第*1*章

## 第1章 竞赛介绍

随着互联网时代的到来，以及计算机硬件性能的提升，人工智能在近几年可以说是得到了爆发式的发展。互联网时代带来了大量的信息，这些信息是名副其实的大数据，性能极佳的硬件也使得计算机的计算能力大大增强，二者结合到一起，人工智能的蓬勃兴盛就变成了自然而然地事情。机器学习作为一种传统的可解释性较强的算法，在人工智能的三驾马车中的算法中也是占有一席之地。本书几经商榷最终定名为《机器学习算法竞赛实战》，意在帮助机器学习初学者通过实战的方法从虽然很优美但是略显枯燥的各种公式和理论当中脱离出来，感受机器学习在实际应用中的奥秘，而竞赛则是一种最特殊的实战。

之所以强烈推荐竞赛作为机器学习实战的重要方式，是因为它实在是快速入门机器学习的极佳方式。对于初学者来说，他们的水平不足以支撑他们直接进到企业接触实际的应用场景，而书里得来的也是终究有些浅薄，这甚至包括本书在内。提起竞赛，大家总不免想起高中时期的各种数学物理化学竞赛，这些竞赛门槛极高且国内国外都有，能拿到好名次的同学甚至能够直接保送国内国外各大知名院校，因此竞赛二字总有一种令人望人生畏的感觉。而近几年人工智能的兴起催生出了各种算法竞赛则相对友好许多也有意思得多，时代的洪流之下各行各业都在寻求生存之道，利用先进的技术进行转型则是一个很好的办法，有些企业就开始寻求人工智能的助力，开始向社会征求优秀的算法解决方案此外，在学术领域的研究者们也渴望获得企业的场景和数据用于算法研究，这就催生出了各种竞赛平台，本书主要给读者介绍机器学习相关算法竞赛经验。

对于有志于进军机器学习相关领域研究或者从事相关工作的初学者来说，竞赛是性价比极高的一个实战选择，可以说是零门槛任何人都能参加，当然，一般主办方自己的员工是不被允许参赛的，即使参赛也不能参与排名。各种各样的竞赛可以说能够覆盖很多行业的典型应用场景，让参赛者不仅能够得到实战的锻炼，也能体会到机器学习在各个行业下沉应用带来的魔力，甚至还能在这里见到很多行业大佬，结交一些志同道合的朋友。

本章将主要从竞赛平台、竞赛流程以及竞赛类型三个部分给大家做竞赛相关介绍，第一节竞赛平台旨在介绍国内外知名的算法竞赛平台，以帮助读者快速了解竞赛渠道。第二节则讲述了完成一次机器学习算法竞赛的大致流程，以及每个模块的功能作用，更详细的内容会在后面章节给出。第三节则会为读者介绍常见的竞赛类型，帮助读者了解机器学习算法竞赛适用场景及业界的需求。

### 1.1 竞赛平台

#### 1.1.1 Kaggle

机器学习领域的大牛吴恩达老师曾经讲过，机器学习大多数时候就只是数学统计，数据相关的特征工程直接决定了模型的上限，而算法只是不断地去逼近这个上限而已。在机器学习领域有一个十分生动的比喻，建模的过程就好比是做饭，数据代表了食材，算法则是烹饪过程，最终饭菜的可口程度就是模型的效果。观看众多的美食影像，如央视著名的美食纪录片《舌尖上的味道》、以及《风味人间》等可以发现，众多的篇幅讲述了新鲜食材的获取，古话里也讲巧妇难为无米之炊，由此可见食材的重要性。同比到机器学习算法竞赛，数据的重要性不言而喻，这也就是下面将介绍的国际竞赛平台Kaggle(https://www.kaggle.com/)对自身的定位：数据科学之家。

打开Kaggle的网站首页，顶部有五个主要部分，即竞赛单元Competitions、数据集Datasets、代码笔记Notebooks、社区讨论Discussion以及在线课程Courses，作为全世界最受欢迎的数据科学竞赛网站，首页也介绍到在这里你可以找到需要的所有数据与代码去完成你的数据科学工作，截至2020年10月5日，这里已经有了50000份以上的数据集以及超过400000的公开代码笔记，本书将着重介绍其竞赛单元。点进竞赛单元Competitions，这里从上到下罗列着历史上所有进行过的比赛，最上面部分永远有着正在进行的各种比赛，随便点开一个比赛，同样可以看见比赛的相关信息，大致有概况Overview、数据Data、代码笔记Notebooks、社区讨论Discussion、排行榜Leaderboard、比赛规则Rules等，接下来将以竞赛Microsoft Malware Prediction为例介绍一场竞赛的主要内容。

##### Overview

Overview即概况，这是了解一场竞赛的开始，包括有四个部分Description、Evaluation、Prizes和Timeline。Description描述是对该竞赛背景的介绍以及主办方信息，竞赛Microsoft Malware Prediction中写到：

恶意软件行业仍然是一个组织良好、资金充足的市场，致力于规避传统的安全措施。一旦电脑受到恶意软件的感染，犯罪分子会在很多方面伤害消费者和企业。微软拥有超过10亿的企业和消费者客户，非常重视这个问题，并投入了大量资金来提高安全性。作为其整体策略的一部分，微软正在挑战数据科学界开发技术，以预测一台机器是否很快会受到恶意软件的攻击。与之前的恶意软件挑战（Malware Challenge，2015）一样，微软正在向Kaggler提供一个前所未有的恶意软件数据集，以鼓励在预测恶意软件发生的有效技术方面取得开源进展。你能帮助保护超过10亿台机器不受损害吗？

简短的一段描述清楚了竞赛的背景和需要解决的问题，可以让参赛者快速了解业务。紧接着就是Evaluation评分，这里会列出本次竞赛的评判标准以及提交文件的格式，竞赛Microsoft Malware Prediction采用的是预测概率与真实标签的ROC曲线下面积即AUC作为模型得分，因此本次竞赛是一个二分类问题。

接下来就是激动人心的Prizes奖项部分，这里展示了本次竞赛总奖金25000美金，其中冠军奖励12000美金几乎占去了一半，通常来说25000美金是Kaggle有奖竞赛里面比较常见的金额，多的能有100000美金。需要注意的是，这个比赛对获奖者会有一定地要求，竞赛结束后需要在规定时间内提交建模方案文档，且不允许微软内部员工参赛。以上这些也基本是大多数比赛都会要求的。

##### **Data**

了解了竞赛的背景与任务之后，参赛者就可以开始熟悉数据了，通常的数据格式都会是csv宽表形式。数据部分也有个单独的Data Description，这里通常会给出所有表格的数据信息，包括采集来源，任务说明，以及详细地各个字段含义等。以竞赛Microsoft Malware Prediction为例，其Data Description如下：

这项竞赛的目标是根据一台Windows机器的不同特性，预测它被各种恶意软件家族感染的概率。包含这些属性和机器感染的遥测数据是通过结合微软的端点保护解决方案WindowsDefender收集的心跳和威胁报告生成的。此数据集中的每一行都对应于一台机器，由MachineIdentifier唯一标识。HasDetections则是机器的标签，表明在机器上是否有检测到恶意软件。参赛者需要利用train.csv中的信息和标签训练集，预测test.csv中每台计算机的HasDetections的值。用于创建此数据集的采样方法旨在满足某些业务限制，包括用户隐私以及机器运行的时间段。恶意软件检测本质上是一个时间序列的问题，但由于新机器的引入，在线和离线的机器，接收补丁的机器，接收新操作系统的机器等等，它变得更加复杂。而这里提供的数据集已大致按时间划分，以上提到的复杂性和抽样要求可能意味着你可能会看到你的交叉验证、公开和私人榜单分数之间的不完全一致！此外，这个数据集并不能代表微软客户的机器，它已经被抽样包含了更大比例的恶意软件机器。

参赛者在参加竞赛时首先应该做到的就是熟悉题目与数据，往往这里面会包含很多重要细节信息，拿上面的题目为例，看上去数据非常简单主办方已经分清楚了训练集与测试集，标准的特征字段与标签字段，赛题任务也很明确就是预测机器是否会被恶意软件感染，但其中不可忽视的是介绍中提到的本质是一个时间序列的问题，训练集与测试集也是大致按照时间进行了划分，而为了突出恶意软件机器更是对正样本进行了一定的升采样，因此这种复杂性与抽样性会给建模带来极大的不确定性，导致其交叉验证、公开和私人榜单分数的上下波动不会完全一致，这一点在竞赛结束后的私人榜单上就有所体现，相比公开榜单，其排名波动异常剧烈。

##### **Notebooks**

Notebooks部分则是本次比赛的开源社区所在，Kaggle能作为全世界最大的数据竞赛平台，其讨论与开源的氛围功不可没，在这里你可以看到各式各样的数据探索（EDA）、特征工程、建模方法以及截然不同的代码风格与个人偏好，有的代码标题下方甚至会显示本代码在榜单上的得分成绩。在这里参赛者可以尽情学习各种工具和代码写法，为了达成同样的目的，也会参赛者会在这里发现更为简洁、优雅、快速的实现方式，同时可以将各种建模方法进行融合博采众长，甚至竞赛圈里流传着只要开源融合得好，得一块银牌不是问题。

##### **Discussion**

和承担代码笔记功能的Notebooks不同，Discussion是真正的参赛者之间交流讨论的地区，这里少有代码，有各种QA以及对赛事的理解发现。参赛者可以在这里自由地和全世界的数据科学共同爱好者讨论竞赛相关心得，甚至是理论上的探索验证，在这里可以见到各种Master乃至Grand Master的身影，他们之间的互动也是十分精彩。

##### **Leaderboard**

Leaderbord是排行榜的展示，所有成功提交过结果文件的参赛者都能在这里找到自己的位置，榜单实时刷新，对于争分夺秒挠破头皮的参赛者来说可以说是十分刺激。Kaggle的比赛通常也会分为Public Leaderboard和Private Leaderboard，即竞赛圈子里常说的A榜与B榜。这展现了机器学习领域是非常重要的一个概念，那就是模型泛化性。实时榜单的存在虽然方便了参赛者不停地验证自己的想法，可以对比出不同方案的得分，但这只是在公开榜上的成绩，机器学习建模很重要的一点就是模型的泛化性能，也可以说是它的鲁棒性，泛化性能好鲁棒性强的模型才可以在未来预测中始终保持着一个良好的效果，这对实际应用来说非常重要，因此才有了A榜B榜的划分，通常A、B榜为同一批数据切分成的两份。通常来说，参赛者需要在比赛的第一阶段不断根据A榜的得分来修正改善建模方案，而最终可以有两次机会选择最终用于计算B榜得分的结果文件，最终的排名依据B榜得分来定。机器学习建模的泛化性是一个比较难的痛点，有些比赛A榜、B榜的排名可以说是翻天覆地地变化，过多的提交、对比A榜的得分往往会使参赛者的模型过拟合A榜，也就是在A榜外的数据模型表现会十分差劲，这也是大家有时会将机器学习深度学习等人工智能建模称为丹炉炼丹或者玄学的原因所在。

##### **Rules**

这里给出了本次竞赛的相关规则，是比Overview部分更加详细地补充，通常需要关注的有几个重要的Timeline，如比赛开始的A榜评测开放时间、队伍合并的截止时间以及B榜的切换时间。此外还有队伍人数及队伍提交总次数的限制，关于竞赛作弊方式的判定以及其他不允许发生的行为等。建议参赛者不不仅要熟悉竞赛背景、竞赛内容，更要对竞赛规则了如指掌，以免不小心违反规则导致辛苦努力白费。

##### **Team**

顾名思义这里就是竞赛队伍的信息窗口，参赛者要加入或者邀请某个人加入自己队伍都在这里完成，众人拾柴火焰高，团队在竞赛中的竞争力通常来说比个人要厉害很多，当然如何根据个人的长短进行团队的搭建也是一门学问。对于一般水平的参赛者来说，组建队伍可以互相交流，取长补短，而由于起点低队伍里每个人的建模方法区别很大的话在最后阶段的模型融合往往能带来很大收益。

##### My Submissions

提交结果文件的地方，使用谷歌浏览器的参赛者需要安装谷歌访问助手才能顺利提交，注意提交格式要与竞赛要求一致，通常都会有提交样例可供参考，成功提交后等待片刻便会计算出此次提交得分并跳转到排名界面，若是比当前参赛者最好成绩还要高，则可以看到参赛者自身的排名蹭蹭往上飞升，此情此景妙不可言。

##### Grand Master

Kaggle竞赛平台还有一项为人所津津乐道的设置便是其荣誉称号制度，含金量高的便是Master与GrandMaster，激励着参赛者一次次去挑战不可能。截至2020年10月5日，全球GrandMaster仅194位，而所有参赛者数量为147906，是名副其实的千里挑一，在这个圈子流传着许多大神传说，等待着一批又一批的参赛者去挖掘故事也去缔造新的传奇。

#### 1.1.2 天池

##### 简介

天池全称为天池大数据众智平台，也叫阿里云天池，是阿里云旗下大数据平台。天池二字，想必和阿里的花名文化同出一脉，颇有古风，武侠小说中多有涉及，现今长白山也有一个神秘的天池，地处中朝边境常年冰雪笼罩。天池竞赛平台和Kaggle也有些类似，有许多开放的数据集、数据挖掘竞赛和AI学习社区等。除了本书涉及到的机器学习算法竞赛，也还有创新应用大赛、程序设计大赛等，奖励也是非常地丰厚。

##### 注册

和大多数的竞赛网站一样，为了防止建立小号等作弊行为保证比赛的公平公正，天池等平台都需要通过邮箱或者手机号注册，并且需要上传个人证件进行实名认证。

##### 赛制

天池的竞赛同样由赛题介绍、开源社区等板块，通常也都设置有AB榜，和Kaggle不同的是，天池通常会设置初赛和复赛，各自都有相应的A、B榜，天池的B榜通常是换数据测验，并且会持续几天，相比于A榜只是时间上有所缩短，而Kaggle的测试数据是预先全部给出，只是在评分的时候只计算A榜部分，最后选定两个作为B榜计算的结果文件。此外，天池的比赛通常会以12点与晚上10点这种类似的方式允许参赛者一天有定点的计算新结果文件得分的机会，同时也会刷新榜单，而不是像Kaggle那样参赛者可以随时提交即时得到评分，Kaggle通常也一天限制参赛者的提交次数。提交次数自然是十分宝贵的，限制次数一方面是为了缩短参赛者自身的资源配置之间的差距，防止有些参赛者以强大的计算资源获取不当优势，同时也是为了避免参赛者过多地依赖测试结果进行建模，导致陷入过拟合的泥沼，使得模型的泛化性较弱白做许多的无用功。

##### 积分

天池设计有积分规则，根据积分或者条件为参赛者设计了五个等级称号，从低到高分别是数据新手、数据极客、数据大神、数据科学家以及数据大师，天池会显示出排名前百的天池科学家，这也是天池的一个特别做法，各个竞赛也基本只显示出排行榜前百名。

#### 1.1.4 DF

DataFountain是CCF即中国计算机学会的指定专业大数据及人工智能竞赛平台，与学界联系紧密，从技术上的数据挖掘、自然语言处理和计算机视觉以及分行业上对竞赛进行区分，将学界与工业界紧密联系在一起，虽然奖金金额可能比不上大平台，但其对行业细分理解以及落地场景的多样化非常诱人。

#### 1.1.3 DC

DC竞赛平台全名DataCastle，即数据城堡，是一家坐落于成都的公司。其网站架构和竞赛举办方式与Kaggle、天池相似，独特的一点在于其专门设有政企办赛的部分，通常参赛者可以在DC平台上看见许多政府以及国企央企扶持的相关创业竞赛项目，除了本书所专注的算法竞赛之外，还有创意赛等。

#### 1.1.5 Kesci

Kesci中文名为和鲸社区，是每年的中国高校计算机大赛-大数据挑战赛的战略合作平台，相比于DC与DF平台，它还能够提供在线的notebook训练环境，这对于一些没有足够硬件资源的参赛者来说比较友好。

#### 1.1.6 JDATA

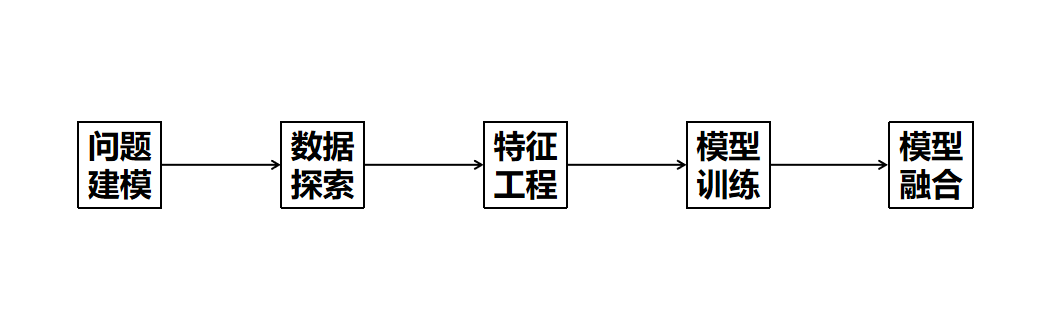
JDATA智汇平台是京东旗下的竞赛平台，板块设置大多也和天池、Kaggle类似，细节方面有所不同，每年的春季是京东自家出赛题的高峰时段，有意思的是，京东的竞赛主要涉及到电商以及物流，它们通常会自定义一些评价指标，参赛者拿到宽表数据之后需要自行考虑建模方案，其中包括训练集与测试集的搭建、样本标签的选取等等，数据质量和赛题难度都极高，当然，奖金也是不少，且还有获得校招绿色通道的机会。

#### 1.1.7 企业网站

除了上述列举的国内外主流竞赛平台，有些企业自家举办竞赛不与平台合作而是自己单独建立一个简易网站，如腾讯社交广告算法大赛，虽然只是有个网站，但竞赛依然十分火爆，此外还有FlyAI、AIChallenger等，参赛者不必面面俱到地了解，可以通过关注一些公众号来了解赛事的最新相关信息，如机器学习助手、麻婆豆腐AI、Coggle竞赛助手等。

### 1.2 竞赛流程

想要成功完成一次竞赛一共需要几个步骤呢？答案是三个，首先下载数据，其次用代码跑出结果，最后进行提交。当然，这只是仿照把大象关进冰箱需要几步罗织的一个笑话，搏君一笑。机器学习算法竞赛也逃脱不了所谓的套路，本书的作者们通过大量的实战经验总结后将整个流程大致分为五个部分，即问题建模、数据探索、特征工程、模型训练、模型融合。当然赛前还有一些准备工作要做，比如注册账号、完善个人信息甚至是实名认证，然后点击想参加的竞赛进行报名，就可以开始接下来的竞赛流程了，本节只简单介绍各个竞赛流程，详细内容会安排在本书第一部分即磨刀事半砍柴功倍的第二章到第六章。



#### 1.2.1 问题建模

相信大家都还记得高考前期，老师们耳提面命地强调审题的重要性，理解题目永远是最先也是最重要的一步。准确理解题目想要表达的意思，能够避免走许多弯路。在机器学习问题建模中，不是所有的数据都是特征加标签这种已经可以直接加入模型训练的形式，很多时候需要从数据当中去进行分析进而抽象出建模目标与方案。虽然通常来说竞赛的目标明确，但也不是所有竞赛的数据是那种可以直接加入训练的格式。有些竞赛如JDATA智汇平台就常常会有一些不同于一般分类和回归评价指标的评估方式，参赛者往往需要根据赛题的理解自行利用主办方提供的数据构造训练集与测试集，这种竞赛极大地考验参赛者的问题建模水平，这也是这类竞赛的难点所在，问题建模方式的选取很大程度影响到了参赛者的成绩好坏。

#### 1.2.2 数据探索

数据探索是机器学习领域最重要的一个概念之一，习惯上被大家称为EDA，即Exploratory Data Analysis的简写，也就是针对数据进行的探索性分析。在理解赛题，大致知道了问题的建模方式后，这时参赛就需要结合赛题背景业务理解去看看数据长什么样子，是否和描述相符、数据包含有哪些信息、数据质量如何等等。首先要对数据有一个清晰地认知，主要是宽表的各个字段取值含义、范围和数据结构等。更深层次的是，结合标签分析特征的分布状态、训练集与测试集的同分布情况、特征之间的业务关联以及隐含信息表征等。总地来说，数据探索是承上启下的一步，帮助参赛者更好地理解问题建模和为接下来将进行的特征工程做好准备。

#### 1.2.3 特征工程

同数据探索性分析一样，特征工程也是机器学习领域重要的一个概念，翻译为Feature Engineering，由它的命名就可以看出这是一项可以被称为工程的模块。机器学习泰斗吴恩达老师在他著名的斯坦福大学CS229机器学习课程上曾经说过，机器学习大多数时候是在进行特征工程，特征决定了的机器学习预测效果的上限，而算法只是不断地去逼近这个上限而已，由此可见特征工程的重要性，事实上无论是在竞赛还是在实际应用中，特征工程也都是花费时间最多的模块，会占去建模者的大部分精力。

#### 1.2.4 模型训练

根据问题建立好模型方案后，根据业务理解进行相关数据探索性分析，继而逐步完善特征工程，就可以得到标准的训练集与测试集结构，接下来就可以考虑如何进行模型训练了，在一般的机器学习算法竞赛上，参赛者大多比较偏爱GDBT类的树模型，当然这也是它们确实效果好，常使用的树模型主要有XGBoost和LightGBM，这两种模型都有Scikit-Learn的接口函数，非常方便使用，此外有时也需要用到的算法有LR、SVM和RF等，有时参赛者需要用到一些深度学习模型如DNN、CNN、RNN及它们的衍生模型，以及广告领域流行的FFM等。如果说之前的步骤花费的是参赛者本人的时间精力的话，这一步则主要依赖于参赛者的计算资源，当然如果不是特别大量级的数据，模型训练一般会很快。模型训练这个模块除了选择合适的模型之外，还有一部分需要花时间的就是参数调优，虽然只要参数不是设置地很离谱效果都差不太大，但在众多的参赛者当中即使是成绩的一点点提升也可能有排名的上升。

#### 1.2.5 模型融合

经过前期繁琐艰辛的各种尝试之后，终于可以来到喜闻乐见的模(xun)型(mi)融(dui)合(you)的阶段了。每一种算法都有其自身的优势和局限性，扬长避短综合各个算法的优势可以使得模型的效果更好。模型融合有许多种办法，诸如Stacking、加权投票等等，第六章模型融合中会详细介绍这部分内容。之所以模型融合也可以称作是寻觅队友，是因为在竞赛当中，参赛者之间的个人差异很大，涉及到问题理解建模、特征工程模型训练等几乎每个模块都会有差别，这就造就了参赛者之间方案的巨大差异，而差异带来的模型融合效果极佳，差异越大提升越大。在这里也建议参赛者如果不是有特别熟悉的队友可以先自己做，一个人走完全流程，这也是对自己的一种锻炼，做到后期实在没有想法了就可以考虑找成绩相近的参赛者进行组队，团队力量在竞赛当中的重要性不言而喻，而且后期组队相当于队伍前期各自进行提交，也是变相多了验证思路的机会，合理的利用规则在规则下进行竞赛也是被允许和提倡的。

### **1.3 竞赛类型**

眼花缭乱的竞赛令人迫不及待地想要跃跃欲试，门类众多的数据竞赛可以满足众多参赛者的不同需求，同时也是鼓励AI+行业的发展，让社会积极探索人工智能使能。因此有必要介绍一下当今常见的数据竞赛类型，下面将分别从数据类型、任务类型以及应用场景进行展开。

#### 1.3.1 数据类型

人工智能领域大致可以分为计算机视觉（CV）、自然语言处理（NLP）和传统宽表的数据挖掘。从数据类型角度就可以大致将三者进行简单区分，计算机视觉领域多是处理图像方面的数据，当然这其中也包括了视频。自然语言处理多是文本数据，涉及到各种语言的分词等等。二者都是近几年随着计算机硬件性能改善支撑着算力的提升，以及宽带网络的快速发展等得到了学术界以及工业界的共同关注。Kaggle上面的竞赛会在题目下方给出数据类型，如图片数据、音频数据、文本数据、宽表数据等，本书将着重介绍传统宽表数据类型的相关竞赛。传统宽表的数据中，通常匹配有样本的唯一id索引以及特征列，特征根据含义又可分为类别特征以及数值特征，类别特征如用户性别，数值特征则如年龄、身高体重等，上述这些特征的形式都是单值特征，此外还有多值特征，如用户的兴趣爱好这列可以同时包含健身、跑步和摄影等，针对这种特殊的多值特征有特别的处理技巧，这部分也将会在计算广告的实战部分为大家进行详细讲解。

#### 1.3.2 任务类型

机器学习相关竞赛以算法为主，偶尔也有方案创新设计赛等，本书将专注讲解机器学习算法相关竞赛，主要是有监督学习的相关内容，即通过已有的带标签的训练集数据根据任务要求进行建模，从而对测试集数据进行预测并给出相应标签的结果进行得分评价。按照问题类型大致可分为分类以及回归，在第二章问题建模中会具体给出相应任务的评价指标。

#### 1.3.3 应用场景

提起应用场景，自然而然想到的是机器学习在各个行业的应用，行业的需求和痛点，纵观各大竞赛平台，涉及到的主要有医学、制造业产品线、金融、电商、互联网等等。其中互联网行业用户数据的丰富性和多样性，以及较少遇到如医学方面的伦理道德等挑战，产生出了很多应用场景，如广告、搜索和推荐等，都是当今人工智能涉足较多的场景。

### 1.4 思考练习

1、请在Kaggle、天池、DC、DF、Kesci以及JDATA等网站分别注册账号，并浏览体会本章介绍内容。

2、完整的竞赛流程包括哪几个主要部分，每个部分在流程中的角色是怎么样的？

3、以日常生活接触到的场景为例，列举出5项可能使用到了机器学习算法的应用。