PRML Page-By-Page Book Review

淡蓝小点Bluedotdot.cn

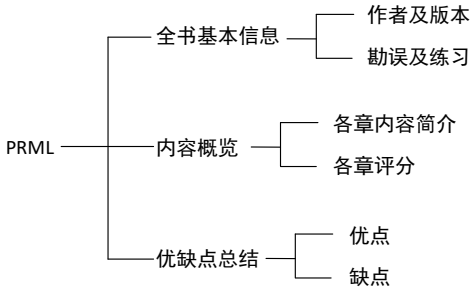
2024 年 4 月 16 日

目录

第零章　Review of whole book

《Pattern Recgonition and Mahcine Learning》是机器学习领域内一本影响力巨大的著作，它内容丰富、坚深翔实，得到了很多从业者、研究者的认可，甚至被誉为机器学习领域的“圣经”。但正因为它又深又博所以阅读起来难度颇大，想要读懂它、读透它是一个不小的挑战。虽然我相信大家只要肯花时间，一定能把它读懂读完，但也免不了要付出巨大的时间成本、消耗巨大的精力。我在淡蓝小点发起了一个名为《PRML Page-by-Page》的项目，目的就是去帮助那些想把PRML读完但又感觉很困难的朋友。在这个项目中我会和大家一起去一页一页甚至一句一句的读这本书。凡是我认为有必要做解释、推导或扩展的地方我都会做详细的说明。全书十四章我标记了几百个这样大大小小的说明点，还整理了一份一千多页的资料。结合这一套视频，我想多少还是会对大家有些帮助。但不得不说即使这样，想要从头到尾把PRML读透仍需要读者付出较大的努力。假设有一名在校研究生，他具有平均水平的基础知识，每周五天、每天投入4小时左右，结合《PRML Page-by-Page》我觉得比较快的话可以在3-6个月左右把PRML读完。如果平时还有其它工作，那需要的时间可能会相应延长。就我个人而言，我读了好几年才把它读完。所以大家不要急，如果能做一个6-8个月的规划已经很不错了。本人水平有限，对这本书的理解肯定也有很多不到位的地方，如果在解读中有错还请大家见谅!

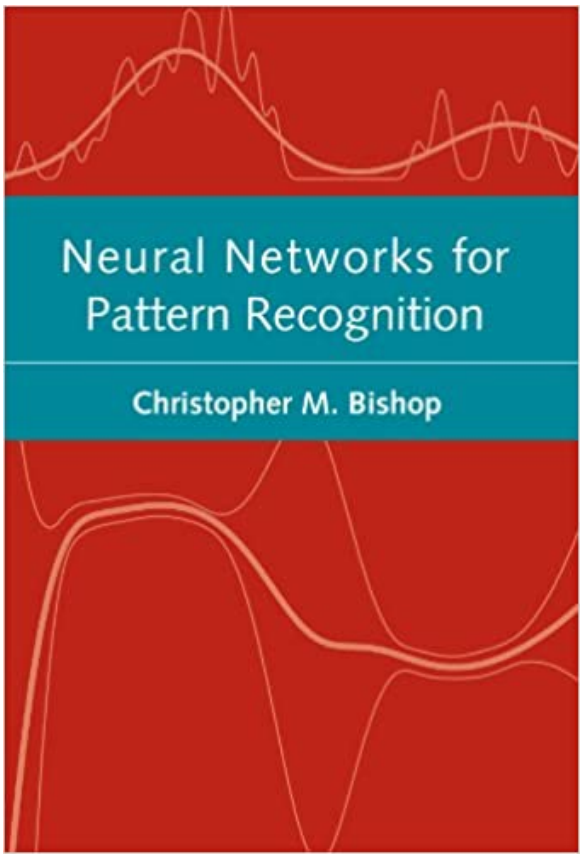
整个Page-by-Page项目是依据PRML的章节设计展开的，在开始阅读这本书之前我们会先对全书做一个概览性介绍，也就是大家现在看到（读到）的内容。在深入到每一章之后，我们也会先对每一章做概览，但那个比较简单一般只有几分钟。好的，让我们开始吧。我们将从以下三方面对全书做概览。

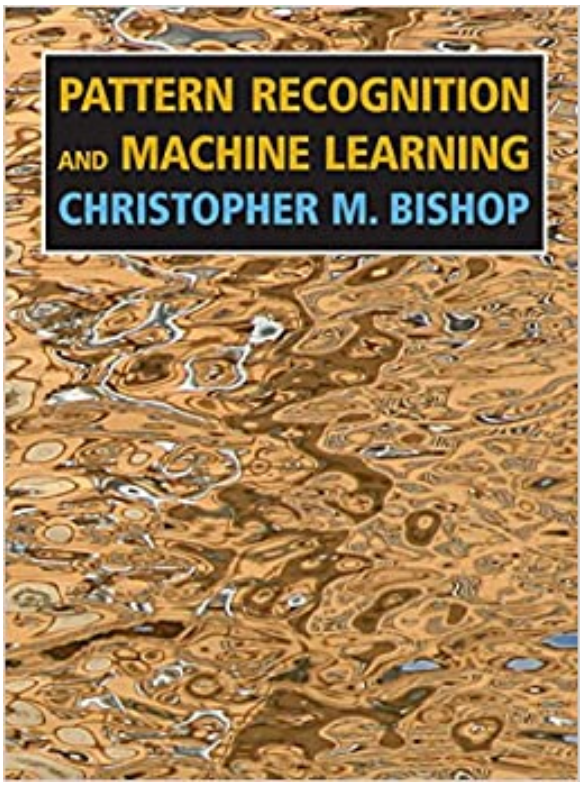


- 作者及版本

本书的作者是克里斯托弗·迈克尔·毕晓普（Christopher Michael Bishop），于1959年4月7日出生于英国。

他其实是物理专业出身，在英国爱丁堡大学获得理论物理学的博士学位，研究方向是量子场理论。后来则致力于模式识别方向的研究。Bishop现在是微软剑桥研究院负责人（英国的剑桥，也叫康桥，不是美国的剑桥），同时也是爱丁堡大学计算机科学的名誉教授。他比较有代表性的研究成果包括PPCA（概率PCA，在第12章会介绍），混合密度网络（Mixture density networks, MDN，在第5章会介绍）等。毕晓普其实写过两本重要的著作，一本是于1995年发表的《Neural Networks for Pattern Recognition》，另一本则是2006发表年的PRML。《Neural Networks for Pattern Recognition》其实也是一本业内非常著名的书，只不过随着时代的发展有更多更好的著作替代了它。如果现在要读一本关于神经网络的书的话，估计很多人的首选会是花书《Deep learning》。相比之下，PRML生命力更持久、名气更大。


Neural Networks for
Pattern Recognition
Christopher M. Bishop



PRML较为全面而深入的介绍了模式识别和机器学习领域的相关内容，它原本是为博士一年级或学有余力的研士生、本科生撰写的教材，当然也适用于那些在相关领域工作的研究者、工程师。虽然在作者看来这本书对读者不做任何前序知识的要求，但实际上我觉得如果缺少必要的知识铺垫读起来会很困难。PRML大概是第一本对模式识别算法、概率图模型、近似算法、贝叶斯视角做系统性介绍的教材，而且写作由浅入深、层层递进，因此造就了PRML在机器学习如今的地位。

PRML本身只是在2006年出版了一版，虽然广大网友想要第二版的呼声很高，但实际上并没有第二版，至少到目前为止还没有。PRML本身是完全开源的，包括教材本身、配套的PPT、插图（PNG/JPG/EPS）都可在网上免费获得。我这里只给一下全书的下载链接。

https://www.microsoft.com/en-us/research/uploads/prod/2006/01/Bishop-Pattern-Recognition-and-Machine-Learning-2006.pdf

但是大家注意，不知道什么原因官方链接下载下来的第12章在排版上有瑕疵。第12章不是很清晰而且部分配图是黑白的，部分页面也无法做文字检索。这一版全书是758页（含封面、封底等）。我从网上下载到了一个排版效果正常的版本，它的第12章排版正常，但我已经不记得是从哪里下载的了。我这一版全书

是749页（含封面、封底等）。

PRML没有真正出版的中文翻译版，但有日语译文版。日语版分上下两卷，卷一含1-5章及附录，卷二含6-14章。日语版的封面如下所示。

网络上有一个公开的PRML汉译版，是一位落款名叫马春鹏的老师或同学翻译的，公开时间是2014年10月26号。这个翻译版中，译者不仅将英文翻译成中文，还直接修正了很多原书中的错误，排版也非常专业。看得出作者花费了巨大的时间和精力。我们应该感谢有这样高水平且热心的研究人员！

有三个与PRML配套的勘误（errata），分别发布于2006、2007和2009年。三个版本的errata中90%以上勘误的都是书写或印刷错误，比如单词拼写错误、大小写错误、加黑不加黑错误等，这些错误几乎不影响对本书的理解。但是也有少量错误非常关键，尤其是公式中的错误。如果这些错误不被纠正的话，读者读到相应章节时可能真的会疑问重重。但是，即使是有三个版本的errata也并没有将书上的关键错误完全识别出来。在我印象中至少还有1-2处关键错误未能被纠正。我们会在Page-by-Page项目中对这些内容作说明。这里给出三个版本的errata下载链接。

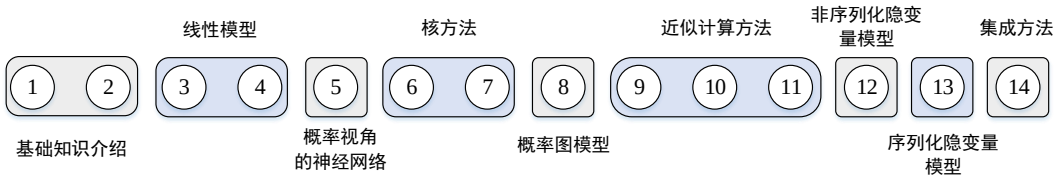
https://www.microsoft.com/en-us/research/wp-content/uploads/2016/05/prml-errata-1st-20110921.pdf

https://www.microsoft.com/en-us/research/wp-content/uploads/2016/05/prml-errata-2nd-20110921.pdf

https://www.microsoft.com/en-us/research/wp-content/uploads/2016/05/prml-errata-3rd-20110921.pdf

另外一份比较重要的资料是PRML的练习答案。网上有一份全体练习的参考答案，是辅助我们理解本书的重要资料。Page-by-Page项目不会去做每一章后面的练习，但我们会介绍大部分在正文中出现且比较重要的练习。因为正文中有些地方的公式推导是被作为了练习。
https://www.microsoft.com/en-us/research/wp-content/uploads/2016/05/prml-web-sol-2009-09-08.pdf

- 内容概览
  PRML全书一共十四章，从头到尾分别是概论、概率基础、线性回归、线性分类、神经网络、核方法、稀疏核机、概率图模型、EM算法、变分法、采样法、连续型隐变量模型、序列化模型和集成模型。日语版翻译把前5章作为一卷、后9章作为一卷，不知道这是否代表在译者眼中可将全书14章按这样分成两部分了？我根据自己的理解给出两种对14章内容的划分方法。第一种是比较细的划分，划分如下。



第一、二两章没有介绍任何具体的模型，只是在为后面的内容做铺垫。第一章介绍了贯穿整个机器学习的一些基本概念，如什么是曲线拟合、什么是概率密度、什么是贝叶斯、什么是模型选择、什么是损失函数等。第二章补充介绍了很多概率统计相关的知识，包括各类常见的分布。其中重点的重点是高斯分布及指数族分布。
第三、四两章介绍了线性模型对于回归问题和分类问题的解法。虽然线性模型是最简单的模型，但PRML中这两章涉及的内容比一般资料更多更深入。
第五章是神经网络。但与现在一般介绍深度学习的资料不同，PRML是从概率的视角介绍神经网络，我个人觉得这是一个不同寻常但又非常有趣的视角。现在流行的深度学习其实在很大程度上是缺乏可解释性的，因为它缺少坚实的理论基础。作者从概率视角看待神经网络一方面可能是历史的局限性，另一方面也可能是想将概率理论作为神经网络的理论基础。
第六、七两章是围绕核方法展开的。第六章是核方法的基础理论，第七章介绍了SVM/RVM两个具体的核方法代表。要读懂第七章一定要先读第六章。另外，

本书从稀疏核的角度来解释SVM也与一般介绍SVM的资料略有不同。核方法是一类对数学基础要求较高、较为抽象的方法，但它也是一类神奇且广泛存在的方法，非常值得大家花时间去学习。例如第六章中介绍的高斯过程，从控制论的角度来看它是一个随机过程；从机器学习的角度来看它是一种核方法；而在深度学习中，它又是量化深度学习中神经网络不确定性的基础（深度学习的不确定性仍是一个开放的待研究课题，但从高斯过程的角度出发、利用高斯分布的方差作为对网络不确定性的量化是一种重要的方法）。

第八章介绍了概率图模型的基本概念，其中最重要的是条件独立性和D隔离。第八章没有介绍任何具体的实用模型，但它是后面几章的总起，地位非常重要。我个人觉得第八章可以说是全书最核心的一章。

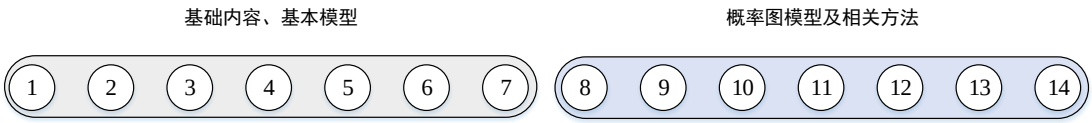
第九、十、十一章都是在介绍近似计算方法，其中第九、十两章是确定性近似，第十一章是随机性近似。第九章的EM算法针对的是有隐变量的模型，第十章的变分法则是对EM算法的进一步拓展。EM算法的计算涉及到后验分布的期望，而变分法则是考虑当后验过于复杂难以参与计算时，用近似分布去替代原有的后验分布。所以说变分法其实是EM算法的一般化推广。第十一章介绍了若干种采样方法，它的重点在于MCMC，但它的难点除了MCMC外还包括混合蒙特卡洛（HMC），主要是其中涉及到哈密顿力学，理解起来有一定困难。这三章的重点不是介绍具体模型，而是求解各类模型的方法。

第十二章以PCA为代表，介绍的是连续型隐变量、非序列化模型。所谓非序列化是指数据呈独立同分步，相互之间没有相关性。PCA是非常常见也非常重要的方法，本书不仅从方差最大化、最小重构误差两个较常规角度介绍了PCA，实际上还提高到流形学习层面做了大量补充。包括MDS（多维尺度）、LLE（局部线性嵌入）等著名模型，能帮读者进一步打开视野。

第十三章以HMM和卡尔曼滤波为代表，介绍的是序列化隐变量模型。所谓序列化是数据不再呈独立同分步，相互之间有相关性。但我们只重点研究一阶模型，即当前时刻的数据只与前一时刻的数据有关。HMM的重要性简直不言而喻，如果你读完全书只能记住一个模型那你就记住它好了（如果你能记住两个那就记住HMM和MCMC，如果能记住三个那就记住HMM、MCMC、EM）。卡尔曼滤也不惶多让，尤其是在这几年兴起的自动驾驶中，对于车辆定位有实际用途。

第十四章是全书最简短的一章，也是难度最小的一章。我甚至觉得它比第一章引论还要简单。

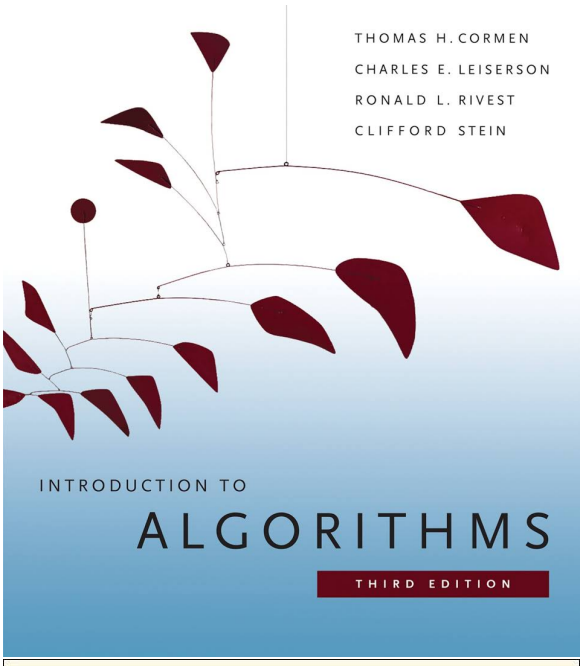
如果我们也按PRML的日译版将十四章分成两部分的话，我会选择将前七章和后七章分开。

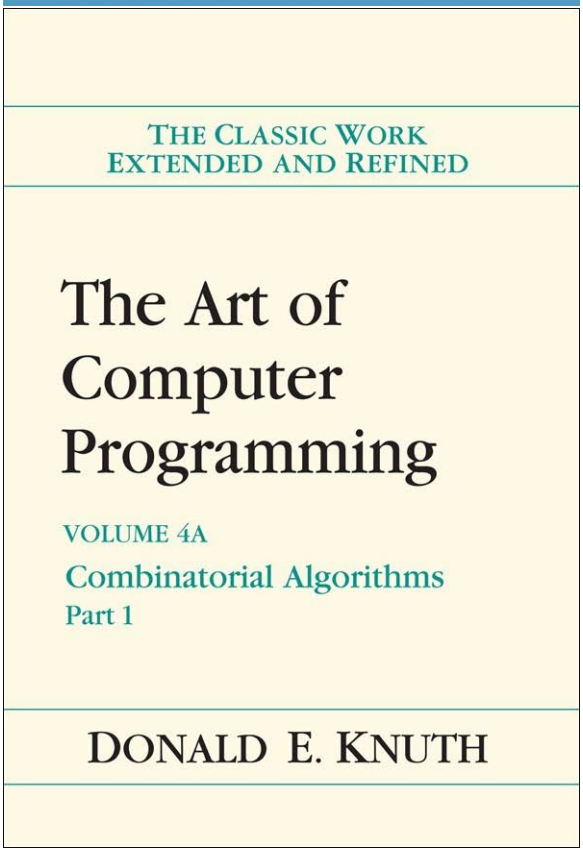


我个人认为前七章倾向于介绍比较基础的内容，如模型选择、Laplace近似、核方法等。前七章也介绍了一些具体模型，如多项式拟合、Logistic回归、SVM等，但总体这些方法都是比较简单、比较基础的。SVM是一个曾经特别流行的模型，但现在风头已被深度学习夺走。我觉得读完前七章就算机器学习的初学者了。

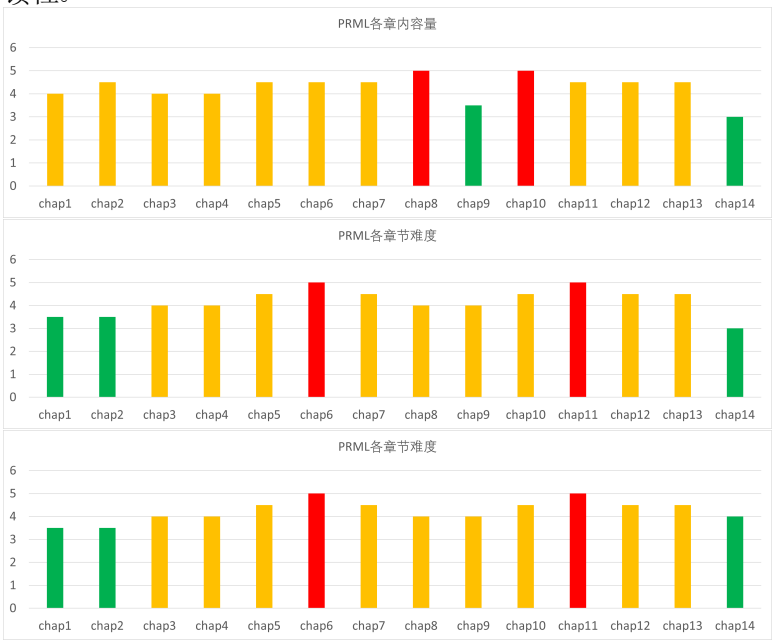
后七章由第八章领衔，集中笔墨介绍了很多非常有用、非常知名的方法和模型。如果说前七章是一些基础性知识的话，后七章的HMM、卡尔曼滤波、PCA、流形学习等方法可用于解决很多复杂的实际问题。我觉得读完后七章就算有一定基础的机器学习学习者了。

说一点题外话。有一本非常有名的书叫《算法导论》(《Introduction to Algorithms》)，它也是一本经典巨著，全书1313页（第3版），可谓深入细致、全面丰富。曾经的我有过这样一个问题，像这样经典深入的书为什么要叫算法“导论”了？所谓导论不就是正式介绍算法前的前导性内容么？如已经如此深入丰富的书还只能叫导论，那到底什么样的书才配得上去掉导论两个字了？直到后来我知道了高纳德（Donald E. Knuth）的《计算机程序设计艺术》(《The Art of Computer Programming》)我才理解为什么《算法导论》要谦虚的称自己为导论了。因为有《计算机程序设计艺术》这样的鸿篇巨制存在，其它关于算法的书似乎都只能算是导论了。所谓“天下武功出少林”，高纳德的这七卷本大概就是算法界的少林吧。题外话，高纳德的七卷本写完了么？是不是还没写完？我有时候就想PRML在机器学习领域的地位到底是《算法导论》了还是《计算机程序设计艺术》了？


THOMAS H. CORMEN
CHARLES E. LEISERSON
RONALD L. RIVEST
CLIFFORD STEIN
INTRODUCTION TO
ALGORITHMS
THIRD EDITION


THE CLASSIC WORK
EXTENDED AND REFINED
The Art of
Computer
Programming
VOLUME 4A
Combinatorial Algorithms
Part 1
DONALD E. KNUTH

下面看几个我个人为各章节的评分，包括三组评分，分别是内容量、难度和可读性。



内容量包括书上本身的内容和我们需要补充介绍的内容。从内容量上来看，第八章概率图模型和第十章变分法是内容最多的两章。这两章主要是原本书上的内容就多，我们补充的内容倒并不多。第六、七两章原本书上的内容并不太多，但需要补充的、关于核方法的知识较多，所以它们的内容量达到4.5分。内容比较多的章节还包括采样、连续型隐变量模型、序列化模型以及第2章概率知识。但很幸运，最多的不一定是最难的、最难的也不一定是最不好读的。全

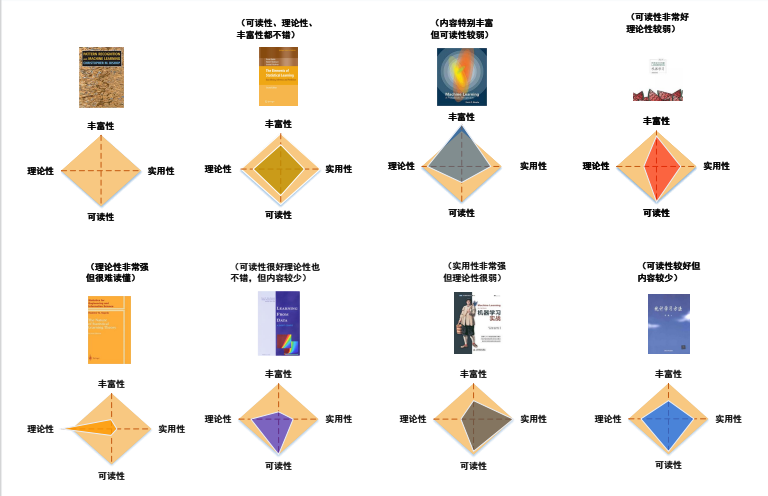
书最困难的章节我觉得是第六章核方法及第11章采样。第六章难是因为有三个概念要补充，分别是：随机过程、RKHS、高斯过程。尤其是其中的RKHS，它是核方法中最核心的概念但本书却省略了它。RKHS涉及到实变、泛函等诸多概念，我们需要先补充介绍这些内容才能真正理解它。第十一章采样原本并不算太困难，但其中11.5节混合蒙特卡洛涉及到哈密顿力学，对于不了解哈密顿力学的同学来说难度较大。大家可以放心的是第八章、第十章，这两章虽然内容很多而且名字听起来很吓人，但其实不算太难，尤其是变分法。这两章的特点是细致繁锁大过困难。而且这两章写作非常到位，读起来会比想像中要轻松一些。读这两章就像吃面包，看着个头很大但吃起来口感还算比较舒适。关于各章节的可读性，我觉得全书写得最好的是第二章。第二章可谓细致、周到、深入、严谨、丰富，我非常喜欢第二章。我个人认为读起来比较困难、不太好理解的是第五、六、七三章。第五章主要是因为视角问题。尤其对于有一定深度学习经验的同学，当你读第五章时可能会觉得这里讨论的重点怎么和一般意义上讨论的神经网络的重点不一样了？这里不太侧重网络结构的设计，也不太侧重各类网络的介绍，而是在介绍如何从概率的角度去看待神经网络。如果说花书是在横向或者向上介绍神经网络的话，那本书的第五章可以认为是在向下介绍神经网络（上是应用，下是理论，尤其是概率统计的理论）。第六七章的可读性得分较低是因为它欠缺的东西有点多，第六章欠了RKHS，第七章欠了对偶问题最优解的原理。但是对于第六、七章甚至第十一章，我第一遍读的时候就一直在想作者为什么要从这样的角度来写，可是等我读完之后回过头来再看我却觉得大概就这应该这样写。因为这样写能省去很多复杂的数学背景，以一种尽量容易被理解的方式把最核心的内容呈现出来。

值得注意的是第9章，它不算太难但又有一定的深度，内容较少且写作易读易懂，并且EM算法本身也非常实用。我觉得它大概会是读者最喜欢的章节之一。上面这些所谓的评分只是我的一家之言。

- 优缺点总结

  PRML的优点不用多说，从它在大家心中的认可度、行业内的地位就能看出它有多优秀。如果一定要总结的话我想概括成以下三方面：一、理论坚实、难度适中；二、内容丰富、模型实用性强；三、行文细腻严谨好读。对于一本如此专业的技术专著，能同时具备以上这三大优点着实不容易。咱们把目前市面上常见的几本有名的书放到一起一对比，大家就能明白为什么即使是出版了这么多年，现在依然没有哪一本能完全超越PRML了。整体上与PRML最接近的应该是ESL（《The Elements of Statistical Learning》）。ESL也非常优秀，但从各方面却又感觉比PRML差了一截。MLAPP（《Machine Learning A Probabilistic Perspectivce》）名气很大，它在内容非富性上甚至超过PRML，但它在可读性上比PRML差远了。我感觉MLAPP更像一本用来查阅的高级技术合辑，而不是一

本好读的书。《学习理论的本质》（《The nature of statistical learning theory》）是一本相对比较薄的书，但它的理论性是所有书里面最强的，重点介绍了PAC理论。从理论性上看它是PRML的老师。但它非常难读也几乎没有实用性。因为比较的时候是从四个方面比较的不太好画图，大家将就着看。花书仅关注深度学习，所以这里没有列入比较。看起来目前的确没有一本比PRML更强的书。



虽然PRML如此优秀，但一本书就像一个人一样，不可能完美无缺，所以我斗胆也说几点PRML的不足之处，说的不对请大家见谅。

第一，PRML在内容上还是不够完整。即使是站在2006年的角度，PRML也缺少了主题模型（LDA）和聚类（Clustering）这两个专题。从如今的视角来看，它还缺少深度学习和图学习这两部分。但是我们也不应苛责作者也不应苛责这本书。PRML毕竟是出版于2006年，那个时候深度学习还不怎么繁荣（那时最流行的还是SVM）。而图学习更是最近几年才逐渐发展起来。缺失这两部分内容是时代的局限性而不是作者的局限性。至于聚类和主题模型其实PRML中也有介绍，只不过没有作为专题介绍。聚类是夹杂在EM算法中介绍的，主题模型则是隐变量模型的一种。没有系统性的对这两类常见算法做介绍也许是作者为了控制全书的篇幅思考之后取舍的结果？我也在想，如果强行加上这两部分内容的话放在哪里比较合适？结果是感觉放在哪儿都不太合适。LDA可能可以放到第十二章中介绍，至于除KNN（KNN在第九章有介绍）以外的聚类算法似乎就无处安放了。要放大概也只能放在第九章了。

第二，PRML还是有点难。虽然和《学习理论的本质》相比PRML算难度适中，但它和MLAPP一起几乎是剩余书籍里面最难读的书了。但PRML的难主要体现在它理论的坚实及视野的广博，MLAPP的难则主要体现在它的行文风格。通读全书还是能比较明显的看出Bishop在非常努力的启发读者，引导读者去思考和理解。而MLAPP中这种努力就很少。它的行文风格就是直接说明问题、然后出给公式和结论。所以说“难”大概不是PRML的一个缺点，而是它的一个特

点。

三部分内容说完之后再说点轻松的，主要是网上一些关于PRML的评价，我们来品评一番，就当茶余饭后的消遣了。还是那句话，说的不好的请大家见谅。
网友们对PRML是又推崇又敬畏，于是给了它一个机器学习“圣经”的封号。如果一定要把现有关于机器学习的书集合起来并横向比较，给其中一本颁布一个“圣经”的称号，那PRML绝对当之无愧。但如果站在现如今的视角来审视PRML的话更准确的称呼应该是“统计学习圣经”。因为随着这十几年来的快速发展，深度学习已经自立门户成为江湖中名声最大、信徒最多、影响力最深远的独立门派。传统定义上机器学习包含了深度学习，人工神经网络只是机器学习众多分枝中的一枝。但现实中机器学习这座庙似乎又早就装不下深度学习这尊大佛。所以很多时候现在的机器学习特指不包含深度学习的那部分内容了。但无论如何，我觉得PRML是受得起“机器学习圣经”这个称号的，如果有一天Bishop能出个第二版，我想它的第二版也是“圣经”这个称号的唯一继承人了。

第二，网上有句幽默的话叫“半部PRML治天下”。说这句话的肯定是个中国人，它的原话大概来自宋代宰相赵普的“半部伦语治天下”。这个典故的原型是：赵普辅佐赵匡胤建立了宋朝，可赵匡胤去世后赵普在一系列政治斗争中被罢了官。他的政敌为了抹黑他散布谣言说赵普这个人其实没什么文化，只读了《伦语》这一本书。后来赵光义想要启用赵普于是就问他：“我听说你只读过《伦语》，是真的么？”赵普机智的答道：“我确实没读过什么书，只读了一本《伦语》。但我辅佐太祖定天下只用了半部《伦语》，现在还剩半部正好可以辅陛下。”宋太宗对他的回答非常满意，于是便重新对他委以重任。从此就有了“半部伦语治天下”的美谈。当然，在现实中，如果你能真的把PRML读懂一半也很不错了。这对阅读别人的论文、理解别的方法和理论都非常有用。但是很明显，在当今这个时代仅凭半部PRML是很难“治机器学习的天下”的。毕竟前面我们说过，现在流行的是深度学习，而这本书受限于成书的时间，对现如今的人工智能行业而言是不够的。

第三是很多人认为作者通过这本书传递了一个“一切皆可贝叶斯”的理念，并把这本书誉为贝叶斯的扛鼎之作。我比较认同这样的观点，比较有代表性的就是书中的第三、四、五章。这三章在介绍完基本模型后总是会再介绍一遍这些模型的贝叶斯版本。但我个人也觉得这句话是不是能泛化成“一切皆可概率化”？贝叶斯在某种程度上等效于正则化，也在某种程度上等效于概率分布的非奇异化（这些概念在PRML中都会有解释）。所以是否贝叶斯有时候不是问题的关键，问题的关键在于

是否能优化模型的泛化能力。作者的确是在贝叶斯方法上着墨很多，而且他自己也说“本书特别强调贝叶斯观点”，但如果我们仔细观察会发现，无论是最基本的线性回归、线性分类、SVM，还是较为复杂的PCA、HMM、卡尔曼滤波，即使是人工神经网络，作者都会从概率的角度去把这些模型解释一遍。这是本书与很多书的一大区别。我觉得作者这样做大概是因为建模是需要理论支撑的，而概率统计是机器学习现有的、最好的理论基础。万物皆有源有本，如果只介绍那些模型本身，读者看到的那些模型可能就犹如精美的空中楼阁，忍不住去想为什么这样建模就可以了？这些模型是凭空设计出来的么？从概率的角度去解释我认为是对模型及其建模思想这二者之间的一种弥和。从“一切皆可概率化”进一步延伸出来的就是近似计算，因为概率计算往往伴随复杂的积分，大多数情况下解析解都不可求。所以本书花了大量的笔墨去介绍近似算法，EM算法、变分法、Laplace近似等等都是典型代表。所以我个人觉得，”一切皆可贝叶斯“这个评价可以进一步放宽或者扩大为”一切皆可概率化“。

最后再说回到Page-by-Page项目上。Page-by-Page的初衷是和大家一页页甚至一句句的把PRML读懂读透，我觉得这个项目对大家的帮助体现在两方面：第一，这个项目能帮你能更容易的读完PRML。因为PRML确实有些地方不太好读，一个人在没有外力的帮助下去读很容易失去信心甚至放弃。尤其是对于还在学校的高年级本科生、硕博研究生。我相信Page-by-Page项目能帮你扫除你在读这本书时会遇见的至少90%（或者更多）的困难。剩余的部分可能由于各人读书角度的不同，使得大家和我对于哪些地方是难点的理解不同。所以有些地方可能我觉得需要花力气去解释而你却觉得很简单、没必要，另一些地方你觉得应该深入进去而我却没有。如果有这种地方大家可以一起讨论，也可以到淡蓝小点来提问。我会尽最大努力去回复大家。也欢迎大家相互贡献知识、互相解答；第二是利用这个项目能加快你读完PRML的速度、节省时间。书中很多需要解释的地方我相信大家凭借自己的力量也能搞明白，但这个过程需要去查资料、思考和总结，这都需要时间和耐心。我觉得整个项目至少能帮大家节省一年或更多的时间。但是也请大家注意，Page-by-Page只是一个读书项目，这个项目中介绍的都是已有的知识。所以本项目能帮你学习但不能替你创新。不过，读完整个PRML大概能把你送到创新的门口。因为在机器学习领域有一个好的理论基础也是非常、非常重要的。如果创新是足球比赛中的临门一脚，那本项目大概相当于将球送到门前并让你学会如何射门。但无论如何临门一脚还要看诸位自己。

再说一点跟着page-by-page读完PRML大概需要哪些基础知识。最重要的包括：高等数学、概率与统计、线性代数。这三门课程达到一般大学70分的水平就够了，不需要很高。也不需要大家完全掌握这三门课程的全部内容，知道一些最重要的概

念就行。例如，知道什么是连续函数、什么是导数、什么是梯度、什么是微积分、什么是收敛和发散，什么是频率、概率，什么是随机变量、期望、方差，什么是矩阵、行列式、转置、可逆、特征值特征向量、什么是拉格朗日乘子法。注意，这里没有“等”。如果我列举的这些问题你都知道，或者你能现在去看书把它们都回忆起来，那你就完全具备了跟着Page-by-page去读PRML的基础。其它的更多知识，例如实变函数、泛函分析、凸优化等我们会在项目中做专门的介绍，几乎不用你提前自学。

整个项目参考了很多资料，原谅我无法在后面的内容中一一指处它们，我只能在这里统一列出一部分名字。还有大部分我参考过的网页或资料已经不可考了。他们包括：

- 《Solution Manual For Pattern Recognition and Machine Learning》, edited by ZHENGQI GAO, the State Key Lab. of ASIC and System School of Microelectronics Fudan University.
  这是由ZHENGQI GAO发布的另一份练习答案，我用的是2017年版的，我借鉴了很多里面的推导和解答。只不过这份资并不全。第八章缺失一部分，第十二章缺失一部分。第十三、十四章没有。这份资料的下载地址是：https://github.com/zhengqigao/PRML-Solution-Manual

- B站上的三位用户：
  shuhuai008的《机器学习-白板推导系列》，主页地址是：https://space.bilibili.com/97068901?spm_id_from=333.337.search-card.all.click
  大旗宛城的《凸优化》，主页地址是：https://space.bilibili.com/364872099?spm_id_from=333.337.search-card.all.click
  派派大星星的PRML解读系列，主页地址是：https://space.bilibili.com/6293151?spm_id_from=333.337.search-card.all.click

- 其它一些我参考较多的视频资料：
  徐亦达老师的课程，在B站、优酷、Youtube上面都有，B站地址是：https://space.bilibili.com/327617676?spm_id_from=333.337.search-card.all.click
  加州理工大学Yaser ABU-Mostafa教授的课程：
  https://www.youtube.com/watch?v=mbyG85GZ0PI&list=PLnIDYuXHkit4LcWjDe0EwlE57WiGlBs08

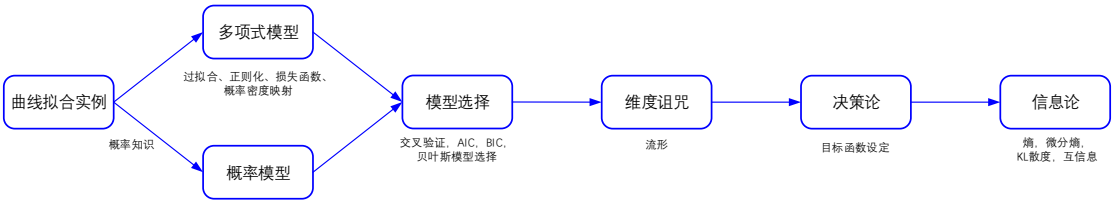
- 前面与PRML作比较的书籍全都被参考过，包括西瓜书，李航老师的第一版、ESL、统计学习的本质、MLAPP、《Learning from data》等。

- 无数的网页及论文，主要来自Wikipedia、Mathematics、Stackoverfow、Stackexchange等。

还有更多的参考资料无法列出，请大家见谅。如有问题请联系我，谢谢！我们马上就要开始PRML这项浩大而卓有意义的工程了，希望完成时还能见到你!

第一章　Review of Chapter 1

第一章是全文的引言，介绍了一些关于模式识别和机器学习的基本概念，这一章的内容量4.0分，难度3.5分，可读性4.0分。整体上，这章没那么难，但读起来也并不觉得多轻松。PRML的理论性从第一章就可见一斑。第一章的架构如下。

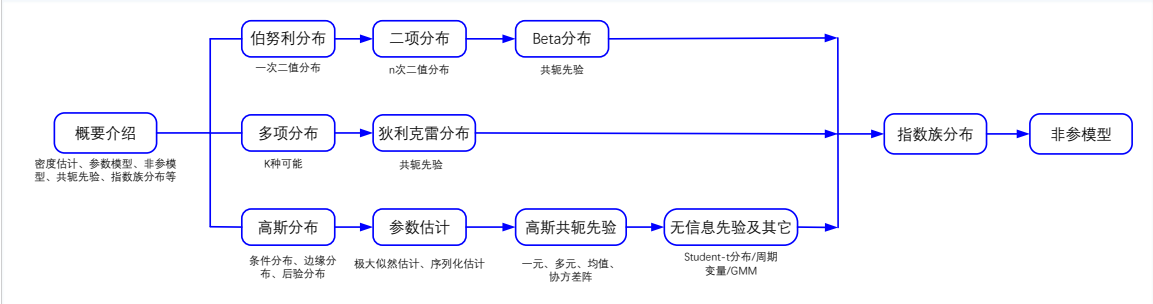


以曲线拟合的例子为切入，介绍了如何构造一个多项式模型去识别数据中隐藏的模式。在这个过程中引入了误差函数、模型训练、模型选择等概念。然后又基于概率理论为该问题建立了概率型模型，并引出过拟合、正则化、极大似然估计、有效参数等概念。模型选择是建立模型时一定会遇到的问题，应对该问题的方法也不少，如交叉验证、AIC、BIC、贝叶斯模型选择等。不过这些概念只在这里简单提到，具体介绍会在各章中进行。

维度诅咒是常见的一个问题，这里说明了什么是维度诅咒以及它为什么可怕，但没有详细说明如何解决。最常见的方法是特征选择或降维。本章介绍决策论其目的在于告诉我们如何设计一个好的、实用的优化目标函数。我们在设计损失函数时必须考虑到问题的现实意义。最后是对信息论简要的介绍。重点理解熵、微分熵、KL散度、互信息这几个概念。尤其注意区别连续型变量的微分熵和熵，它们在定义上有微小不同。总体而言第一章是相对简单的，几乎不需要补充介绍其它内容。比较困难的点包括：变量密度函数映射，有偏估计、无偏估计等。第一章中的很多概念会在后面章节中被反复提及和解释，所以第一次阅读时留有一些疑问也问题不大。

第二章　Review of Chapter 2

第二章主要补充介绍了概率分布的相关知识，这是非常基础但也非常重要的内容，大家不要轻视。第二章内容的结构如下图所示。

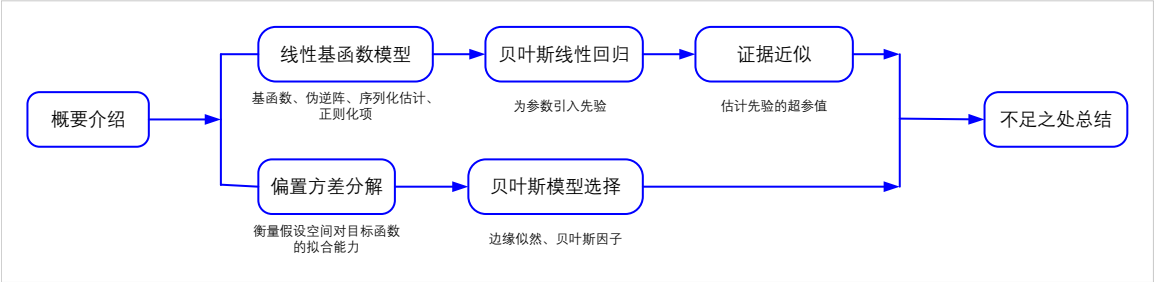


本章开头总起之后，首先介绍了最简单的分布——伯努利分布及其推广N重伯努利分布，也叫二项分布。所谓伯努利分布就是扔硬币，它只有两个结果。而二项分布就是将扔硬币重复N次。Beta分布是这两个分布的共轭先验。多项分布是有K种取值情况的离散分布，例如扔有六个面的骰子就属于一次多项分布。多项分布的共轭先验是狄利克雷分布。本章的重点在于2.3节高斯分布。在这一节，作者非常耐心细致的介绍了高斯分布、高斯条件分布、高斯边缘分布及高斯后验分布、高斯共轭先验。这些分布的推导结果在后面章节中用的非常多，大家一定要重视这些内容。有一些内容可以不作为本章内容的重点，但了解这些内容也是多多益善，如Student-t分布、周期性随机变量等。GMM是高斯分布的一个变种，它也非常常用，第二章对它做了一定的介绍。但关于GMM的具体的参数推导会在第9章EM算法那里介绍。但有一个重点是无信息先验。无信息先验部分有点抽象，需要补充介绍较多的内容。前面介绍的这些分布以及更多的常见常用分布，如泊松分布、指数分布（注意区分指

数分布和指数簇分布，指数分布是一种具体的分布，它也属于指数簇分布）等其实都属于同一簇分布：指数簇分布。指数簇分布是对所有前面介绍的分布的总结，它是一类具有某些相同特性分布的抽象概括和总结。指数簇分布几乎囊括了所有我们平时常见的分布，具有很多良好的性质。但请注意，也有两个明显的例外：GMM和一致分布不属于指数簇分布。一般情况下混合分布都不属于指数簇分布。最后，本章还介绍了什么是非参模型。所谓非参模型是指无法用确定的参数来控制和描述的模型。非参模型的计算通常是基于全部或部分样本数据完成的，而不是某个特定的公式。

第三章　Review of Chapter 3

第三章介绍了线性回归模型的相关内容。线性回归模型几乎最简单的模型，它的研究意义甚至大于其实用意义。本章在全书中也算是内容相对较少、相对简单的问题。但作者还是介绍了一些比较深刻的概念，如偏置方差分解、贝叶斯因子等等。第三章内容结构如下所示。



作者首先对第三章内容做了简单的概要介绍。在3.1节中，作者介绍了线性基函数模型。它是指对变量$\boldsymbol{x}$做关于基函数的变形，再基于基函数建立参数的线性模型。这就是线性基函数，也是我们本章要重点研究的线性回归模型。3.1节会介绍好几种不同的基函数。在估计参数的值时，我们会遇到一个非常重要的概念——摩尔-彭若斯伪逆阵。我们知道，极大似然估计其实就等同于最小平方误差函数的优化问题，作者在3.1.2对最小平方误差优化给出了几何解释。作者不仅介绍了极大似然估计，还介绍了序列化估计方法。关于正则化项，我们要注意次数为1和次数为2的两种特殊情况。它们次数比较低，我们用的比较多。而且它们一个是凸的、一个不是凸的。如果为模型的参数引入先验就得到了贝叶斯线性模型，如果先验参数也未知的，也就是模型的超参也需要基于训练集估计，那就要用到3.5节中介绍的证据近似方法。

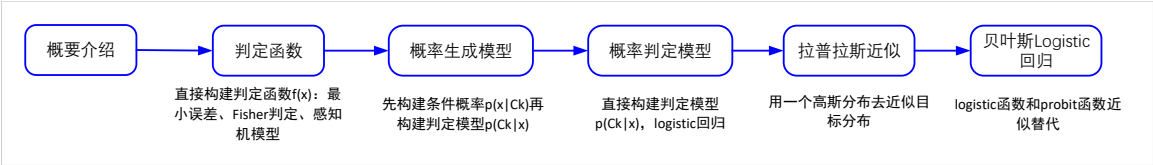
模型的偏置方差分解是量化的衡量整个假设空间的对目标函数的拟合能力的。请大家注意，它衡量的是整个假设空间而不是从假设空间中学习到的某一个特定的模型。这部分内容我们会引用《Learning from data》这本书2.3节对这个问题的解释，那里解释的很清楚。

贝叶斯模型选择是跟交叉估计不同的另一类模型选择方法，其中的重点是贝叶斯因子。

最后作者对线性回归模型的不足之处作了简单总结。

第四章　Review of Chapter 4

三、四两章讲的都是线性模型，但第三章是线性回归，第四章是线性分类。三、四两章从内容、篇幅上来看都差不多，都属于相对难度不太大的章节。但第三章后面部分更侧重于理解，第四章后面部分更侧重于推导。第四章内容结构如下所示。



线性分类方法整体上可分三类（在第一章就概括性介绍过）：直接构建判定函数、构建生成式模型、构建判定模型，分别对应本章的第4.1、4.2、4.3小结。4.1节直接建立判定函数$y(x)$。对于给定的x，模型直接给出x所属的类别。若$y(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x}+w_0$，就是最基本的线性判定函数。4.1节一共介绍了三类判定函数：最小平方法判定、Fisher判定、感知机算法。其中感知机算法在线性模型的基础上引入激活函数。4.2节先建立条件概率模型$p(x|Ck)$，再根据贝叶斯公式得到判定模型$p(Ck|x)$。这类方法得到的信息最多，但也难度最大、计算量最大。4.3节直接建立判定模型$p(Ck|x)$，这类方法中最重要、最典型的代表是logistic回归模型。logistic函数虽然简单，但求它参数的估计值没有解析解，所以要用近似方法。

4.4节介绍了Laplace近似，这是一种比较简单、常用的近似方法。它用一个高斯分布去近似目标分布。但它要求目标分布有良好的单峰形态，否则近似效果不佳。这一方法在本书后面几章中会被经常用到。4.5节为logistic模型参数引入先验得到logistic后验，但由于logistic模型没有解析解，所以我们只能用近似算法求解。最

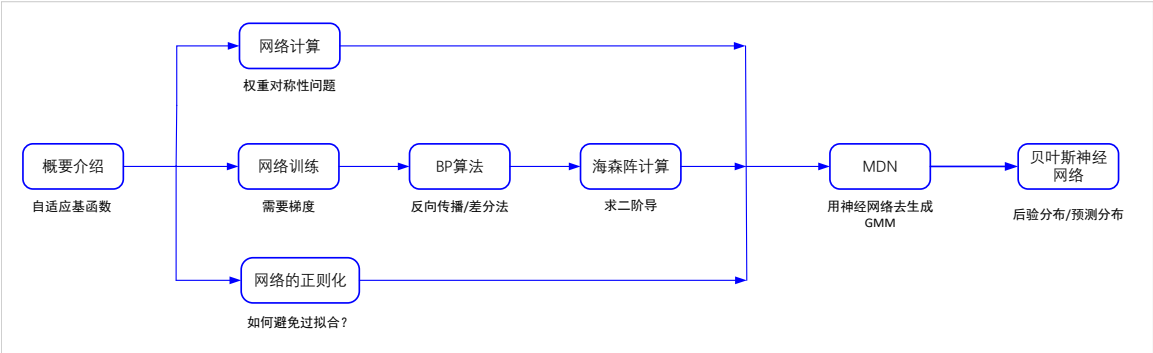
后一节的近似算法略显麻烦，需要大家仔细一点。

第五章　Review of Chapter 5

当下正是深度学习大行其道的时候，但从这本书发行的2006年到现在已经过去十多年了。这十多年正是神经网络飞速发展、重新崛起的十多年。应该说，现今业界对神经网络的理解已经和这本书发行时人们对神经网络的理解有了天翻地覆的区别。因此，当我们站在如今的时间坐标顺着作者的视角来审视当年人们对神经网络的理解时，我们会发现那是一种完全不同的角度。现如今的神经网络人们关注的是网络的结构、网络的效果，或者网络的分布式训练、布署等等。这些问题与应用密切相关，而且它有着如今这个时代数据量爆炸、计算能力指数级增长、互联网应用无处不在的时代烙印。而本书的第五章侧重介绍的是如何利用梯度训练模型的参数、如何高效的求雅各比矩阵、海森阵、如何为网络中的参数引入先验、超先验等，而这些都是统计模型传统的标准问题。神经网络在诞生之初只是众多模型识别方法中的一种，它跟其它统计模型是排排坐的兄弟或者说是同一座庙宇下排排摆着的罗汉僧。从这本书的视角和内容来看，虽然当年已经到了属于神经网络的时代前夜，但它仍没有越过突破这种身份限制的临界点，让深度学习成为跟机器学习平起平坐的江湖新派。人们还是习惯用传统的思维去研究它、看待它。

不过，这并不妨碍我们去阅读本书的第五章。现代意义上的深度学习虽然带有强烈的、天生的创新性，但传统研究者却有着坚实的理论基础。从概率视角、统计视角去重新认识一遍神经网络不仅能帮我们补充很多关于神经网络的基础知识，还能让我们更清晰的看到这两个时代神经网络的巨大差别，丰富我们的视野，让我们更充分的体会这十多年来深度学习野蛮生长的路径。

第五章的内容结构如下图所示：



神经网络并不是凭空出现的，它完全可以和前面介绍基函数模型关联起来。无论是分类问题还是回归问题，我们在第三、四章中都是提前指定了基函数。包括用多少个基函数、用什么样的基函数、每个基函数的参数是多少，我们只是利用训练数据学习了连接基函数的参数$\boldsymbol{w}$。从基函数的角度来看，神经网络与此十分类似，只不过各个基函数的关系比之前略显复杂。通过调整与神经元相连的边w_i的值趋向于0，可以视作是改变了模型基函数的数量。如果有需要，我们也可以将基函数中的参数视作待定参数，利用训练数据进行训练（虽然很多时候我们并没有这样做）。

经过简单的内容概要之后，5.1节以前向神经网络为例介绍了神经网络是如何计算的。还顺带讨论了神经网络的参数对称性。

网络的计算是比较简单直观的，它的重点在于训练。5.2节按照一般的思路——极大似然估计法分析了网络参数的求解。当然，这里最重要的一点是说明我们需要利用误差函数关于$\boldsymbol{w}$的梯度来更新$\boldsymbol{w}$值。至于如何求得误差函数关于每个w_i的导数则要到5.3节才会介绍。另外这一节还介绍了误差函数的二次近似。也就是用一个$\boldsymbol{w}$的二次函数去近似E($\boldsymbol{w}$)。在这种情况下，误差函数的等高线也会成为一个被旋转了之后的椭圆（跟二维高斯分布类似）。

5.3节是非常重要的一节，它具体回答了网络训练的最重要问题：如何求得梯度。当然，使用的方法是大家应该都知道的反向传播算法。一个小细节请大家注意：BP算法严格来说只是用来求梯度的方法，并不直接等同于参数训练方法。因为拿到梯度之后如何更新参数还可以有不同的做法。相当于参数训练可以分两步：第一步求得梯度，第二步更新参数。BP算法只是帮我们完成了第一步。BP算法其实并不难，只不过需要大家多一点耐心和细致。简单来说，我们利用网络从输出元向输入元把梯度一层层传播回去就叫BP算法。除了利用BP算法求梯度外，作者还介绍了用

差分法求梯度。相比之下，差分法比BP算法的复杂度更高，但它的逻辑更简单。它不需要做反向计算，纯粹的前向计算即可。因此差分法可用来做训练软件编写正确性的校验。而且，基于BP算法我们能高效的求雅各比矩阵。

5.4节用很大的篇幅说明了如何求神经网络的海森阵，这是一般的深度学习不会刻意介绍的。海森阵是一个非常常用的量，例如，它可以用来做二阶近似、可以实现训练数据微小变化时的快速重训练等。本书在这里讨论了有关海森阵的几个问题：1.海森阵对角化近似后的求法；2.海森阵近似为向量外积后的求法；3. 基于向量外积近似的逆海森阵求法；4.用差分法求海森阵的每个元素并近而求海森阵；5. 海森阵的快速精确求法，是基于梯度的反向算法；6. 如何快速求$\boldsymbol{v}^T\boldsymbol{H}$，即一个行向量和海森阵的乘积。我们会介绍一种快速、精确计算法。

如果说5.4节的内容已经跟我们一般印象中深度学习里讨论的话题相去甚远的话，那5.5节的内容可能更是许多深度学习研究者从未曾注意过的问题。5.5节讨论了网络的正则化，按照一般化思路，最先想到的当然是向误差函数中引入正则化项。但以权重衰减为代表的正则化项会带来一致性问题。除了提前停止训练外，作者还介绍了一些技巧性比较强的方法，如切线传播、引入数据变化的正则化项等。过拟合之所以被现代化的深度学习讨论较少是因为现代化的网络规模往往都特别大，远超当时人们的想象，所以现在更常见的问题是有效数据量不够导致的欠拟合而非过拟合。5.4节介绍的部分内容现在即不那么热门，推导起来还比较繁锁。请大家根据自己的需要和兴趣阅读。

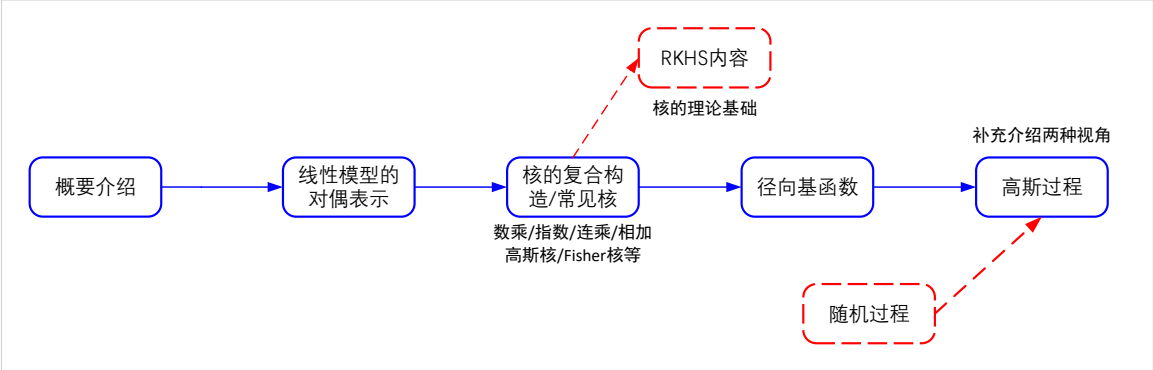
MDN是一种仍然具有生命力的算法，它是利用一个神经网络学习出一个混合模型。注意，它得到的是混合模型的参数而非x对应的标签值。

最后一节贝叶斯神经网络，重点在讲如何求网络中参数的后验以及预测分布。当然，由于网络模型的非线性、非凸性，我们主要采用近似计算。用到的方法包括拉普拉斯近似以及logistic、高斯卷积的近似求解。注意，此“贝叶斯神经网络”非现代意义上的贝叶斯神经网络。

第六章　Review of Chapter 6

第六章介绍了核方法相关内容，从这一章开始我们要进入本书最精华、最高能的部分了，并将一直持续到第十三章。虽然前面五章作者写作的也十分精彩，尤其是第二章，论述之细腻、内容之详尽，都让人十分喜爱。但受限于前五章的题材——第一章导论、第二章概率知识补充、第三四两章最简单的线性模型、第五章神经网络但介绍的内容已和如今业界的理解相去甚远，因此从本章开始PRML将为我们输出更具体、更实用、更具理论深度的内容。

第六章篇幅并不长，是全书第二短的一章（最短的是第十四章），但想要完全理解本章的内容却并不容易。核方法是一类比较抽象的方法，理解它需要一定的数学理论基础，因此page-by-page 项目会在适当的地方对某些问题展开，做轻量化的系统性介绍。我们来看本章的内容结构。



经过简单的概要介绍后，首先介绍了线性模型的对偶表示。对偶表示是对线性模型作变形，将其变成核函数的形式，以引出本章要讨论的内容。接着介绍了核函数的复合构造方法，也就是基于已知的或简单的核函数利用数乘、指数运算、连乘等操作生成更多更复杂的核函数。作者还介绍了一些常见的核函数，如高斯核、Fisher核

等。按本书的设计，接下来就是6.3径向基函数了，但我觉得在进入后面内容之前我们要回答一个非常关键的问题：如何判定一个函数是否是合法的核函数？书上对这个问题介绍的非常简单，只是一两句话一带而过，我觉得这是不够的。因此，我们会在这里对RKHS的相关内容展开介绍。这部分内容相对抽象，但并不需要大家具备很多的先序知识，所需的知识我们都会现场补充。RKHS是核方法的理论基础，我觉得不理解RKHS的话不能算真正理解了核方法。RKHS的补充内容最终将汇集成Mercer条件（核函数判定条件）。

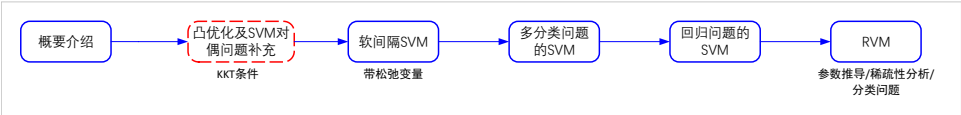
径向基函数是一种非常常见的核函数，它本身并不是很复杂。但作者在介绍它时提到了“格林函数”。针对格林函数，我们也会补充一部分知识。

最后一节是高斯过程。高斯过程是随机过程的一种，因此在深入高斯过程之前我们会先补充介绍随机过程的定义。另外，关于高斯过程我们会借助《Gaussian process for machine learning》这本书。我们先从这本书介绍的两种视角：权重视角和函数空间视角来理解高斯过程，然后再来看书上的内容。当我们能从这两种视角理解高斯过程时我们就能轻松理解书上的内容。

第七章　Review of Chapter 7

第七章主要介绍了两个模型：SVM和RVM，其中SVM是重点。SVM在深度学习再次崛起前曾非常流行，也许现在大家感受不到，不过当年它的火热程度可以用铺天盖地来形容。甚至在十年前，它的余热仍然还在。SVM最大的特点是它只利用少量的支持向量参与计算。RVM和前面介绍过的ARD（自动相关性确定）密切相关。

本书对于SVM的讨论主要集中于应用面，一个不足之处在于缺少了SVM最重要的理论基础——对偶问题转换。为了更好的理解SVM，我们会在本章较为靠前的位置对凸优化、对偶问题的相关内容作补充。本章的内容结构如下图所示。



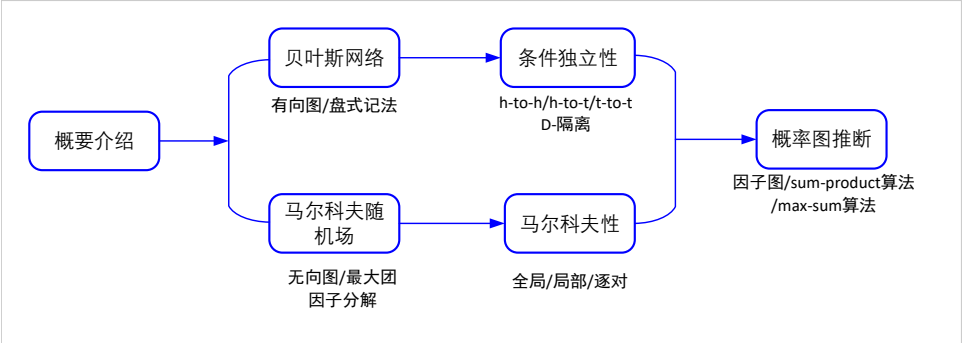
在概要介绍里说明了本章和第六章的区别。第六章介绍的是核方法，而本章讨论的是利用核方法的两个特例，它们最大的区别重点在于“稀疏”二字。所谓稀疏是指只有少量的、部分的数据参与计算，这样做最大的好处是能节省计算量。对偶问题转换是理解SVM最重要的理论基础，它也是凸优化问题中的最核心的内容之一。我们会补充介绍关于凸集、凸函数、凸优化、保凸运算法等相关内容。在对偶问题转换中我们将重点讨论KKT条件和Slater条件。理解了这些内容几乎可以说就理解了本章内容的一多半了。然后会介绍一些SVM在应用层面的拓展，例如为模型带上松弛变量、如何用SVM解多分类问题、如何用SVM解回归问题等。

RVM也是一种稀疏核机，而且它的稀疏性可能比SVM更强。本书主要讨

论RVM的三个问题：1.如何实现RVM的参数推导（这里我们会做较多推导）；2.分析RVM的稀疏性；3.将RVM用于分类问题。最后会对RVM和SVM做一个简单的对比、小结。

第八章　Review of Chapter 8

从第八章开始，我们要进入PRML的下半场了。第八章集中介绍了一些概率图模型的基础概念及算法，这些内容不针对任何特定模型，是普遍适用于全体概率图模型的。

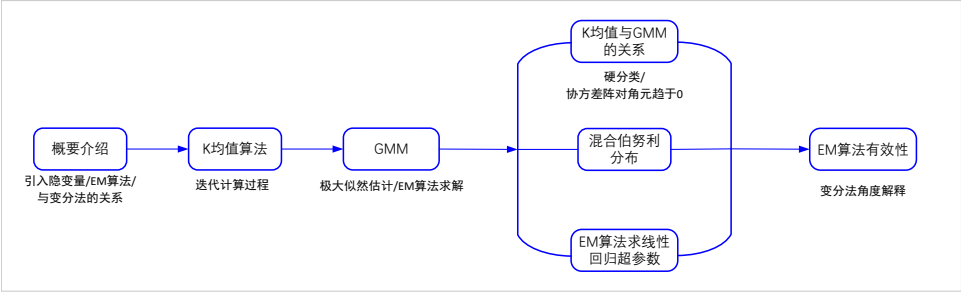


在概要介绍阶段，作者说明了PGM的分类及为什么要建立PGM。紧接着，8.1节首先讨论了有向图即贝叶斯网络，这其中最重要的是关于有向图的条件独立性判断。任意贝叶斯网络（无环）都由三种基本情况构成：head-to-head、head-to-tail、tail-to-tail，判断图中任意两个结点是否条件独立主要就是看这两个结点的连接关系符从哪种情况并且在当前观察情况下它们的相关性是被阻断了还是连通了。无向图又称为马尔科夫随机场，无向图中结点间的条件独立性判断依据是马尔科夫独立性，也分成三种：局部马尔科夫独立性、全局马尔科夫独立性、逐对马尔科夫独立性，实际上这三种独立性是等价的。最后一节介绍了对PGM的概率计算，也叫推断。关于推断，书上重点讲了三个问题：因子图、sum-product算法、max-sum算法。在做概率推断时，无论有向图还是无向图，首先都会转换成一种特殊的无向图——因子图。基于因子图，我们重点讨论两个推断问题：1.求任意某个结点或某些结点子集的边缘概率，这利用的是sum-product方法；2.求使得某个结点或结点子集概率值最大的变量取值，这利用的是max-sum算法。

本章的内容和概念相对较多较杂，而且为了便于理解我们会补充介绍一些实例（本章不会补充介绍更多概念，只会列举一些实际例子）。但相对前几章，如第六章核方法、第七章稀疏核机等，本章的内容不算太抽象，都是比较具体、易于理解的概念，只不过需要大家更有耐心。

第九章　Review of Chapter 9

第九章介绍了一个非常重要、非常常用、非常有名但同时又较好理解的算法——EM算法。本章内容既非常实用又有一定的深度，但同时又比较好理解且内容量不大，所以第九章可以说是PRML中最令人喜爱的章节之一。这一章的内容结构如下所示



首先作者说明了为什么要在模型中引入隐变量，表面上看引入隐变量使模型更复杂了（因为多了一个变量且这个变量的取值还是不知道的），但实际上这样做反而是为了帮助我们更容易的解决问题。而EM算法就是专门用于解决含隐变量模型的参数学习问题的。EM算法其实只是第十章马上要介绍的变分法的一种特例。

然后作者首先介绍最简单的隐变量模型——K均值算法，并推导了如何学习模型中的参数。然后就介绍了GMM算法。GMM算法是非常实用的一个算法，但是注意它不属于指数族分布。实际上一般的混合算法都不属于指数族分布。理论上，如果不考虑计算时间和存储空间的限制，用含足够多分量的GMM可以以任意精度拟合任意分布。当然，这里最重要的还是如何利用EM算法学习GMM中的参数。之后，作者介绍了三个拓展问题：K均值与GMM的关系，实际上K均值是一种特殊的GMM算法，它的特殊性表现在K均值算法是硬分类并且它相当于对各分量协方差阵取值做了限制的GMM；混合伯努利分布，这只是给我们多举一个例子，加深理解；以及如何利用EM算法求线性回归模型中的超参数α,β。前面我们都是在介绍什么是EM算法以及

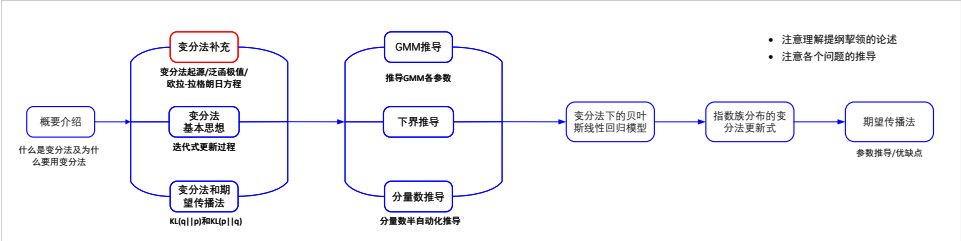
它的计算原理，最后一小节作者从更一般的角度说明为什么EM算法作为一种近似算法是有效的。虽然这里的推导并不涉及到变分法，但实际上等我们介绍完第十章变分法后就能体会到，这里的有效性说明其实是在变分法的框架下进行的（毕竟EM算法只是变分法的一个特例）。

EM算法总结起来有三个特点：1.它是近似计算方法，但能达到较高的精度和可靠性；2.它是迭代式算法；3.它适用于含隐变量的模型。

第十章　Review of Chapter 10

第十章介绍了机器学习中一类非常有名的方法——变分法。变分法源自于物理、工程控制专业，经过演化后应用到了机器学习中。相比于前面介绍的其它一些方法，变分法自成体系、框架完整，第九章中的EM算法就是变分法的具体实现。其实计算机专业（包括电子、信息、生物工程等专业）无论是本科生还是研究生，很少开设变分法这门课。在部分学校的部分专业（如数学专业、物理专业、控制专业等），变分法可能会和泛函或实变一起开设。因此，大家如果之前不了解变分法也并不奇怪。这一章内容相对较多，但相比核方法等章节，并不十分抽象，应该说理解起来难度不大，只是需要一些耐心。阅读本章时我建议大家注意两个方面：

1. 要注意理解那些提纲挈领的论述句
2. 要注意对具体问题的推导



在概要介绍中，作者简略说明了什么是变分法以及为什么会有变分法。紧接着，我们补充介绍有关变分法的背景知识。包括变分法的起源、为什么说变分是泛函的微分、欧拉-拉格朗日函数等，然后作者从一般角度介绍了变分法的计算过程。变分法的关键其实是近似，它根据KL散度KL(q||p)，寻找近似于p的分布q。我们在第一章介绍过，KL散度是不对称的，因此将p和q反过来会得到完全不同的结果。也就是基于KL(p||q)会得到期望传播法。

介绍完这些内容之后，作者介绍了好几个变分法应用的实例。包括用变分法推导GMM、用变分法推导贝叶斯线性回归、变分法推导Logistic回归等。这其中最值得

大家留下深刻印象的我觉得是如何用变分法实现GMM分量数的半自动推导。作者还介绍了如何用变分法对一般形式的指数族分布进行推导。注意，我们前面说过，一般的混合分布都不属于指数族分布，我们这里的意思是完整数据{X,Z}属于指数族分布。

最后，作者介绍了期望传播法。期望传播法和变分法思路类似，只不过在具体实现上略有差别（一个是KL(q||p)一个是KL(p||q)）。

还是那句话，本章的推导工作较多、较细致，请大家尽量保持耐心。

第十一章　Review of Chapter 11

第十一章介绍了非确定性近似方法的核心问题：采样。从某个已知分布中获取独立同分布的样本是一件看似理所当然但实际上却没有那么容易的事。本章将集中介绍若干种采样方法，其中最重要的方法是MCMC，而最难以理解的方法可能是HMC。

从整体上看，这章的内容可分为三大部分。第一部分介绍了一些常规采样方法，例如标准采样法、拒绝采样法、自适应采样法、重要性采样法等等。这些方法都比较简单，也很好理解。但这些方法适用性都不太强，要么是对被采样分布的密度函数要求过高（例如要求可逆、可导等），要么是效率太低采样太慢（例如拒绝率过高或找不到好的提议分布）。这些方法我们会逐一介绍。第二部分介绍了MCMC，它也是本章的重点。MCMC指的是一类方法，或者说一类方法的思想。它具体代表包括Metropolis算法、Metropolis-Hasting算法、Gibbs算法等。大家在读本章的时候，应以这部分内容为主，重点理解MCMC的思想及Metroplis算法的采样过程。这部分内容我们会做一定的补充，我认为难度不是很大。第三部分介绍了一种较新的采样方法——混合蒙特卡洛算法或者也称为哈密顿蒙特卡洛算法。说它较新是相比于MCMC而言，实际上它也被提出了几十年了。这部分内容理解起来难度比MCMC稍大，因为HMC的理论基础是哈密顿力学。非力学或物理学专业的同学可能都只学过牛顿力学，不太了解哈密顿力学。虽然作者说理解HMC可以不了解哈密顿力学，但是我觉得先补充介绍一部分HMC的背景知识总是好的。HMC其实也是一种MCMC，或者说也是一种Metropolis算法，只不过它有独特的接受率。HMC相比其它方法的特点是它更快、更高效，因为它的游走过程利用了梯度信息

（一般MCMC中的游走是随机的）。对于HMC，如果大家不能理解它的原理和思想的话，也可以先只记住它的采样过程。它的采样过程其实是相对简单的，跟一般的Metroplolis算法没什么区别。最后还有一小节介绍了一下配分函数，因为我们经常用配分函数定义概率分布。这部分内容较少。总体而言，本章的重点是MCMC，难点是HMC。

第十二章　Review of Chapter 12

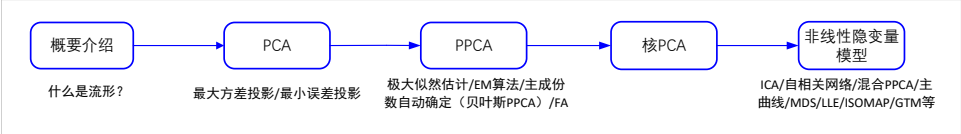
第十二章介绍了连续型隐变量模型。根据隐变量、观察变量取值类型的不同，一共可以分为四类。我们在第九章介绍的EM算法其实是离散型隐变量、连续型观察变量模型。本章介绍的主要是连续型隐变量、连续型观察变量模型。当然，在本章的最后会简单介绍一下连续型隐变量、离散型观察变量模型，不过只是简单的一代而过。离散型隐变量、离散型观察变量模型会在第十三章介绍，典型代表就是HMM。本章虽然名为“连续型隐变量模型”，但其实我个人觉得基本上也可以算作是流形学习基础了。

我们在概要阶段会做一点补充介绍，主要是介绍什么是流形。它的定义略显复杂，涉及到拓扑。本章的重点是第二节和第三节，分别介绍了标准PCA和概率版PCA。PCA是一个非常有名的模型，在数据降维、特征提取、可视化等方面都有广泛的应用。简单来说就是将高维的$\boldsymbol{x}$转变成低维的$\boldsymbol{z}$，转变后的坐标基向量是数据集协方差阵的特征向量。对于PCA，书上给出两种角度：最大方差投影和最小误差投影。但其实这两种角度很相似。

PPCA是概率版的PCA，可以通过极大似然估计求模型中的参数。或者，为了提高计算效率、减少计算量，还可以借助EM算法进行求解。无论是PCA还是PPCA，一个很重要的问题就是要确定主成份向量的数量（也就是降维后的维度数）。这类似于确定混合高斯分布中的分量数。我们可以借助贝叶斯方法来自动推导主成份数。将$\boldsymbol{x}$服从的高斯布分布的协方差阵从各向同性的矩阵放松为对角阵就得到FA模型。

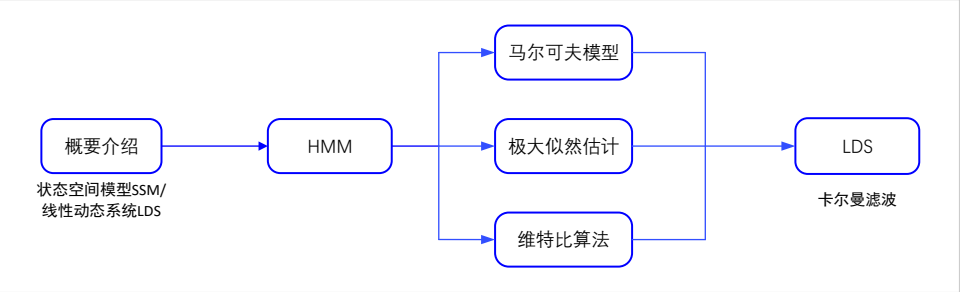
核PCA其实就是将原先的$\boldsymbol{x}_n$转变成$\boldsymbol{\phi}(\boldsymbol{x}_n)$，其它的几乎都不变。

最后一节，作者罗列了大量的非线性隐变量模型（PCA、PPCA、核PCA，隐变量和观察变量都可以看作是线性关系）。我们对作者提到的方法做简要介绍，其中比较著名的有ICA、自相关网络、MDS、LLE、ISOMAP等。MDS、LLE、ISOMAP是流形学习中非常经典、非常具有代表性的算法。



第十三章　Review of Chapter 13

第十三章介绍的是序列化数据模型。这一章跟前面介绍的内容有一个很大的不同：前面所介绍的模型，它的数据基本都是独立同分布的；而这一章要介绍的序列化模型，它的数据之间是相关的。当然我们这一章讨论的是最简单的关联关系：即序列化关联关系，也就是下一时刻的数据只和上一时刻的数据有关。

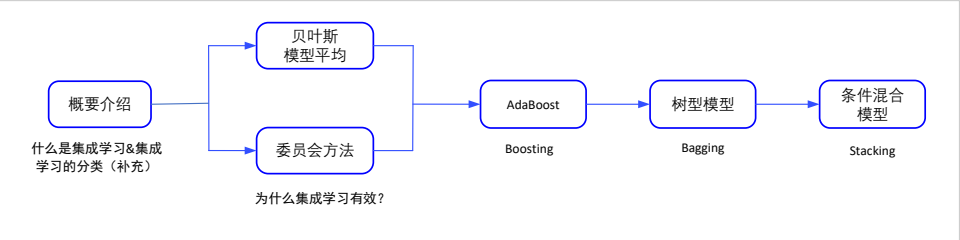


整个第十三章可分为两部分。第一部分主要介绍了隐马尔可夫模型，它是状态空间模型的典型代表。隐马尔可夫模型含有离散型隐变量，它的观察变量可以是离散型的也可以是连续型的。但是HMM中隐变量之间存在序列关系，观察变量也受隐变量的直接影响。在PRML中，我们重点讨论有关HMM的三个问题：HMM的由来、HMM的极大似然估计以及用于找到最优隐变量取值的维特比算法。第二部分介绍的是线性动态系统，简称LDS。线性动态系统顾名思义，变量之间不仅是序列化关系而且这种序列化关系是线性关系。如果我们假定每个变量都服从高斯分布，那么就有了卡尔曼滤波。卡尔曼滤波是LDS中的典型代表，它也是一个非常实用的模型，它被广泛的用于位置估计。例如，自动驾驶中系统实时感知自己的位置就应用了卡

尔曼滤波；卫星或飞行器也利用了卡尔曼滤波给自己定位。卡尔曼滤波的数学基础其实就是我们一直在用的高斯分布的线性变换，我们会给出它的具体推导过程。

第十四章　Review of Chapter 13

第十四章大概是PRML全书最简单的一章，它甚至比第一章Introduction还要简单。“Combining Models”可翻译为融合模型或模型组合，但实际上这一章介绍的就是集成学习“ensemble learning”。集成学习是机器学习中非常有名的一类，它在多个机器学习竞赛中都取得了很好的成绩。集成学习是指用多个模型同时进行预测，并将多个结果结合起来作为最终的预测结果。本书这一章介绍的主要是基于统计学习的集成学习，随着深度学习的崛起，现在也有很多专门针对深度学习的集成学习方法。本章的内容结构如下所示。



在概要介绍阶段，作者简单说明了什么是集成学习。集成学习（基于统计学习方法）其实可分为三大类分别是Boosting、Bagging、Stacking。关于这一点书中没有明确说明，我们会作一点补充介绍。接下来，作者首先说明了集成学习和贝叶斯模型平均间的差别；然后说明了为什么委员会方法是有效的，委员会方法其实就是集成学习，这一小节相当于从理论上说明了为什么集成学习能进一步提高预测的准确性；接下来还有三小节，其实这三小节就分别介绍了Boosting、Bagging、Stacking方

法。Boosting的典型代表是AdaBoost，Bagging的典型代表是树型模型例如随机森林。不过本书并没有介绍随机森林，而是介绍了随机森林的基础——如何构造树型模型。随机森林相当于由多棵树型模型构成的模型组合。Stacking就是将多个模型按比例组合起来，书上是以条件混合模型为例进行说明的。