"十一五"国家重点图书出版规划项目

10000个科学难题

10000 Selected Problems in Sciences

信息科学卷

Information Sciences

"10000个科学难题"信息科学编委会

科学出版社
www.sciencep.com

目　录

科普篇

机器学习之模型选择

Model Selection for Machine Learning

机器学习致力于建立计算模型 q 以发现并描述已知数据集 X 中的规律，并用 q 进行预测推断和问题求解。q 是一个计算程序或数学模型。X 中数据量是有限的，可以被一个复杂度 k 足够的 q_k 及任何一个 $q \supseteq q_k$（包含 q_k 为子集）来完全描述。不过，X 中数据通常受到噪声干扰，即使 q_k 能完全表述 X，也分不清所描述的有多少是规律，多少是噪声，且 k 越大，描述噪声的机会也越大。相反，一个复杂度 k 太小的 q_k，不能充分表述 X 中规律，且 k 小则相应的描述误差（称为经验误差或风险）r_f 也越大。考虑模型簇 $q_1 \subset q_2 \subset \cdots \subset q_k \subset \cdots$，如图 1 所示，曲线 r_f 随 k 增加逐步下降至零，因此，难以由 r_f 确定合适的 k^*。我们更感兴趣的是，将 q_k 用于 X 以外但仍反映 X 中规律的未来数据，相应的描述误差称泛化误差或风险 r_g，它随 k 变化有个最小点 k^*。以图 2 为例，经验误差 r_f 在 $k=1$ 处较大，而在 $k \geqslant 2$ 各处都变得很小，泛化误差 r_g 在 $k=1$ 处较大，在 $k=2$ 时最小，而 $k>2$ 又增大。模型选择之经典概念就是在模型簇 $\{q_k\}$ 中选择合适的 k^* 以致相应模型 q_{k^*} 之泛化风险最小。实际上，并没有未来数据可用，只能根据经验风险 r_f 和 q_k 的结构，在给定数据集 X 上去预测它在未来数据上的表现 r_g，这正是模型选择的困难所在。

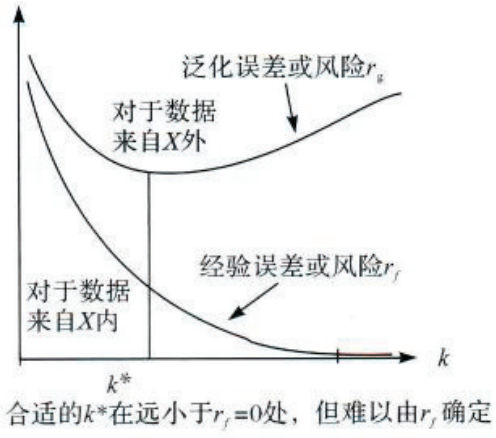


图 1　经验风险与泛化风险

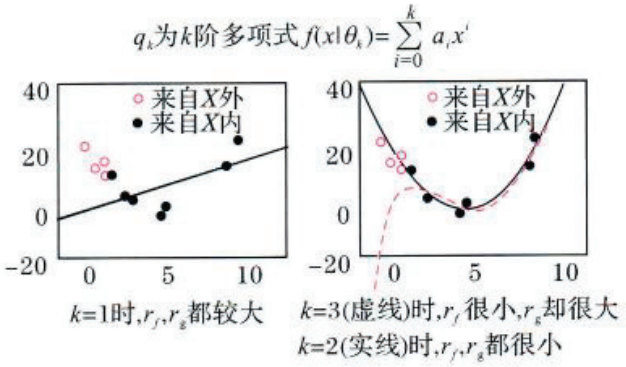
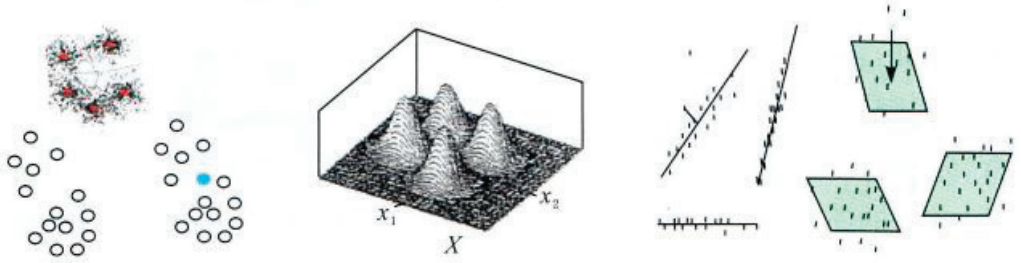


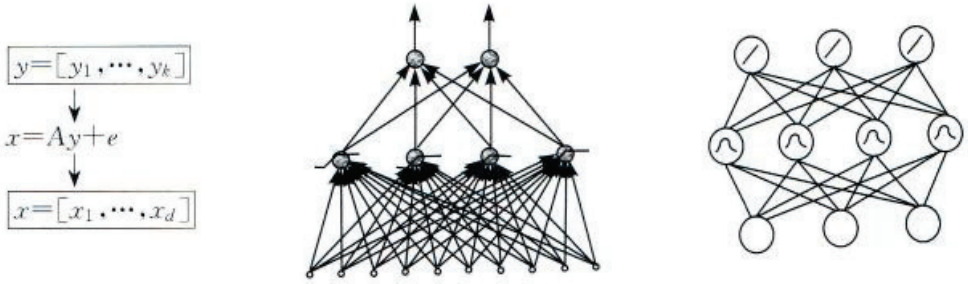
图 2　多项式曲线拟合

在机器学习中，q_k 通常由 k 个简单单元经过简单组合形成，如图 3 所示。最常见是图中第一个例子，每个单元为一红点，表示一个类，称为聚类分析，模型选择的任务是回答有几个类。同样，此问题难以由曲线 r_f 回答，看似是一、

二类的数据集，却可以在把每个数据点都看成是一个类时出现 $r_f=0$，而且，究竟是 $k=1$ 或 $k=2$，并非仅看似就可解决，一个数据点的去留就能影响判断(如图 3 中实心点所示)。考虑到数据的随机性，每个单元进一步可以是一个概率分布，也可以是线、面或图 3 中的其他结构。更一般地，q_k 可以表示成一个参数概率分布 $q(X|\theta_k)$。模型选择通常以两步法实现。首先，对每个 q_k，基于 X 求解 θ_k，称为参数学习，并得到图 1 中的 r_f 曲线。然后，按某准则近似估计 r_g，并选择合适的 k^*。这个两步法，枚举 k 并对应于每个 k 值需学习 θ_k，计算量巨大。不仅如此，当 k 值越变越大时，θ_k 中未知参数的数目将骤增，θ_k 和 r_f 曲线的估计都将变坏，并造成 k^* 估计变坏。这两个缺点都让模型选择变得更加困难。



q_k 可以简单的是几个点、线、面、高斯分布



q_k 也可以是一个 k 维线性系统或是 k 个隐单元构成的非线性网络

图 3　简单单元经过简单组合形成的结构

下面以两个发展阶段，回顾数十年来对这些问题的主要努力：

第一阶段从 20 世纪 60 年代中到 90 年代初，以两步法实现为主，探求各种理论下的选择准则，可归纳为两个方向。方向一是寻求近似估计泛化风险 r_g，兵分三路。一是基于数据集 X 重复取样的经验估计，要点是每次从 X 获得两个不同子集，分别用于建立模型和评估它的 r_g，并重复多次后获得平均值，典型例子是 Stone 于 1974 年提出的交叉验证(cross-validation，CV)[1]，这条路计算量巨大，但做法简单，故仍有广泛应用。第二路是把 X 看成随机集，而参数估计 $\theta_k(X)$ 和相应的 $r_g(X)$ 都是随机变量，通过理论分析，近似估计 $r_g(X)$ 的期望值作为泛化风险。始于 1960 年出现的 Cp 准则，用于在平方误差 r_f 最小的线性回归分析中选择变量[2]。随后是以 $-\ln q(X|\theta_k)$ 为 r_f 的准则 AIC(Akaike

information criterion)[3]，用于线性自回归中的阶次分析，它包括 Cp 为特例，还渐近地等价于 CV[1]。AIC 的提出引发了对模型选择的广泛研究，三十余年后，Akaike 获得类似诺贝尔奖的日本 Kyoto 奖。第三路则基于 Vapnik-Chervonenkis(VC)理论，估计任何 θ_k 下 $r_g - r_f$ 的上界[4]，对定性分析有用，也给局限于两类识别的支持向量机提供了理论支持，但得到的上界通常含有未知系数，表达式也复杂，一般不便于实际应用。方向二不是估计泛化风险，而是寻求对数据 X 之最短编码。最早由Solomonoff 于 1964 年提出，1 年后，Kolmogorov 也独立给出，后来被称为 Kolmogorov 复杂度，即表述 X 所要的最短计算程序[5]。真正被用于参数学习和模型选择，是局限在概率模型下的最短编码，始于 MML(minimum message length)[6]，随后有 BIC(Bayesian information criterion)[7]和 MDL(minimum description length)[8]。这些研究相互影响，有共同特点，细节又不同。如图 4(a)所示，MML 和早期 MDL 的思路皆是两部分编码，即用 $q(X|\theta_k)$对 X 编码，并估计 $q(\theta_k)$用于对 θ_k编码，而 MML 和 MDL 的不同点在于对 $q(\theta_k)$的不同考虑，此办法也等价于基于非规范先验 $q(\theta_k)$的最大验后(MAP)估计。BIC 和后期的 MDL 则都是考虑单部分编码。如图 4(b)所示，BIC 或 Marginal Bayes 用$q(X|k)$对 X 编码。后期的 MDL 则用归一化后的 $q(X|\theta_k)$对 X 编码。最广泛应用的是这些研究所共同含有的主要部分。实际上，许多应用都涉及图4(c)所示隐变量模型，上述研究所致力考虑的都是消去隐变量后得到的 $q(X|\theta_k)$，实质上如同考虑一个无隐变量模型。

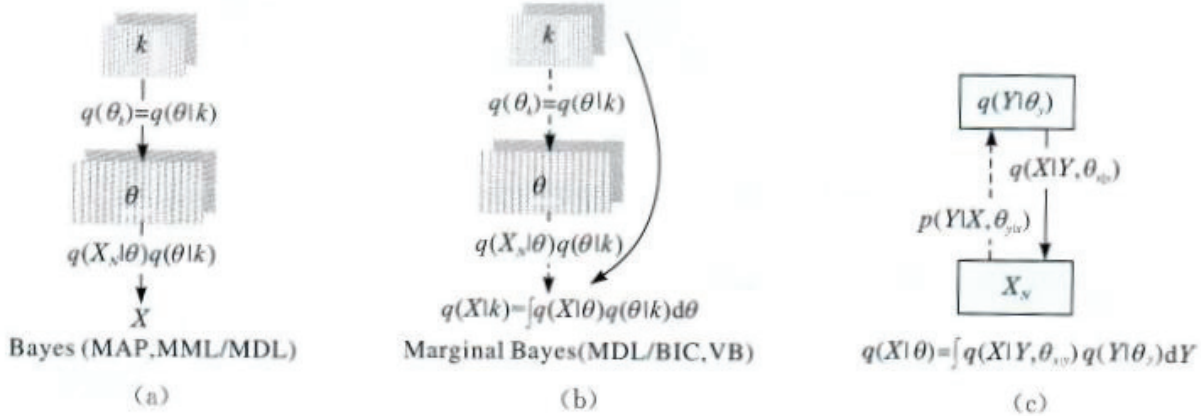


图 4　最短编码与 Bayes 方法

从 20 世纪 90 年代初至今是研究发展的第二阶段。除了上述成果的应用和延伸外，还出现了第三个方向，即 BYY 阴阳和谐学习。考虑图 5(a)所示的系统，内部表示 R 不仅有参数θ 为长期记忆，还包括隐变量 Y 为短期记忆。阴模型从内向外描述或重建数据 X，而阳模型作为阴模型的逆，从外向内将观察 X 映射到R，这些都在阴阳和谐原理下进行。阴阳和谐指阴阳以最默契(交换信息最小)

的方式达到最大共识。数学上通过图 5(a)中所列的和谐度 $H(p\|q)$之最大化来实现，不仅寻求阴能很好描述 X，而同时阳成为阴之最优逆，并且促使 BYY 系统复杂度尽量小。另一理解是，为了让阴能很好描述 X，阳需从 X 向 R 输送给阴的信息为 $-H(p\|q)$，它越小越好。在图 4 中，先验 $q(\theta_k)$是模型选择的仅有动因，$q(Y|\theta_k)$则躲在 $q(X|\theta_k)$的积分背后，实际上对模型选择没有帮助。而在图 5(a)中，Y 的复杂度 k_Y 是 k 的一部分；甚至在许多应用中，模型选择就是只选择复杂度 k_Y。$q(Y|\theta_k)$远比 $q(\theta_k)$容易准确获得，且在 $H(p\|q)$中占有与$q(\theta_k)$同等的地位，可以显著地改进模型选择。

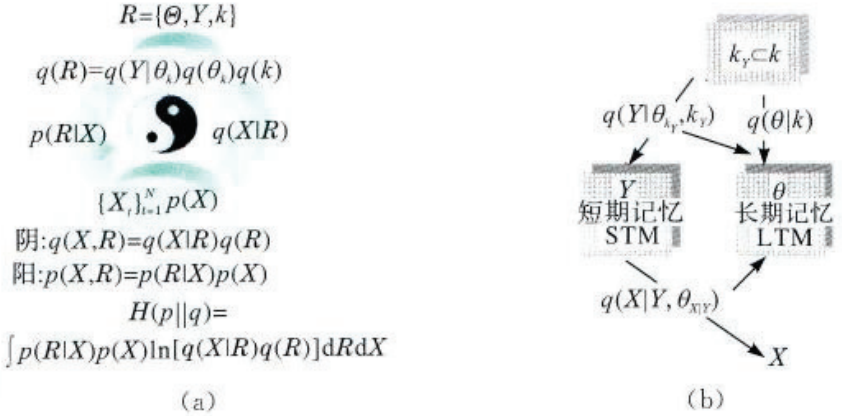


图 5　BYY 系统与最大和谐原理

第二阶段的另一方向是针对两步法实现的两个缺点(参见本文第二段内容)，寻求在参数学习中自动实现模型选择，即所依据的学习理论或机制能够驱使某参数子集在学习中趋向零，导致多余的模型复杂度被删除[9]。20 世纪 90 年代初提出的 RPCL 学习在竞争学习各类中心的同时能自动选择类别[10]。90 年代中期开始的 Lasso 和基于拉普拉斯先验的 Bayes 学习，在线性回归分析中能自动删除多余变量[11]。通过挑选先验 $q(\theta_k)$，图 4(a)的 MML[6] 和图 4(b)中基于 VB(variational Bayes)近似的 $q(X|k)$[12]，皆在 21 世纪初用于在学习中自动选择类别。在学习过程中自动实现模型选择，也是 BYY 阴阳和谐学习的重要特点之一，与 MML 和 VB 相比，优点是 $q(Y|\theta_k)$与 $q(\theta_k)$具有同等地位，有力地加强了学习中自动模型选择的能力。

虽然经过近半世纪的努力，可以认为模型选择仍是未决难题。已有各种方法中，没有哪一个被公认是最好的。机器学习的目的，即学习现有的去推断未来的，是造成困难的关键。此难题实质上涉及从哲学层面看什么是最好的学习理论。若相信现有的和未来的都遵循同一真理，则追求估计泛化风险是有道理的，应当致力改进基于 VC 理论的研究，令其便于实际应用，并建立它与最短编码和 BYY 阴阳和谐的内在联系。其实，许多人不再信奉这一经典哲学，转向寻

求能对数据最短编码的模型或等价地最可能产生这一数据的模型[5,8]。值得注意的是，Solomonoff 于 1964 年提出的基本思想近年来已成为新人工智能的理论基础[5]。基于多年研究，笔者相信作为中华核心文化之一的阴阳和谐哲学，是解决此难题的一个出路[9]。

参考文献

[1] Stone M. An asymptotic equivalence of choice of model by cross-validation and Akaike's criterion. J. R. Statist. Soc. B, 1977, 39: 44－47.

[2] Gorman J W, Toman R J. Selection of variables for fitting equations to data. Technometrics, 1966, 8: 27－51.

[3] Akaike H. A new look at the statistical model identification. IEEE Trans. on Automat. Contr., 1974, 19(6): 714－723.

[4] Vapnik V. The Nature of Statistical Learning Theory. New York: Springer, 1995.

[5] Schmidhuber J. The new AI is general and mathematically rigorous. Journal of Frontiers of Electrical and Electronic Engineering in China, 2010, 5(3): 347－362.

[6] Wallace C S, Boulton D M. An information measure for classification. Computer Journal, 1968, 11(2): 185－194.

[7] Schwarz G. Estimating the dimension of a model. Annals of Statistics, 1978, 6(2): 461－464.

[8] Rissanen J. Basics of estimation. Journal of Frontiers of Electrical and Electronic Engineering in China, 2010, 5(3): 274－280.

[9] Xu L. Bayesian Ying-Yang system, best harmony learning, and five action circling. Journal of Frontiers of Electrical and Electronic Engineering in China, 2010, 5(3): 281－328.

[10] Xu L, Krzyzak A, Oja E. Rival penalized competitive learning for clustering analysis: RBF net and curve detection. IEEE Transactions on Neural Networks, 1993, 4(4): 636－649.

[11] Williams P M. Bayesian regularization and pruning using a Laplace prior. Neural Computation, 1995, 7(1): 117－143.

[12] Corduneanu A, Bishop C M. Variational Bayesian model selection for mixture distributions//Jaakkola T, Richardson T. Artificial Intelligence and Statistics 2001. New York: Morgan Kaufmann Publishers, 2001: 27－34.

撰稿人：徐　雷

香港中文大学计算机科学与工程系