权益量化研究

申万宏源研究 SWS RESEARCH

2020 年 09 月 01 日

量化投资新起点

——机器学习系列报告之一

相关研究

证券分析师
邓虎 A0230520070003
denghu@swsresearch.com
于光希 A0230520060002
yugx@swsresearch.com

研究支持
于光希 A0230520060002
yugx@swsresearch.com

联系人
于光希
(8621)23297818×转
yugx@swsresearch.com

本期投资提示：

- 机器学习是人工智能的一个分支，也是人工智能的核心领域。机器学习的目的在于推理，推理的过程是学习，研究计算机如何模拟人类的学习行为。从 1930 年代至今，机器学习逐渐发展成为一门独立的学科，已有超过数百种算法被提出。《Do we need hundreds of classifiers to solve real world classification problems?》对 17 大类共 179 个分类器，在 121 个数据集上进行了测试。结果显示，随机森林和支持向量机（高斯核）效果最好，其次是神经网络和 Boosting 集成方法。
- 机器学习的一大发展趋势是大众化。早期的机器学习研究人员不仅需要对算法有深刻的理解，还需要具备较强的 C++实现能力。如今随着 Scikit-Learn、Tensorflow 等开源库的出现，机器学习的应用难度大大降低。
- 机器学习的发展与硬件和数据密不可分。CPU 速度的提升、GPU 的出现，使得计算机的计算能力实现了飞跃。大数据时代的到来和大量非结构化数据的出现，使得传统统计方法遇到了瓶颈，带动了机器学习的进一步发展。
- 机器学习的两大特点是自动学习的能力和数据驱动（data-driven）。对于某些类型的任务，很难通过显式编程指令完成。在这种情况下，最好的方法是使计算机具有从数据中学习的能力。以监督学习为例，算法学习的是一个从输入到输出的函数 f，而学习过程需要大量的数据。数据量越大，算法的学习效果越好。
- 越来越多的金融公司尝试将机器学习算法应用于金融市场，并已经在算法交易、智能投顾、反欺诈、风险管理、投资预测等方面取得了突出的成果，但仍有一些问题值得重视。首先，金融数据含有较多噪声。其次，金融数据的结构易发生变化。并且，主流的深度学习算法是一种“黑箱模型”，可解释性较差。最后，人工智能离人的智能还存在较大差距。
- 机器学习可以分为监督学习、无监督学习和强化学习三大类。监督学习是一种有标记的学习方法，输入是带标签的数据，学习目标是由输入到输出的函数。无监督学习的输入是不带标签的数据，学习目标是描述数据结构的函数。强化学习研究的是个体如何基于环境而行动，以取得最大化的预期奖励。
- 最后我们用上海地区二手房数据，以房价预测为例说明机器学习的标准工作流程。一个标准的机器学习项目工作流程主要包括以下几个方面：定义问题、数据预处理、建立基准模型、建立比较模型、交叉验证和参数调整等。相对于基准的线性回归模型，一个简单的双层神经网络模型在测试集的预测效果有明显提升，尤其对于异常值的拟合效果更佳。

申万宏源研究微信服务号

目录

图表目录

1．发展历程

2016 年 3 月，经过数以万计的自我对弈练习，AlphaGo 以 4：1 击败世界冠军李世石，同时也成为第一个在不让子情况下击败围棋职业九段棋手的电脑程序。围棋中的落点极多，计算量极大，“弃子争先”、“势孤取和”等术语也说明围棋中对局势的判断较为复杂。因此早在 1997 年，电脑程序“深蓝”就已经击败国际象棋世界冠军，而一直难以战胜围棋职业棋手。AlphaGo 的胜利是里程碑式的，而其最终版 AlphaGo Zero 仅训练 40 天就超越了所有旧版本。伴随着 AlphaGo 的横空出世，忽如一夜春风来，人工智能再一次进入了大众的视野。

1.1 算法简介

机器学习是人工智能的一个分支，也是人工智能的核心领域。机器学习的目的在于推理，推理的过程是学习，研究计算机如何模拟人类的学习行为。在几十年的发展过程中，机器学习已成为一门涉及数学、统计学、计算机等多领域的交叉学科，目前已广泛应用于数据挖掘、计算机视觉、自然语言处理、模式识别、反欺诈等领域。

虽然机器学习在近十年才逐渐走入人们的视野，但其基本算法可以追述到上世纪。1936 年，英国数学家 Turing 提出图灵机，尝试将人的计算行为抽象为数学逻辑。同年 Fisher 提出线性判分析（Linear Discriminant Analysis），通过线性变换将数据投影到低维空间。朴素贝叶斯(Naive Bayes)可以追溯到 1950 年代，是一种以贝叶斯定理为基础，假设特征之间相互独立并建构分类器的简单方法。1958 年，Cox 提出逻辑回归（Logistic Regression），是一种广义线性模型，可以预测样本属于不同类别的概率，同样适用于分类问题。朴素贝叶斯和逻辑回归形式简单、算法稳定，至今仍是分类问题的标准选择。

1957 年，Rosenblatt 提出感知机（Perceptron）。感知机是一个简单的二元线性分类器，甚至无法处理异或问题。但感知机可以被看做形式最简单的前馈神经网络，也是后来深度神经网络的雏形。1967 年，K-近邻算法 (K-Nearest Neighbors) 出现，是一种非参数统计方法，也是最简单的机器学习算法之一。K-近邻算法在对新样本进行分类时只依据最近邻的 K 个样本的类别来决定新样本的类别，可以进行简单的模式识别。

从 1960 年代末到 1970 年代末，机器学习乃至人工智能经历了一段时间的冷却器。由于当时计算机的内存容量和处理速度都有限，算法在实际问题的应用过程中遭遇巨大瓶颈。直到 1980 年，第一届机器学习国际研讨会在美国卡内基梅隆召开。机器学习进入迅速发展的新阶段，真正成为一个独立的研究方向，大量算法提出。

决策树是机器学习中的一个重要模型，至今仍被广泛使用。决策树是一个有向无环图，树中每个节点表示某个决策规则，每个分支路径则表示在该决策规则下可能的属性值，而每个叶节点则最终输出一个决策。1986 年，Quinlan 提出了著名的 ID3

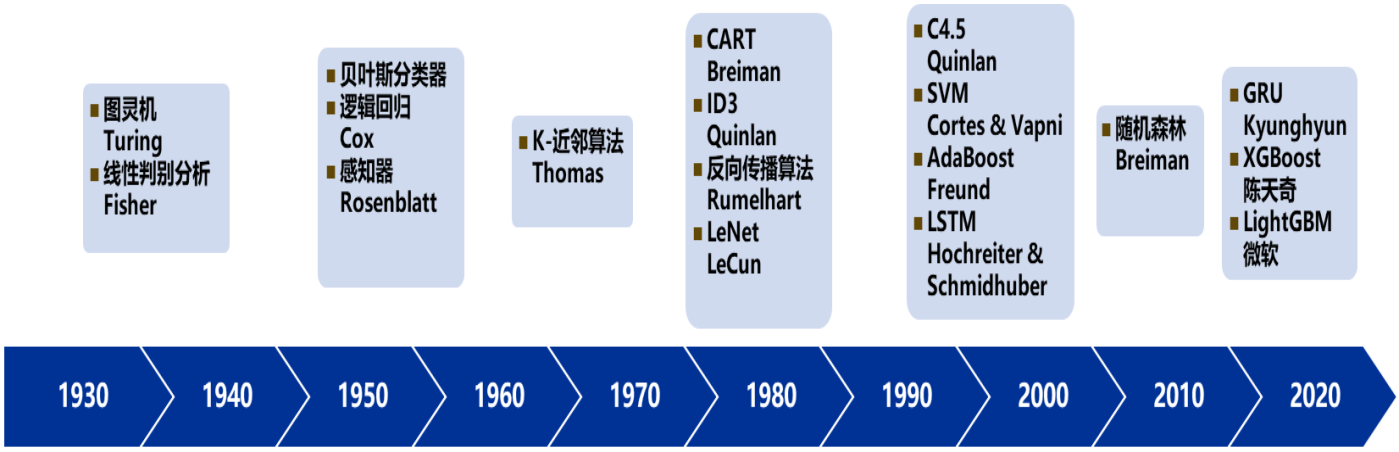
算法。之后的 1993 年，在 ID3 算法的基础上，Quinlan 又提出了改进的 C4.5 算法。ID3 算法和 C4.5 算法构造的都是分类树，1984 年，CART（Classification and Regression Trees）算法出现，CART 既可以构造分类树，也可以构造回归树。决策树的优点在于易于理解和实现，对缺失数据不敏感，模型分裂过程也非常直观。这类模型通常形式简单，且具有较好的可解释性，因此被称为“白盒模型”。

与“白盒模型”相对的是“黑盒模型”，主要指深度神经网络。这类模型通常形式复杂，具有很高的准确性。但这些模型的内部工作机制难以理解，并且不能估计每个特征的重要性，不具备可解释性。1986 年，用于训练深层神经网络的反向传播算法出现，目前仍然是深度学习中被使用的训练算法。1989 年，LeCun 提出第一个真正意义上的卷积神经网络，也是后来 LeNet 的原型。由于当时计算机硬件条件的限制，LeNet 主要被应用于读取邮政编码等字符识别任务。1997 年，Hochreiter 和 Schmidhuber 提出了 LSTM（Long Short-Term Memory），是一种特殊的循环神经网络。LSTM 克服了循环神经网络中的长期依赖问题，至今被广泛应用于自然语言处理。2014 年，在 LSTM 的基础上，Kyunghyun 提出了更为简化的 GRU（Gated recurrent unit）。

1995 年，Cortes 和 Vapni 提出了现代版本的支持向量机（Support Vector Machine）。支持向量机的目标是寻找一个决策边界，决策边界是一个超平面，在决策边界两侧的数据分属于不同类别。支持向量机的原理非常简单，将数据映射到一个更高维的空间里，同时让决策边界与每个类别最近的数据间隔最大化，这样决策边界对新数据具有更好的分类能力。除了进行线性分类之外，支持向量机还可以使用函数有效地进行非线性分类。以支持向量机为代表的浅层学习和以神经网络为代表的深层学习一直处于交互竞争的状态。由于支持向量机同时适用于各种非线性问题，决策边界可解释性强，并且有严谨的数学理论支持，在很长一段时间内令神经网络黯然失色。

同样在 1995 年，Freund 提出一种集成学习算法 AdaBoost（Adaptive Boosting）。集成算法分为 Bagging、Boosting 和 Stacking，AdaBoost 是 Boosting 集成算法的代表，前一个分类器的误差被用来训练下一个分类器。Bagging 集成算法的代表是随机森林，由 Breiman 在 2001 年提出。随机森林以决策树为基础，用随机的方式生成多棵相互独立的决策树，并且以决策树输出类别的众数决定最终输出。集成学习的思想非常简单，综合多个简单的弱分类器的输出结果，能够达到强分类器的效果。并且集成模型不会过度依赖某一个分类器，抑制了决策树中常见的过拟合问题，大大增加了模型的泛化能力。由华人学者陈天奇提出的 XGBoost（2014）和微软开发的 LightGBM（2016）至今仍然是机器学习竞赛的首选。

图 1：机器学习算法发展简史（1930-2020）



资料来源：申万宏源研究

1.2 算法比较

从 1930 年代至今，机器学习逐渐发展成为一门独立的学科，已有超过数百种算法被提出。《Do we need hundreds of classifiers to solve real world classification problems?》对 17 大类共 179 个分类器（包括判别分析、贝叶斯、神经网络、支持向量机、决策树、集成方法、广义线性模型、K-近邻，偏最小二乘、主成分回归、逻辑和多项式回归等），在 121 个数据集上进行了测试。结果显示，随机森林效果最好，最高准确度达到 94.1%，并且在 84.3%的数据集中超过 90%。其次是高斯核的支持向量机，最高准确度 92.3%，且与随机森林之间的差异没有统计显著。其次是神经网络和 Boosting 集成方法。

1.3 大众化

机器学习的一大发展趋势是大众化。早期的机器学习研究人员不仅需要对算法有深刻的理解，还需要具备较强的 C++实现能力。如今随着 Scikit-Learn、Tensorflow 等开源库的出现，机器学习的应用难度大大降低。因此越来越多的非专业研究人员加入机器学习，使用机器学习解决领域问题，促进了机器学习的推广。

2. 机器学习与金融

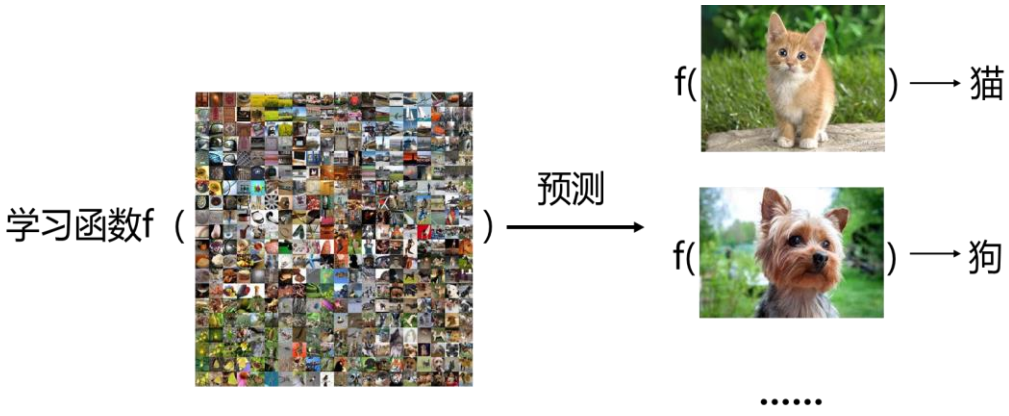
2.1 为什么是机器学习?

机器学习的一个特点是自动学习的能力。Arthur Samuel（1959）将机器学习定义为“研究如何在不输入显式编程指令的情况下，使计算机获得学习的能力”。Tom Mitchell（1997）在定义机器学习时提到，“机器学习是对能通过经验自动改进的计算机算法的研究”。对于某些类型的任务，很难通过显式编程指令完成。在这种情况下，最好的方法是使计算机具有从数据中学习的能力。

比如你希望建立一个垃圾邮件过滤系统，你可以根据过去垃圾邮件的内容，制定一些筛选规则（如正文内容包含特定字符、发送人邮箱地址等）。但这样做的缺点是你需要不断添加新规则，因此程序会难以维护。同时由于显示指定了筛选规则，垃圾邮件的内容只需微小的变化（或避开某些关键词），便很容易躲过过滤系统。而如果使用机器学习对邮件内容进行建模，可以自动检测垃圾邮件中的异常模式，学习筛选规则。这样的过滤系统不仅易于维护，而且是自适应的，可以根据增量数据进行自动更新。

机器学习的另一个特点是数据驱动（data-driven）。机器学习研究的是一种特殊算法，能够让计算机在数据中学习从而进行预测。以监督学习为例，算法学习的是一个从输入到输出的函数 f，而学习过程需要大量的数据。数据量越大，算法的学习效果越好。以大型视觉数据库 ImageNet 为例，包含超过 1400 万张手动标记的图片，共 2 万多个类别，每个类别都包含数百个图像。

图 2：机器学习的学习过程需要大量数据



资料来源：ImageNet，申万宏源研究

2.2 为什么是现在?

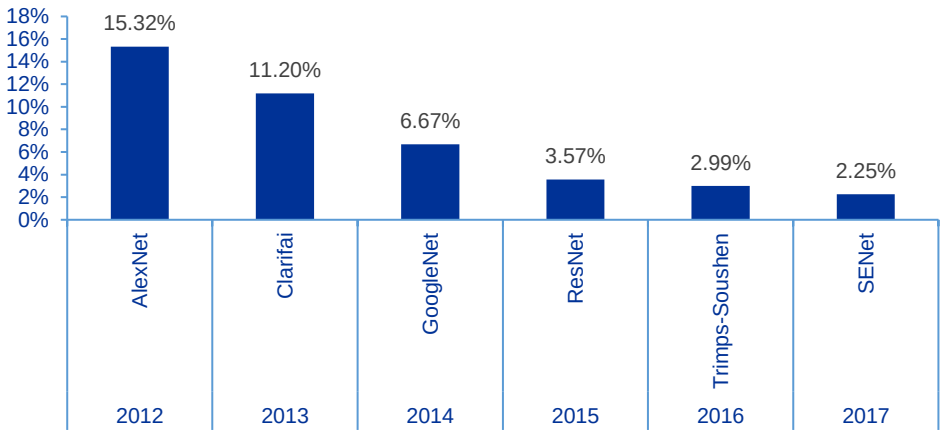
从上文可以看出，很多经典的机器学习算法在上世纪就早已提出，但为什么机器学习近十年才兴起？机器学习的发展与硬件和数据密不可分。

与 1990 年代相比，CPU 的速度大约提升了 5000 倍。更重要的是，出现了用于游戏 3D 渲染的 GPU，更加适合训练深度学习算法，使得计算机的计算能力实现了飞跃。硬件的发展使得复杂模型的训练成为可能，在 2015 年 10 月，AlphaGo 就使用了 1202 块 CPU 以及 176 块 GPU 进行训练。

没有大数据的兴起，人工智能只是空中楼阁。伴随着存储硬件的指数增长和互联网的迅猛发展，各种形式的数据已经渗透到每一个行业。大数据具有 3V 特点，即数量多（Volume）、速度快（Velocity）、多样性（Variety）。在互联网时代到来之前，计算机视觉、自然语言处理等领域的研究数据根本无从收集。大量非结构化数据的出现，使得传统统计方法遇到了瓶颈，带动了机器学习的进一步发展。

在具备了硬件升级和数据增量的条件之后，深度学习在近十年迎来飞速发展。自从 2010 年以来，ImageNet 每年举办一次大规模视觉识别挑战赛 ILSVRC。AlexNet 在 2012 年的挑战赛上取得了巨大的突破，之后历年冠军基本都基于深度神经网络，并且表现越来越好。

图 3：ILSVRC 历年冠军基本都基于深度神经网络



资料来源：ImageNet，申万宏源研究

2.3 如何应用于金融?

随着机器学习在计算机视觉、自然语言处理等领域取得巨大进展，学术界也对机器学习在金融领域的应用进行了大量研究。《Empirical Asset Pricing via Machine Learning》基于 1957-2016 年间近 30000 只股票数据，比较了线性回归、正则化的广义线性回归、梯度提升树、随机森林、神经网络等机器学习算法的效果。实证结果显示利用机器学习进行预测可以为投资者带来超额回报，其中树模型和神经网络的效果要优于回归模型。

同时越来越多的金融公司开始使用科技进行金融创新，尝试将机器学习算法应用于金融市场，并已经在算法交易、智能投顾、反欺诈、风险管理、投资预测等方面取得了突出的成果。2017 年 10 月 18 日，全球第一只人工智能基金 AI Powered Equity ETF 诞生，只应用人工智能和机器学习进行投资决策。2019 年 3 月 7 日，高盛又推出 5 只依靠人工智能算法的 ETF。

表 1：国内外各大金融公司积极探索机器学习在金融中的应用

应用领域	公司
机器学习交易策略	英仕曼集团
生成量化模型信号	纽约人寿投资公司
深度学习预测股票成交量	道富集团
空间另类数据分析	高盛
剖析财报电话会议	美国世纪投资公司
信用资产投资组合管理	中国人寿资产管理公司

处理保险索赔以及评估企业风险	平安集团
市场情绪分析	彭博

资料来源：CFA Institute，申万宏源研究

虽然机器学习在金融中的应用逐渐被普及，但仍有一些问题值得重视。首先，金融数据含有较多噪声，越来越多的非结构化数据更加突出了这个特点。金融市场的定价机制本身就很复杂，在涉及交易问题的时候，影响因素更多。除了市场本身，还有心理、博弈等因素。因此如何从大数据中去除噪声，寻找真正的信号，对机器学习是一大挑战。

其次，金融数据的结构易发生变化。对于正在发展中的国内金融市场，这一点更加明显。机器学习从历史数据中学习规律，并预测未来，因此模型存在着失效风险。并且，主流的深度学习算法是一种“黑箱模型”，可解释性较差。而金融比较看重行为背后的逻辑，这使得深度学习的应用受到一定限制。

最后，人工智能离人的智能还存在较大差距。机器学习比人更加擅长的是大数据量的简单运算，但内在的逻辑认知不是强项。比如一个小孩子，在看到几十幅动物图片后，就可以进行图像识别。但对于机器而言，这个简单的任务仍然需要数万张图片的训练才能完成。

3．机器学习分类

机器学习可以分为监督学习、无监督学习和强化学习三大类。

3.1 监督学习

监督学习是一种有标记的学习方法，输入是带标签的数据，学习目标是由输入到输出的函数。函数的输出可以是连续值（回归），也可以是离散值（分类）。在监督学习中，每一个输入数据都有一类标签（特征），标签的存在可以让算法更好地“学习”输入与输出之间的内在逻辑关系。多数机器学习算法属于监督学习，包括线性回归、决策树、支持向量机、神经网络等。

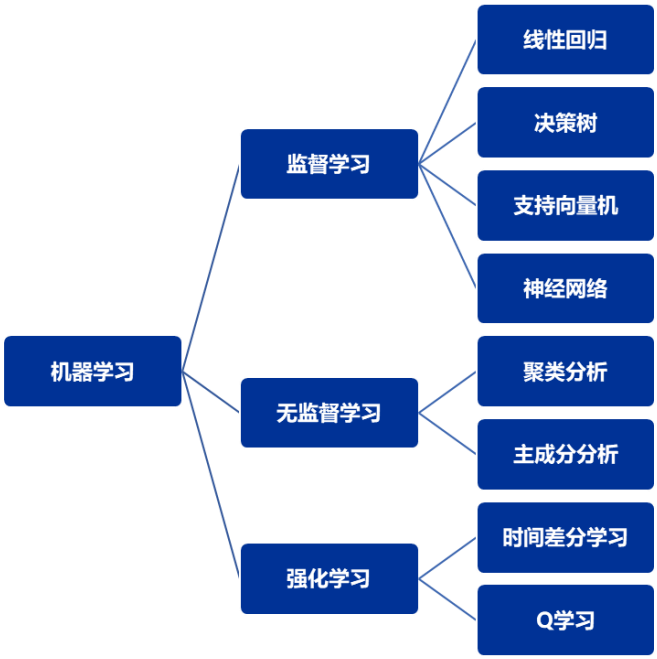
3.2 无监督学习

与监督学习不同的是，无监督学习的输入是不带标签的数据，学习目标是描述数据结构的函数。无监督学习从训练数据中自动推断结论，探索数据的内部结构。最典型的无监督学习包括聚类分析和主成分分析。聚类分析可以对不带标签的数据进行分组，相同子集中的数据具有相似属性，可以发现数据间隐藏的模式。主成分分析利用正交变换，将原始数据转换为一组线性不相关的变量，是一种常见的降维方法。

3.3 强化学习

强化学习研究的是个体如何基于环境而行动，以取得最大化的预期奖励。强化学习是一个动态的过程，个体通过与环境进行交互，并通过环境反馈的奖惩来决定下一步行动，学习最优策略。强化学习算法包括时间差分学习和 Q 学习等，AlphaGo 就运用了深度强化学习技术。

图 4：机器学习主要分类



资料来源：申万宏源研究

4. 标准工作流程

下面我们用一个实际例子说明机器学习的标准工作流程，以及在机器学习项目中需要用到的一些知识。一个标准的机器学习项目工作流程主要包括以下几个方面：定义问题、数据预处理、建立基准模型、建立比较模型、交叉验证和参数调整等步骤。在机器学习项目中，首先需要明白面对的问题是什么，这样才能确定预测的目标。数据预处理包含特征工程、缺失值和异常值、划分训练集和测试集、标准化等，干净的数据是模型成功的必要条件。在正式建立模型前，首先需要建立一个基准模型。在建立比较模型后，根据模型在验证集上的拟合效果，我们需要对模型进行评估，并调整模型的超参数。在数据较少的情况下，一般通过交叉验证完成。

图 5：机器学习的标准工作流程

资料来源：申万宏源研究

4.1 定义问题

在机器学习项目中，首先需要明白面对的问题是什么，这样才能确定预测的目标。在大型机器学习项目中，这一步尤其关键。你不仅需要明确输入数据和输出数据的形式，还需要确认数据的可得性。同时在逻辑层面上，思考一些更宏观的问题。机器学习模型的训练是建立在历史数据基础上的，识别的模式也是一种对历史数据规律的总结。使用历史规律预测未来，这样的逻辑架构在面对的实际问题上是否合理？

在正确定义了问题之后，便可以开始数据的收集和处理。在这个例子中，我们使用的是从网络爬取的上海地区二手房信息，共 92510 条数据。输入数据是房屋的相关特征（包括厅室、朝向、建筑时间等），输出数据是房屋的均价（万元/平方米），因此是一个监督学习中的回归问题。

4.2 数据预处理

数据预处理也被称为数据清洗。顾名思义，在实际问题中可以获得的原始数据大多是不规范的，比如数据不完整（含有缺失值或 0）、包含较多噪声（异常值）、格式不一致（文本型数据或包含乱码）。因此需要对原始数据进行清洗，使之成为算法适合的数据类型。算法是烹饪方法，模型是锅碗瓢盆，而数据则是做菜原料。良好的食材是美食的基础，干净的数据一样是模型成功的必要条件。尤其是机器学习问题中的数据量较大、特征较多，数据预处理就更加重要。数据预处理主要包括特征工程、处理缺失值和异常值、划分训练集和测试集、标准化等步骤。

4.2.1 特征工程

特征工程指对原始特征进行合并、变换等操作，进行特征提取和特征选择。在一些实际问题中，原始特征的维度非常高，很容易导致模型拟合不佳，因此需要提取有效特征。虽然深度学习端到端的特性，一定程度上降低了特征工程的要求。但在金融数据的处理中，特征工程仍是必不可少的一步。筛选有效特征，建立在深入理解数据特性的基础上，因此特征工程一般需要借助领域知识的帮助。特征选择需要分析特征之间的有效性，如相关系数、条件熵等。也有一些算法具有特征选择的功能（如 LASSO、逐步回归等），可以自动进行特征选择。

在这个例子中，原始数据是房屋的信息，所有特征都不是数值型。因此需要提取特征中的有用信息，并转换为数值型数据。比如对于建筑时间，其实我们更加关心的是楼龄，因此可以将建筑时间转换为楼龄这个新特征。对于面积和总价，可以计算均价并作为模型的输出目标。而经纪人则是一个无关特征，可以删除。

表 2：特征处理方法

特征名称	示例	处理方法	处理后特征
厅室	2 室 2 厅	拆分	厅、室
楼层	中层（共 18 层）	one-hot 编码	中层、高层、低层
朝向	南北向	one-hot 编码	南向、北向、南北向等 13 个新特征
建筑时间	2014 年	转换	楼龄
经纪人		删除	-
面积/总价	121.2 ㎡/468 万	计算	均价

资料来源：申万宏源研究

楼层和朝向这两个特征是类别变量，且包含了较多的类别（如高层、中层、低层和南向、北向、南北向等）。在机器学习中，对于类别变量，一般有整数编码和 one-hot 编码两种编码方式。整数编码为每个类别都赋予一个整数值，如 1、2、3 等。而 one-hot 编码为每个类别都创建一个二值变量，在该类别时二值变量为 1，否则为 0。

整数编码的缺点在于，类别值之间存在严格的次序关系。即假定类别值越高，该类别更好。因此在机器学习，尤其是自然语言处理相关问题中，one-hot 编码更加合适。因此使用 one-hot 编码转换将楼层和朝向这两个特征转换为数值型变量。

图 6：one-hot 编码和整数编码

特征	楼层
样本1	高层
样本2	中层
样本3	低层

one-hot编码

特征	南向	北向	南北向
样本1	1	0	0
样本2	0	1	0
样本3	0	0	1

特征	楼层
样本1	高层
样本2	中层
样本3	低层

整数编码

特征	朝向
样本1	1
样本2	2
样本3	3

资料来源：申万宏源研究

在经过以上处理后，特征数量由 5 个变为 18 个。

4.2.2 缺失值和异常值

实际的金融数据中大多含有缺失值，比如固定资产投资数据 1 月份不公布、财务数据部分科目缺失等。对于缺失数据，有很多处理方法。可以直接删除缺失值，也可以对缺失值进行填充（平均值、插值法、K 近邻等），还可以对缺失值进行拟合（线性回归、最大期望算法等）。对于一些基于决策树的算法，如 XGBoost 和 LighGBM，对缺失值不敏感，也可以忽略缺失值。

对于异常值的处理，一般使用 3sigma 原则：若数据服从正态分布，则与平均值偏离超过 3 倍标准差（sigma）的数据被判定为异常值。因为在正态分布的假设下，数据在平均值 3 倍标准差之外的概率很小。有时候也需要人为设定阈值，进行异常值的判定。

由于本例中缺失值的数量并不是很多，所以对于含有缺失值的数据，可以直接删除。在经过以上数据预处理后，还有 76764 条数据。

4.2.3 划分训练集和测试集

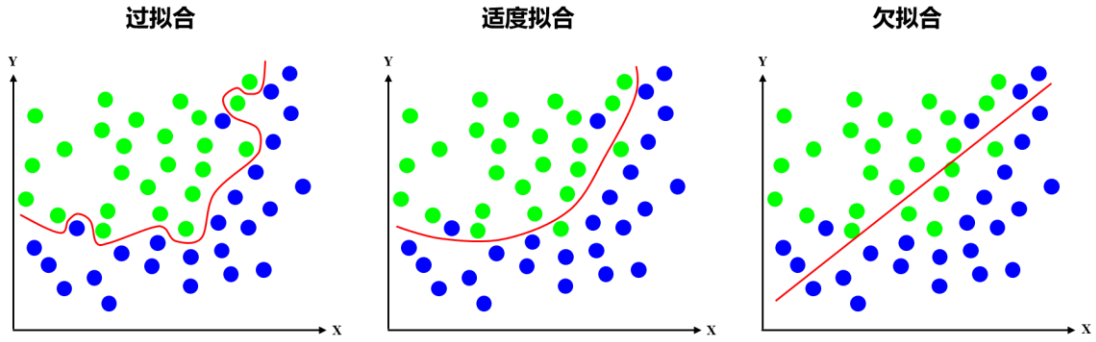
在监督学习中，一般要将数据集划分为训练集和测试集。在本例中，我们按照 75%和 25%的比例划分训练集和测试集。训练集的数据用来训练模型，测试集的数据用来测试预测效果。

为什么要这样做呢？因为在机器学习中，我们关注的是训练得到的模型在新数据集上的预测能力，这种能力被称为模型的泛化能力。为了验证这种泛化能力，需要将原始数据集分层。形象地讲，训练集起到“老师”的作用，训练模型学习数据的内在特征，并寻找输入与输出之间的内在逻辑关系。测试集起到模型评估的作用，将输入带入模型“学习”到的逻辑关系中，即可获得预测值，并与实际值进行对比，进行模型评估。

除了模型训练过程中可以学习到的参数（如神经网络中各层神经元的权重），还有一些模型之外的参数被称为超参数（如每层网络神经元的个数），通常需要预先指定初始值。因此可以进一步将训练集进一步划分为训练集和验证集，其中验证集的作用是进行超参数的选择。

因为在机器学习中更加关注模型的泛化能力，所以通常最优模型并不是完美拟合训练集的模型。因为训练集的数据同时包含了随机的噪声，泛化能力强的模型应该将噪声过滤掉。正如在考试之前，我们会做很多模拟题，但如果只是记住模拟题的答案，在考试中也无法取得好成绩。

图 7：训练模型的过程就是寻找适度拟合的模型



资料来源：申万宏源研究

所以在训练过程中需要避免模型过度（完美）拟合训练数据，这种情况被称为过拟合。如果在训练过程中使用一个参数过多的复杂模型，在训练集上可以得到完美的拟合效果。但由于同时拟合了噪声，这样的模型泛化能力通常较差。过拟合是模型训练过程中的一个常见问题，而防止过拟合是参数调整过程中的重要工作。

而如果在训练过程中仅仅使用一个简单的模型，或许不足以描述输入与输出之间的复杂关系，这种情况被称为欠拟合。在模型复杂度和泛化误差之间寻找一个平衡点，训练模型的过程就是寻找适度拟合的模型。

4.2.4 标准化

机器学习中的数据维度通常较高，特征间的数值差异相差很大。为了保证模型的拟合效果，需要对原始数据进行标准化处理。标准化的方法有很多，常用的有 min-max 标准化、z-score 标准化、log 函数、归一化等。

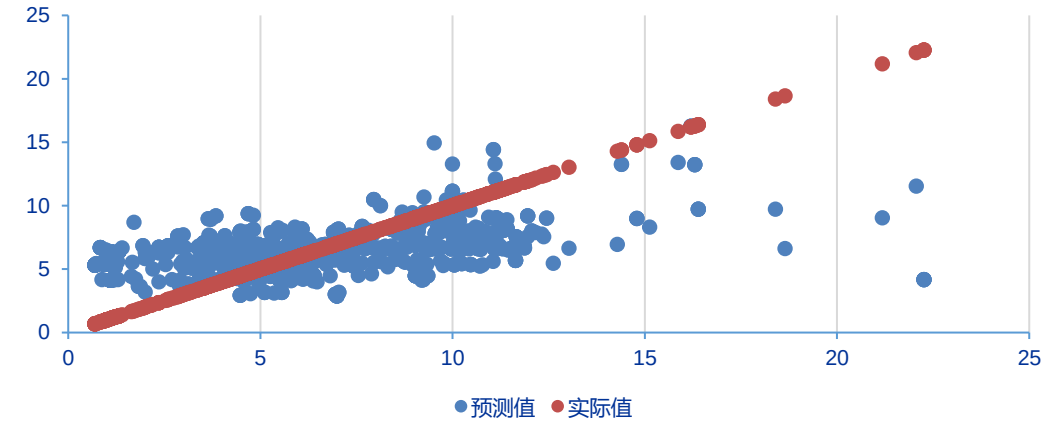
本例中我们使用 z-score 标准化，即对每个特征，减去其平均值，再除以其标准差。需要注意的是，这里的均值和标准差都是训练集数据的。在模型的整个训练阶段，不应该含有任何测试集数据的信息。

4.3 建立基准模型

在正式建立模型前，首先需要建立一个基准模型。这里我们使用线性回归建立一个基准模型，供后续模型比较。由于目标预测房屋的均价，因此使用 MAE（Mean Absolute Error，平均绝对误差）表示预测值和观测值之间绝对误差的平均值，作为模型的评价指标。

从拟合效果看，线性回归的 MAE 较高，达到了 2.49。可以看到模型出现明显的欠拟合，说明线性回归无法学习到输入输出之间复杂的函数关系。

图 8：线性回归部分测试集数据预测值与实际值的偏离程度（横坐标表示实际值）



资料来源：申万宏源研究

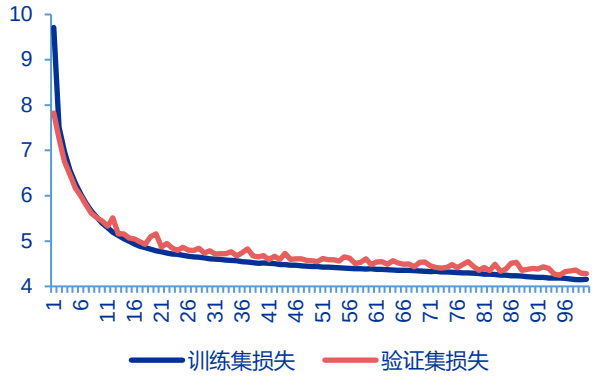
4.4 建立比较模型

接下来我们使用一个双层神经网络，对这个问题进行优化。我们构建的神经网络每一层有 64 个神经元，使用 ReLU（Rectified Linear Unit）激活函数。线性回归的模型简单，最优参数可以求得解析解。而对于大多数机器学习模型，最优参数通常没有解析解，因此需要使用数值计算方法近似求解。最常用的求解方法是基于梯度下降的优化方法，梯度项的计算需要遍历数据集中所有样本，遍历整个训练数据集的过程被称为一个训练时期(epoch)。

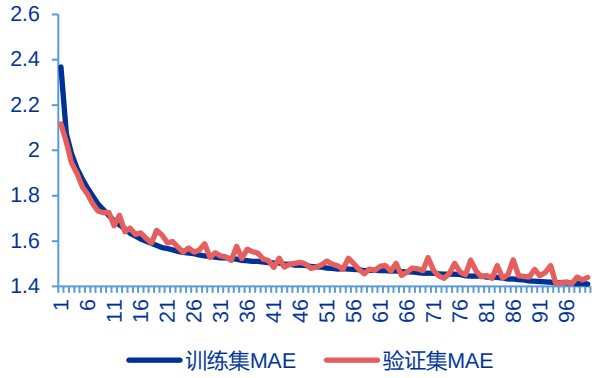
由于在训练中需要使用梯度下降算法，因此神经网络的训练过程是不稳定的。我们需要不断观察模型在训练集和验证集上的误差，直至出现明显的过拟合迹象。可以看到，在进行了 100 次训练之后，验证集的损失函数和 MAE 逐步稳定，没有出现明显的下降。

图 9：神经网络模型不同训练时期的损失函数

图 10：神经网络模型不同训练时期的 MAE



资料来源：申万宏源研究

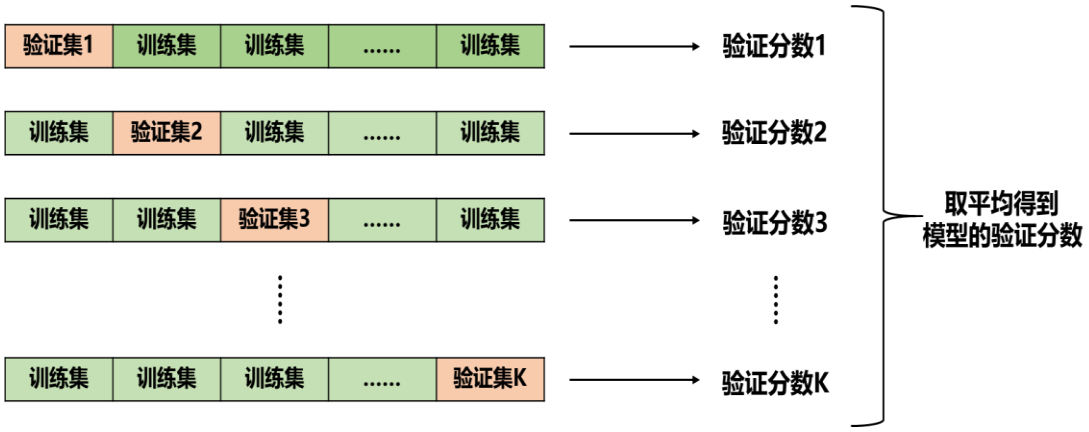


资料来源：申万宏源研究

4.5 交叉验证和参数调整

本例中的训练次数可以看做模型的一个超参数，因此需要将训练集划分成训练集和验证集，对超参数进行选择。在模型评估的过程中，由于复杂机器学习的算法训练并不稳定，在验证集数据较少的情况下，采用单一的验证集评估可能波动很大。所以多数时候我们采用交叉验证的方法对模型进行更加稳定的评估，其中最常用的是 K 折交叉验证。

图 11：K 折交叉验证步骤

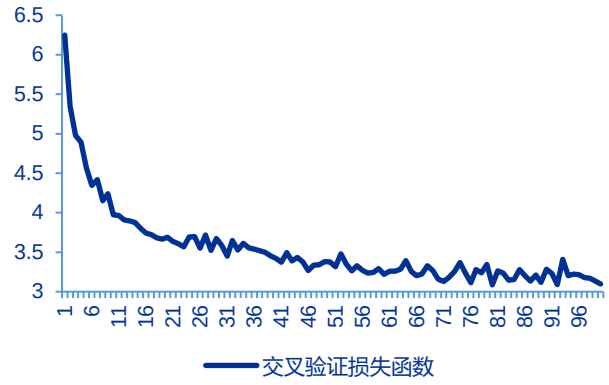


资料来源：申万宏源研究

K 折交叉验证将训练集数据划分为 K 个子集，使用 K-1 个子集进行训练，另一个子集进行验证。如此交叉验证重复 K 次，保证每个子集进行一次验证，得到 K 个验证分数，模型的最终验证分数等于 K 个验证分数的平均值。在数据较少的情况下，交叉验证可以划不同子集，实现了在同一数据集上训练多个不同模型，相当于增加了数据量。同时重复运用随机划分的子集进行训练和验证，在一定程度上也增强了模型的泛化能力。

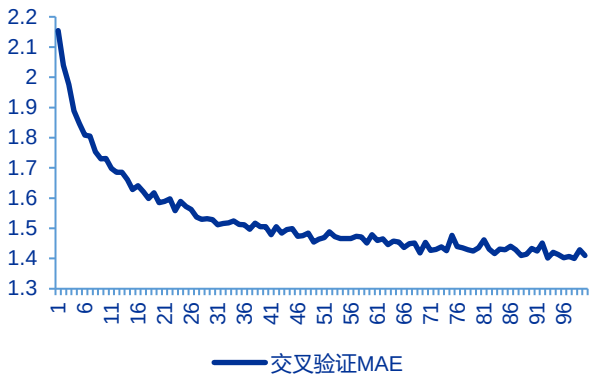
在本例中，由于训练数据较多，验证集本身评估效果较为稳定，因此使用交叉验证的评估结果和原始评估结果较为一致。

图 12：交叉验证得到的验证集损失函数



资料来源：申万宏源研究

图 13：交叉验证得到的验证集 MAE

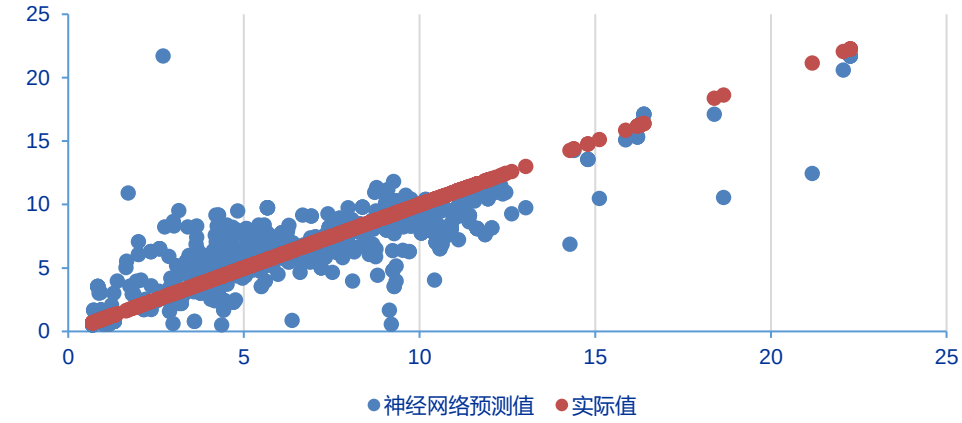


资料来源：申万宏源研究

在完成模型评估和参数调整后，就可以在测试集上进行预测。神经网络模型在测试集的 MAE 下降到 1.12，相对于线性回归有明显的提升，尤其对于异常值的拟合效果更佳。

但同时我们也注意到，测试集上目标的均值为 6.57，因此即使神经网络模型仍有 17%左右的误差。一是因为任何模型本身都存在一定的泛化误差，否则会过拟合。二是因为我们只选择了 18 个特征，而影响房价的其他因素还有很多，如配套设施、交通条件、小区环境、房产税等。在增加有效输入特征的情况下，模型会取得更好的效果。

图 14：神经网络部分测试集数据预测值与实际值的偏离程度（横坐标表示实际值）



资料来源：申万宏源研究

信息披露

证券分析师承诺

本报告署名分析师具有中国证券业协会授予的证券投资咨询执业资格并注册为证券分析师，以勤勉的职业态度、专业审慎的研究方法，使用合法合规的信息，独立、客观地出具本报告,并对本报告的内容和观点负责。本人不曾因，不因，也将不会因本报告中的具体推荐意见或观点而直接或间接收到任何形式的补偿。

与公司有关的信息披露

本公司隶属于申万宏源证券有限公司。本公司经中国证券监督管理委员会核准，取得证券投资咨询业务许可。本公司关联机构在法律许可情况下可能持有或交易本报告提到的投资标的，还可能为或争取为这些标的提供投资银行服务。本公司在知晓范围内依法合规地履行披露义务。客户可通过 compliance@swsresearch.com 索取有关披露资料或登录 www.swsresearch.com 信息披露栏目查询从业人员资质情况、静默期安排及其他有关的信息披露。

机构销售团队联系人

华东	陈陶	021-23297221	chentao1@swhysc.com
华北	李丹	010-66500631	lidan4@swhysc.com
华南	陈左茜	755-23832751	chenzuoxi@swhysc.com
海外	朱凡	021-23297573	zhufan@swhysc.com

法律声明