(clustering)，或者在一个输入空间中确定一个数据的分布，称为密度估计(density estimation)，或者将高维数据空间投影到两三维，

目的是可视化(visualization)。

最后，加强学习(reinforcement learning, Sutton and Barto, 1998)的技术关注的问题是在给定的语境下采取合适的方法去最大化奖励。对于监督学习，学习算法不会输出一个最优的样本，但必须通过实验和错误来发现他们。通常情况下存在一系列的状态和动作，其中学习算法与它的环境交互。在很多例子中，现有的动作不仅仅影响现在的奖励，并且也会对所有的子序列时间步骤产生影响。例如，通过加强学习的技术，神经网络可以学习到下西洋双陆棋达到一个很高的标准(Tesauro,1994)。在这里，网络必须学会将棋盘的一个随机位置作为输入，接着产生一个强的的移动作为输出。这通过将网络与其自己的一个拷贝来对抗来完成，或许会进行百万次。一个主要的挑战是西洋双陆棋会包含很多种移动，只有在游戏结束的时候，在取得胜利的情况下，奖励才得以实现。奖励必须适当地对所有的移动产生贡献，即使一些移动会有好的情况但其他的会差些。这是一种信贷分配(credit assignment)问题。加强学习的一个通常的特征是探索(exploration)和开发(exploitation)之间的权衡,其中探索是系统尝试新的方法来观察他们将会有怎样的效果，而开发系统会使用已知的动作来产生高的奖励。太关注探索或者开发都将会差生较差的结果。加强学习任然是机器学习研究中一个活跃的课题。然而，详细的讨论会超出本书的范围。

尽管每个问题都需要自己的工具和技术，但是很多关键的想法都支撑着所有的这些问题。这章主要的目标是用相关的非形式化方法去介绍介个最重要的的概念和用简单的例子描述他们。在书的后面，我们将会看到这些想法在再出现在更加复杂的模型中，这些模型会适用于真实世界的模式识别中。这章还会包含三个重要工具的介绍，它们会在整本书中都用到，叫做概率理论，决策理论和信息理论。尽管这些听起来像令人畏惧的话题，但是它们事实上是直接的，并且如果将机器学习技术用到实践应用中，清晰地理解它们是必要的。

1.1 例子：多项式曲线拟合

我们通过引入一个简单的回归问题来开始，我们将会用它作为一个运行的例子来贯穿整章来激励一些关键概念。假设我们观察一个真实值输入变量x，我们希望使用这个观察值去预测真实值目标变量t的值。对于目前的目的，使用合成的方法产生人造样本是具有启发性的，因为当我们进行学习模型比较时候，我们能知道生成数据的精确过程。样本的数据是由在目标值中包含随机噪声的的函数$sin(2\pi x)$产生的，具体描述见附录A.

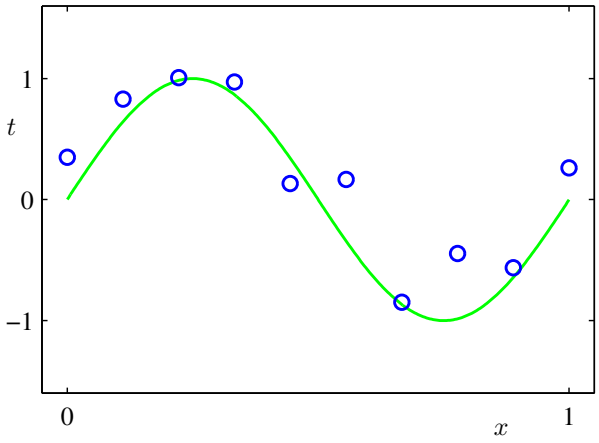

图 1.2: Plot of a training data set of N = 10 points, shown as blue

现在假设我们有一个训练集$\mathbf{x} \equiv (x_1, \ldots x_N)^T$，包含N个x的观察值，对应的t的观察值表示为$\mathbf{t} \equiv (t_1, \ldots t_N)^T$。图1.2展示了包含 $N = 10$ 的训练集的点图。图1.2中的输入数据集$\mathbf{x}$是选择x_n的值产生的，对于$n = 1, \ldots, N$，取值范围为[0,1]，目标数据集$\mathbf{t}$是通过计算函数数$sin(2\pi x)$得到的对应值，并对每一个点添加一个低水平的服从高斯分布随机噪声（高斯分布会在1.2.4节介绍）可得到t_n。通过这种方式产生数据，我们可以得到很多真实数据集的性质，也就是说这些数据存在潜在的规律，我们希望去学习这种规律，但这种独立的观察值受到随机噪声的干扰。这种平噪声可能从本质上体现随机过程，例如放射性衰变，但是通常是因为变量本省是不可观测的。

我们的目标是通过利用训练集，对于一些新的输入变量$\hat{x}$做出预测变量的目标值$\hat{t}$.我们后面可以看到的，这包含隐含地试图找到潜在函数$sin(2\pi x)$.当我们从有限的数据集中概括时，在本质上是一个棘手的问题。更多的，观察数据集受到噪声的干扰，因此对于一个给定的$\hat{x}$，对应的近似值$\hat{t}$是不确定的。在1.2节中讨论的概率理论提供了一个用精确和定量的方法来表达这种不确定性的框架，在1.5节中讨论的决策理论重现了这种概率的利用，为了根据似然准则做出优化决策。

就目前而言，我们将会非正式地进行和考虑基于曲线拟合的的简单方法，我们使用多项式函数来拟合数据得到形式

$$y(x, \mathbf{w}) = w_0 + w_1 x + w_2 x^2 + \ldots + w_M x^M = \sum_{j=0}^{M} w_j x^j \tag{1.1}$$

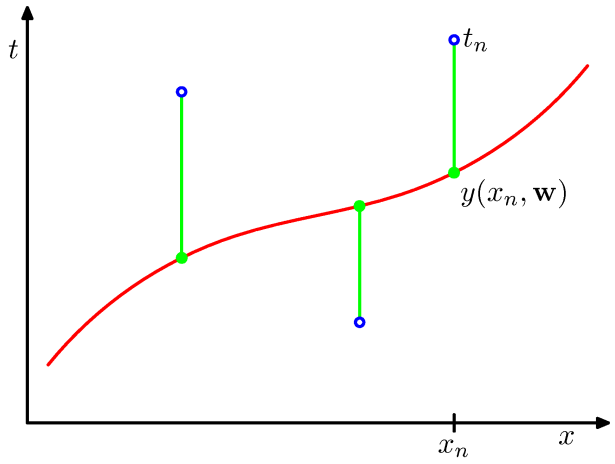

图 1.3: The error function (1.2) corresponds to (one half of) the sum of the squares of the displacements (shown by the vertical green bars) of each data point from the function $y(x, \mathbf{w})$

其中M是多项式的阶，x^j表示x的j次幂。多项式的系数$w_0, \ldots, w_M$可以使用向量$\mathbf{w}$表示。注意，尽管多项式函数$y(x, \mathbf{w})$是x非线性的函数，但对于w是线性函数。如多项式之类的函数，对于未知参数是线性的，具有重要的性质，称为线性模型，我们将会在3.4节来扩展介绍。

参数的值由对训练数据的多项式拟合确定。这可以表示为通过最小化误差函数（error function），对于任意的值(w)和训练集数据点，误差函数用来测量函数$y(x, \mathbf{w})$的失配。简单选择被广泛地使用误差函数，通过对每个数据点x_n的预测值$y(x, \mathbf{w})$和对应的目标值t_n误差平方和,因此我们最小化

$$E(w) = \frac{1}{2}\sum_{n=1}^{N}\{y(x_n, \mathbf{w}) - t_n\}^2 \tag{1.2}$$

其中包含的因子1/2是为了后面计算的方便。我们将会在这个章节的后面讨论使用这个误差函数的原因。现在我们简单地注意到它是一个非负的的等式，当且仅当函数$y(x, \mathbf{w})$精确地通过每一个训练数据点的时候为0。函数平方误差和的几何解析表示为图1.3。

我们可以通过选择$\mathbf{w}$的值来解决曲线拟合问题，其中$E(\mathbf{w})$尽可能小。因为误差函数是参数为$\mathbf{w}$的一个二次函数，它的导数是相对于系数在元素$\mathbf{w}$下是线性的，因此最小化误差函数有唯一的解，可以用紧凑的形式表示为$\mathbf{w}^\star$。多项式的结果表示为函数$y(x, \mathbf{w}^\star)$。

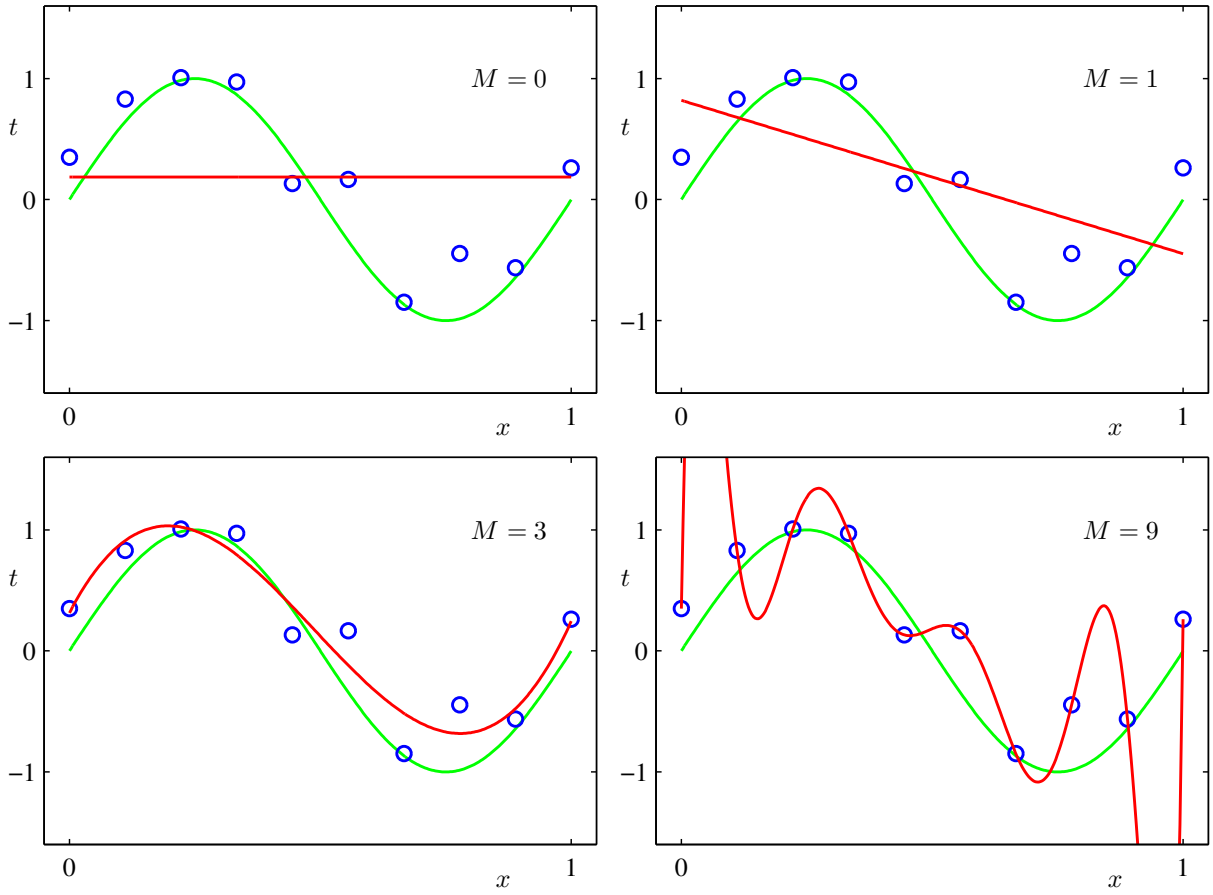

图 1.4: Plots of polynomials having various orders M, shown as red curves, fitted to the data set shown in Figure 1.2.

这里存在的问题是如何选择多项式的阶M，我们会看到这会转化为一个重要概念的例子，称为模型对比和模型选择(model comparison or model selection)。在图1.4中，我们展示了对于图1.2中展示的数据集的多项式拟合结果的4个例子，其中多项式的阶数为M = 0,1,3,9。

我们注意到常量（M = 0）和一次（M = 1）的多项式对于数据得到相当糟糕的拟合,并且对于函数$sin(2\pi x)$表示也表现很差。如图1.4所示的例子，三次多项式对于函数$sin(2\pi x)$似乎得到了最好的拟合。当我们采用更高次的多项式(M = 9)时，对于训练数据，我们得到了一个较好的拟合。事实上，多项式精确地通过每个点，并且$E(\mathbf{w}^\star) = 0$。反而，拟合曲线较大范围的波动对于函数$sin(2\pi x)$会得到较差的表示。后面这种行为称为过拟合（over-fitting）。

正如我们前面所描述的，我们的目标是实现对于新数据的精确预测的归纳。我们可以通过考虑单独的测试数据集定量地对这些在M上独立的归纳模型的性

能进行观察，其中每个数据集精确地使用和生成训练数据集相同的方法生成的100个数据点，但在目标值中使用新的随机噪声。对于每个选择的M，我们可以对训练数据通过(1.2)式子来评价得到的值$E(\mathbf{w}^\star)$，并且我们也可以对于测试数据集来评价$E(\mathbf{w}^\star)$。通常我们使用均方根误差(RMS)来定义更加方便

$$E_{RMS} = \sqrt{2E(\mathbf{w}^\star)/N} \tag{1.3}$$

其中除数M允许我们可以同等地比较不同大小的数据集，并且平方根确保E_{RMS}可以和变量t在同一尺度上进行测量。对于不同的M值，训练和测试数据集的REM误差可以在展示在图1.5中。测试数据集是用来测量对于新数据x的观察结果，得到的预测值t的效果如何。我们从图1.5中可以观察到，对于小的M值，会得到相对较大的测试误差。这可以归因于得到的多项式是相当平滑的，并且不能捕获函数$sin(2\pi x)$的震荡。真如我们所看到的对于图1.4中 $\mathrm{M} = 3$ 的例子，M在$3 \leq M \leq 8$范围内时会得到较小的测试误差，并且这也合理地给出了生成函数$sin(2\pi x)$的表示。

对于 $\mathrm{M} = 9$，正如我们所预料的，训练集误差为0，因为多项式对应的10个系数包含10个自由度$w_1, \ldots, w_9$，所以可以精确地被协调到10个训练数据点。然而，如图1.4所示，测试集误差变得非常大时，对应的函数$y(x, w^\star)$波动很大。

这似乎是矛盾的，因为一个给定阶数的多项式包含所有低阶的的例子。$\mathrm{M} = 9$ 的多项式因此能和 $\mathrm{M} = 3$ 的多项式产生一样好的结果。此外，我们或许可以假设对于新数据的最好预测可能是函数来自于函数$sin(2\pi x)$，其中的数据是由函数生成的（我们后面可以看到这确实是这样的）。我们知道函数$sin(2\pi x)$的一个幂级数展开包含所有的阶，因此我们期望当我们增加M时，性能会单调地提升。

我们可以通过检验从各种阶数的多项式得到的系数值来得到一些深入的问题，如表1.1所示

	M = 0	M = 1	M = 6	M = 9
$w_0^\star$	0.19	0.82	0.31	0.35
$w_1^\star$		-1.27	7.99	232.37
$w_2^\star$			-25.43	-5321.83
$w_3^\star$			17.37	48568.31
$w_4^\star$				-231639.30
$w_5^\star$				640042.26
$w_6^\star$				-1061800.52
$w_7^\star$				1042400.18
$w_8^\star$				-557682.99
$w_9^\star$				125201.43

表 1.1: 不同阶数的.

第 2 章The Second Chapter

2.1 密度估计

密度是一个个嘎发哦发哦合法

2.1.1 高斯密度

2.1.2 高斯密度

第 3 章The First Chapter

3.1 daodaodaodoa