



章节列表

1. [介绍](#)
2. [机器学习是什么](#)
3. [成为机器学习工程师必备的5项专业技能](#)
4. [机器学习的最佳编程语言？](#)
5. [机器学习工程师面试宝典](#)
6. [企业如何通过 Kaggle 来寻找最优秀的机器学习人才](#)
7. [机器学习的未来](#)
8. [结论](#)



对机器学习感兴趣？并非只有你有这种想法！每天有越来越多的人对机器学习产生兴趣。实际上，最近很难找到其他领域能够像机器学习这样引起轰动了。机器学习如此的侵入我们的集体意识，这既是一种创造历史（例如当 AlphaGo 打败了全球的最佳围棋棋手）的现象，也是一种疯狂的发展趋势。无论你是如何发现机器学习这一领域的，有一点值得肯定的是：机器学习的时代到来了。

要如何加入这场21世纪前沿科技的浪潮，成为高薪抢手又能改变世界的机器学习工程师？

来自硅谷的《机器学习工程师终极职业指南》就可以帮你实现理想！在 Udacity，我们非常有幸能与世界上最具前瞻性的公司合作，并能够接触到机器学习领域最具创新思维的人士。我们的合作伙伴，如 Facebook, Google, Kaggle, 滴滴出行等，提供这一领域最吸引人的职位机会，并且是了解企业对机器学习人才有何期望的重要信息来源。

我们汇总了所有资料，并整理出了这份指南，我们的专业导师提供了独特的见解和经验。如果你想进入大数据、人工智能领域，成为行业抢手技术精英，那么花15分钟时间阅读这份指南，将能够帮助你事半功倍实现职业理想！



机器学习是一个真正独特的领域，因为它即复杂又简单。例如，你可以比较一下这两段描述：

“机器学习是计算机科学的一个分支，从人工智能中的模式识别和计算学习理论研究发展而来。

机器学习探索能够学习和预测数据之算法的研究与构建。”

“机器学习是让计算机无需明确编程便可采取行动的科学。”

第一个来自维基百科，第二个来自斯坦福大学线上机器学习课程的描述。完全是不同的风格，对吧？那么机器学习为何会如此既复杂又简单呢？原因就在于它无所不在。

那什么是机器学习呢？

它从哪里来，是什么意思，以及它为什么重要？从根本上来说，机器学习的目的是理解大量数据。注意，我们所说的“**大量**”是真正意义上的**大量**——数以百万计的可计算、量化和分析信息：上百万的患者、学生、交易、推文。仅就现代世界所产生的数据量来说，机器学习是非常必要及有可能实现的。

当然，统计和算法等领域一直以来就旨在总结数据，以助力决策和预测。而且，机器学习中使用的许多公式和技术是早在几个世纪前的数学家开发出来的，唯一一个新方面是数量。计算能力的提高使我们能够用几小时的时间，完成手动需要几个世纪的分析。

结果是：现在我们拥有比以前多十亿倍的数据，而计算它们的能力也同样提高了十亿倍。这一切是怎么发生的？答案就是机器学习！从字面上看，即机器从数据中“学习”概念。它的学习方式就和我们的日常学习一样：查看经验与以往观察并识别有用的信息。不同之处在意，人类往往从几十个经验中学习，而机器学习可以从数以百万计的经验中学习，而这些经验均使用严密的数值定义。

机器学习是什么的？

机器学习是一门如何用大量数据来解决问题的学科。物理学家们总是要找到事物的内在规律（例如牛顿三大定律），然后用实验去验证定律。但在机器学习中，定律是什么并不重要，定律可能太复杂，也可能不存在。例如我们观测到过去的每个黑夜之后都会有白天，每个白天之后都会有黑夜。虽然我们并不知道日心说或任何其他天体物理学，我们依旧可以断言，下个白天一定会到来。机器学习只需要收集大量大量的数据然后归纳出模型来进行预测，分类和决策。

机器学习算法可以对事物进行分类。例如根据我的种种淘宝浏览记录，算法会把我分类成“电子产品爱好者”，但根据另一个人的浏览记录，算法会输出分类“美妆爱好者”。如何将一个人准确的分类，从而给他显示最适合的广告，是一个很有商业价值的机器学习问题。无论facebook，淘宝，小红书都在不遗余力的改进这个算法。

机器学习也可以对某个数量进行预测。例如星期一下午5~6点，上海市徐家汇地铁站附近会有多少乘客用滴滴叫车？会有多少司机在3千米范围内响应？对这些数量的精准预测可以帮助提前调配空车，提高服务的可用性。不光滴滴，大型的互联网公司例如饿了么，摩拜单车都需要不断的改进他们的机器学习算法来进行精准预测，他们对高水平机器学习工程师永远属于渴求的状态。

机器学习还可以直接输出一个决策。例如在某种股市行情下，应该买进500股IBM。例如在某种围棋棋盘下，应该下三三位抢角。例如在某种路面情况下，应该向右变道准备下高速。例如在某些学习状态下，应该给学生更多的某个类型的题目进行强化训练。所有这些决策都是机器学习在多种可选决策中挑选的最优决策。而决策的优劣也正是在一次次决策后的反馈中总结而来的。大型的基金公司，汽车公司，教育公司都需要大量的机器学习人才来改进产品，降低成本，否则只能慢慢走向灭亡之路。

机器学习工程师的实际工作是什么？

很简单！机器学习工程师建立程序，用来动态执行以前数据科学家手动执行的分析。那么它的重要性何在？花一点时间思考一下数据发挥重要作用的领域：医疗、教育、天文学、金融、机器人学等。机器学习已经对所有这些行业产生了影响，事实上，几乎没有哪个领域不受机器学习的影响！

这是为什么机器学习如此让人着迷的一个关键原因，因为它无处不在。很多时候，我们甚至意识不到它在运行。你使用过 Google 翻译吗？Siri 呢？你的 Facebook 信息流呢？正是机器学习使它们全部变为可能！如果你对 Udacity 有所了解，你应该知道我们的创始人兼总裁 Sebastian Thrun 本人在这个领域拥有较长时间且非凡的历史——他在卡耐基梅隆大学创立了一个硕士课程，后来演变成了机器学习博士课程；他是斯坦福大学人工智能实验室的主管；还是 Google 无人驾驶汽车开发的领导者。

事实上，Google 翻译可能是最著名的（和利用率最高！）的机器学习实践例子，而 Google 对其工作原理的描述非常经典地阐释了这个概念的应用：

机器翻译是 Google 将前沿研究和世界级基础设施相结合的一个很好例子。我们专注于研究能够借助获得更多数据加以改进，并很好地推广到新语言的统计翻译技术。凭借大规模计算基础设施，我们能够快速实验在网络规模的数据上培训的新模式，从而显著提高翻译质量。

这里最重要的一句话是“借助获得更多数据加以改进”，这正是机器学习的本质所在。

在 2006 年，Tom Mitchell 发表了机器学习中的规则 (*The Discipline of Machine Learning*)。在这篇文章中，他提出这样一个问题：

“我们如何构建能够根据经验自动改进的计算机系统？

机器学习便是这个问题的答案，这也是为什么 **Udacity 联合 Google、Kaggle、滴滴出行 推出了“机器学习工程师”纳米学位项目**，旨在让你在家就能学习全球最高水平的机器学习课程，零基础入门，掌握如何用机器学习在大数据、金融、人工智能领域进行预测分析，成为被 Google 认证的工程师！



成为机器学习工程师 必备的5项专业技能

AI的浪潮已经来临，人工智能的市场在不断扩大。从 Google、Facebook、亚马逊到 BAT，以及国内新兴的科技企业，都在极力投入人工智能领域的产品或研究，而机器学习作为人工智能最前沿的技术之一，越来越多的企业开始加大对它的运用。

根据《人民日报》报道，中国目前需要500万人工智能开发人才，但现在在 LinkedIn 上有资料的仅有5万。在招聘网站上，你也可以搜索到很多顶尖企业在招聘相关人才，看到诱人的薪酬待遇。

但是，对机器学习感兴趣是一回事，真正地开始在这一领域寻求职业发展机会则是另一回事。**这一章将帮助你了解成为机器学习工程师需要具备的总体思维和特定技能。**

首先，如果你想成为机器学习工程师，则要了解两个非常重要的事项。第一，它并非一个“纯粹”的学术岗位。你不一定要具备研究或学术背景。其次，仅具有软件工程经验或数据科学经验并不足够，最好能同时具备这两种经验。

数据分析师与机器学习工程师

了解数据分析师和机器学习工程师之间的区别也很关键。在最简单的情况下，关键区别在于最终目标。

作为数据分析师，你要分析数据以便讲述故事，并生成可行动的见解。重点在于信息传播：图表、模型和可视化。由人类来进行分析，并向其他人展示分析结果，后者然后可能会根据所展示的信息作出业务决策。一定要注意的是，输出的“受众”是人类。

但是，作为机器学习工程师，你的任务主要是通过数据训练算法来完成分类、预测或者输出决策的任务。你的最终“输出”是可以运转的软件（并不是你可能会随之创建的分析或可视化内容），该输出的“受众”经常包括能够自主运行、几乎不需要人类监督的其他软件组件。产生的信息依然要能采取行动，但是对于机器学习模型，由机器做出决策，它们会影响产品或服务的行为。因此对于机器学习这一职业领域来说，软件工程技能非常重要。

理解生态系统

在学习具体技能之前，还需要了解一个概念。要成为机器学习工程师，则有必要了解你要处理的整个生态系统。

假设你在连锁杂货店工作，公司想根据消费者的购买历史记录发放购物券，并生成购物者实际上会使用的购物券。在数据分析模型中，你可以收集购物数据，分析趋势，然后提议策略。机器学习方法则写出自动的购物券生成系统。但是它会如何写出该系统，并使其能正常运转呢？你需要了解整个生态系统：库存、目录、定价、购物单、账单生成、销售点软件、CRM 软件等。

最终，这一流程重点不再是了解机器学习算法，或者何时及如何应用这些算法，而是侧重于了解系统性关系，并编写将成功地集成和形成接口的软件。注意，机器学习的输出实际上是能运转的软件！

现在我们详细了解下成为机器学习工程师需要具备的技能。我们将分成两大部分进行介绍：技能总结以及语言和库。在这一章中，我们先从技能总结开始，在接下来的内容中，我们将讨论机器学习的语言和库。

技能总结

1. 计算机科学基础知识和编程

机器学习工程师务必掌握的计算机科学基础知识包括数据结构（堆栈、队列、多维数组、树、图表等）、算法（搜索、排序、优化、动态编程等）、计算能力和复杂度（P 与 NP、NP 完全问题、大 O 记法、近似算法等），以及计算机架构（内存、缓存、带宽、死锁、分布式处理等）。

在编程时，你必须能够（相应地）应用、实施、调整或处理这些知识。练习题、编程竞赛和黑客马拉松是磨炼技能的好方法。

2. 概率与统计学

概率的正式特性（条件概率、贝叶斯规则、似然、独立性等）和推出的技巧（贝叶斯网络、马尔可夫决策过程、隐藏式马可夫模型等）是很多机器学习算法的核心；我们可以通过多种方式来解决现实生活中的不确定性。与此领域关系紧密的是统计学，统计学提供了构建和验证通过观察的数据得出的模型所需的各种方法（均值、中位数、方差等）、分布（均匀分布、正态分布、二项分布、泊松分布等），以及分析方法（ANOVA、假设检验等）。

3. 数据建模和评估

数据建模是估算给定数据集的底层结构的流程，旨在找到有用的规律（相互关系、集群、特征向量等），和/或预测之前未知实例的属性（分类、回归、异常检测等）。该估算过程的关键部分是继续评估给定模型有多好。根据手头上的任务，你需要选择相应的准确性/错误衡量方法（即分类模型的对数损失、回归模型的误差平方和等）。迭代学习算法经常会指定利用广义误差来调整模型（例如神经网络的反向传播），因此了解这些衡量措施很重要，即使仅仅应用标准算法也是如此。

4. 应用机器学习算法和库

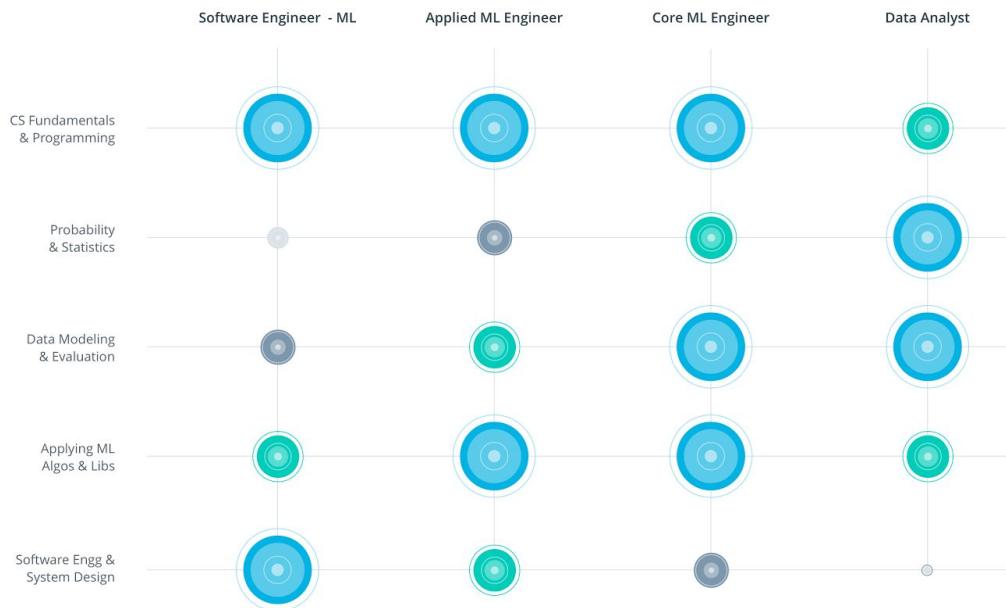
机器学习算法的标准实现可以广泛地通过库/包/API 获得（例如 scikit-learn、Theano、Spark MLlib、H2O、TensorFlow 等），但是有效的应用它们则需要选择合适的模型（决策树、近邻取样、神经网络、支持向量机、多个模型的组合，等）、拟合数据的学习流程（线性回归、梯度下降、基因算法、装袋、推进和其他特定模型方法），以及了解超参数对学习有何影响。你还需要了解不同方法的相对优势和不足，以及会带来阻碍的各种陷阱（偏差和方差、过拟合和欠拟合、缺少数据、数据泄露等）。数据科学和机器学习挑战（例如 [Kaggle](#) 上的挑战）是了解不同类型的问题和细微差别的很好的方式。

5. 软件工程和系统设计

最终，机器学习工程师的典型交付内容是软件。经常是纳入更大型的产品和服务生态系统的一个小组件。你需要了解这些不同组件是如何协同工作的，与它们通信（使用库调用、其他 API、数据查询等），并为你的组件构建其他组件将依赖的相应接口。可能有必要仔细设计系统，以避免瓶颈问题并使算法能够在数据量增多时灵活扩展。软件工程最佳做法（包括需求分析、系统设计、模块化、版本控制、测试、文档等）对生产力、合作性、质量和可维护性而言异常宝贵。

机器学习工作岗位

与机器学习相关的工作机会正在飞速增多，企业正在努力充分利用这一新兴技术。下图描绘了这些一般职责类型的核心技能的相对重要性，并与典型的数据分析师职位进行对比。



不同机器学习岗位的核心技能的相对重要性



在上一章中，我们了解了在这一领域走向成功需要掌握的5项关键技能。现在，我们将回答对机器学习感兴趣的学员最常提出的问题：我需要了解哪门编程语言？

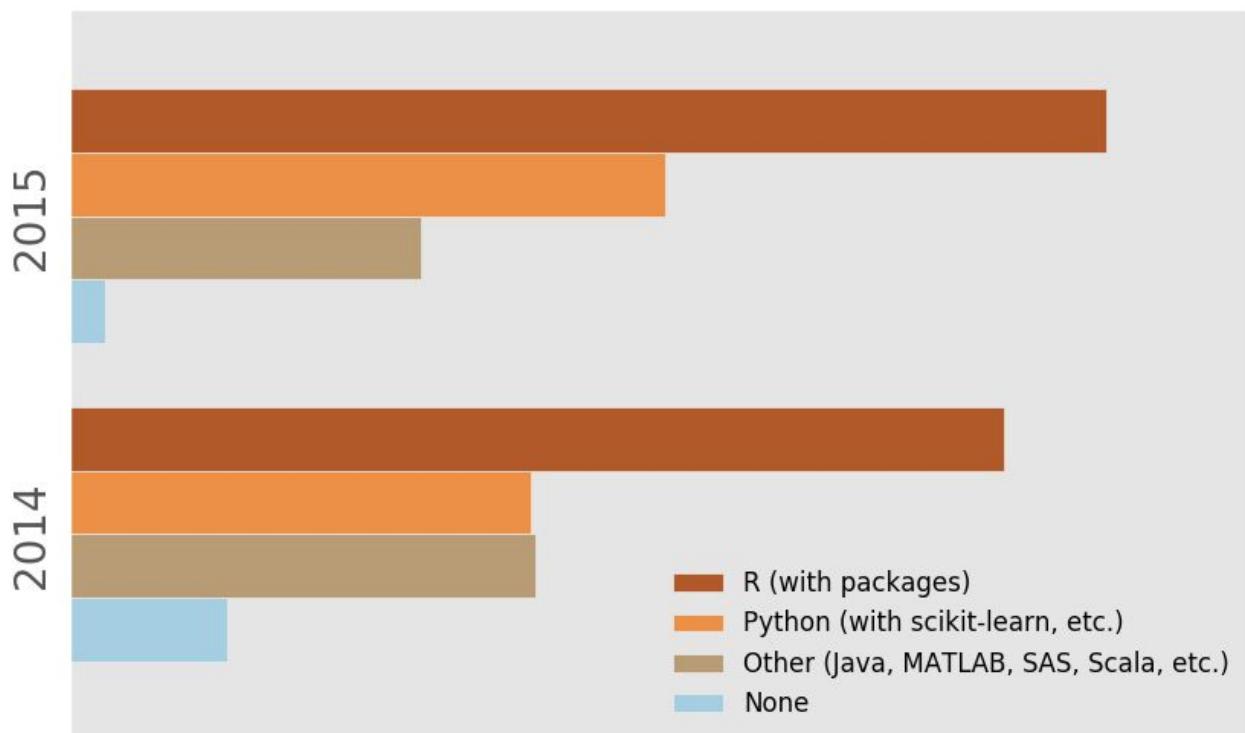
答案可能会令你吃惊。**语言真的不重要！**

只要你熟悉所选语言提供的机器学习库和工具就可以了，语言本身不重要。不同的编程语言提供了各种不同的机器学习库。只不过根据你在公司内的职责和你要处理的任务，某些语言、库和工具可能比其他的更有效。

R 语言

R 语言是一种专门用于统计学计算的语言，非常适合大规模数据挖掘、可视化和生成报告。你可以轻松访问大量的软件包（通过[CRAN 资源库](#)），使你能够应用几乎所有种类的机器学习算法、统计学测试和分析流程。该语言本身使用规整、但有些深奥的语法来表达关系、转换数据和进行并行操作。

KDNuggets 最近开展的一项问卷调查发现在 2015 年，R 语言是用于分析、挖掘和数据科学任务的最热门语言，不过，Python 在过去几年越来越流行。



KDNuggets 2015 年的问卷调查：数据挖掘、数据科学任务的主要编程语言

MATLAB

MATLAB 在学术领域很流行，因为它能够运算复杂的数学表达式，充分支持代数和微积分、符号计算，并且具有大量适用于多种学科（例如数字信号处理和计算生物学）的工具箱。它经常用于为机器学习算法进行原型设计，在某些情况下，还可以生成完整的解决方案。对于商业用途，依然需要高昂的许可费用，但是用途也很大，因为它能够显著减少研究和开发工作。Octave 是 MATLAB 的免费替代版，语法几乎一样，但是只有有限数量的工具箱，IDE 成熟度也更低。

Python

虽然 Python 是更加通用的编程和脚本语言，但是它在数据科学家和机器学习工程师人群中越来越流行。与 R 语言或 MATLAB 不同，数据处理和科学计算习语并没有集成到该语言本身当中，但是 NumPy、SciPy 和 Pandas 等库提供了对等的功能，语法却更容易懂。

[scikit-learn](#)、[Theano](#) 和 [TensorFlow](#) 等专门的机器学习库使你能够训练各种机器学习模型，可能会用到分布式计算基础设施。这些库的大部分关键性能代码依然通常用 C/C++ 甚至 Fortran 编写而成，Python 包充当封装器或 API（很多 R 包也是这样）。

但是最大的优势在于 Python 生态系统使我们能够轻松地构建复杂的端到端产品或服务，例如使用 Django 或 Flask 的网络应用，或使用 PyQt 的桌面应用，甚至是使用 ROS 的机器人代理。

这种多功能性正是我们在机器学习工程师纳米学位课程中主要使用 Python 的原因！

Java

Java 是软件工程师的首选语言，因为它能清晰一致地实现面向对象的编程，并且通过使用 JVM 而独立于平台。它为了清晰与可靠性而牺牲了简洁、灵活性，从而适合实现关键的企业级软件系统。为了保持这一级别的可靠性并避免编写杂乱的界面，一直使用 Java 的企业在满足机器学习需求时或许最好能继续使用 Java。

除了可用于分析和原型设计的库和工具（例如 [Weka](#)）之外，使用 Java 编写大规模分布式学习系统也有很多很棒的选择，例如 [Spark+MLlib](#)、[Mahout](#)、[H2O](#) 和 [Deeplearning4j](#)。这些库/框架非常适合符号行业标准的数据处理和存储系统，例如 Hadoop/HDFS，使得它们更易于集成。

C/C++

C/C++ 非常适合低级别的软件，例如操作系统组件和网络协议，其中计算速度和内存效率至关重要。出于相同的原因，还适合实现机器学习程序的关键部分。但是，它缺少惯用的数据处理抽象方法，并给内存管理增加了开销，因此不适合初学者使用，并且给开发完整的端到端系统带来了负担。

对于嵌入式系统（例如智能汽车、设备和传感器），或许有必要使用 C/C++。在其他情形下，由于现有的基础设施和应用特定的代码，可能只是为了便利才使用 C/C++。无论是何种情况，C/C++ 都不缺乏机器学习库，例如 [LibSVM](#)、[Shark](#) 和 [mlpack](#)。

企业解决方案

除了这些语言和库之外，还有好几个用于建模和商务分析的其他商业产品，它们会在更加具有管理性的处理环境中应用机器学习模型。这些产品（包括 [RapidMiner](#)、[IBM SPSS](#)、[SAS+JMP](#) 和 [Stata](#)）旨在为数据分析提供可靠的端到端解决方案，并且经常会提供可编程的 API 和/或脚本语法。

该领域最近的一项发展状况是出现各种基于云端的机器学习服务平台，例如 [Amazon Machine Learning](#)、[Google Prediction](#)、[DataRobot](#)、[IBM Watson](#) 和 [Microsoft Azure Machine Learning](#)。这些平台可以帮助你扩大你的学习解决方案，以便处理大量的数据并快速实验不同的模型。只要你掌握了扎实的机器学习知识，那么使用新的产品或平台就像学习使用一款新的工具。

*专业提示：在选择语言/库时，要考虑的一个重要事项是执行时间和开发时间之间的平衡性。一个速度极快的学习通道能够在几分钟内处理完数据，但是需要几个月的开发时间，那么就毫无用处。一定要能够快速构建和测试原型，因为第一次尝试肯定会上当。

因此大多数企业都在物色能熟练使用自己所选的工具/语言/库的机器学习工程师。用 Python 或 R 语言等高级语言设计算法原型，然后根据需要将解决方案移植到 Java 或 C/C++ 以进行生产是很常见的做法。

更多参考资料

1. [机器学习的最佳编程语言](#) – Jason Brownlee, Machine Learning Mastery
2. [Kaggler 最喜爱的工具](#) – Ben Hamner, Kaggle (详细论坛帖子)
3. [Python、机器学习和语言战争](#) – Sebastian Raschka, 密歇根州立大学
4. [用于分析、数据挖掘、数据科学任务的主要编程语言](#) – Gregory Piatetsky, KD Nuggets



机器学习工程师 面试宝典

对准备工作面试的比喻多种多样，譬如准备上战场、鼓足勇气向异性邀约。最佳情况是精神紧张，而最坏的情况则是感到恐惧。做好机器学习面试准备也不例外。你知道要么会非常出色，要么就很糟糕。但是如何确保效果很棒呢？一切在于思维能力和准备工作。

公司和职位

了解即将参加的面试的背景，即为何存在这一空缺职位将是准备工作必不可少的一部分。了解为何接受面试有助于了解你对公司的价值。例如，如果公司想要寻找一名机器学习工程师，那么很清楚的是他们正在尝试解决很难应用传统算法或者根本行不通的复杂问题。同时还清楚他们非常想要解决该问题。

发现核心问题

在申请此类岗位时首先要做的是想象下你处在该岗位的情形。为此，你需要尽量搜索关于该公司和职位的信息。要准备调查研究，问问自己：我可以为该公司解决一个什么样的核心问题？回答该问题的过程应该能够让你很兴奋，并促使你进一步了解关于该问题的信息：现有的方法、该领域最近的发展状况，并得出一系列更加具体的挑战。如果你知道你要加入的团队，那么选择相应的问题可能很轻松；否则，选择对公司来说很关键的问题。换句话说，思考一下公司面临的挑战，然后尝试确定他们可能会问的问题。

探索潜在数据来源

准备过程的下一步应该是思考要回答这些问题，你需要哪些数据。有些也许能够立即获得，但是你可能需要构建其他关联性来收集某些信息。深入探索公司的基础设施和运营状况，他们运转所用的堆栈是什么，他们具有什么样的 API，他们已经在收集什么样的数据，等等。如今的大部分企业都拥有一个博客，他们会在博客上讨论他们的挑战、方法、成功和失败情况。这些博客可以帮助你深入了解他们的运转情况，他们可能拥有什么样的产品和服务。

准备讨论机器学习解决方案

现在你需要作出非常大的概念性跳跃：机器学习如何适应所有这一切？根据你要实现的目标，以及你认为可能可用的数据，能够变成学习问题吗？可以使用什么样的合适模型？举个例子，很多推荐系统（例如 Netflix 和 Amazon）遇到的主要挑战是分类归并，而不是预测，即一旦你能够整理似乎具有相似的偏好和行为的用户群，就很容易推荐他们可能觉得实用的产品。

这种思维流程将帮助你讨论对公司来说最关键的问题。没人会指望你一来参加面试，就能够针对他们数个月或数年来苦苦解决的问题给出完整的解决方案！但是大家都喜欢对他们的核心问题展现出真实的兴趣、激情和好奇心的应聘者。

面试者可能会向你提出更加技术性的问题，具体取决于你的面试官和面试阶段，但是你应该抓住一切机会表示你对公司和职位都思考研究过。当他们提出更加开放性的问题（例如“描述当你处理一个项目时遇到的技术性挑战，以及你是如何解决这一挑战的？”）时，试着给出和公司的兴趣保持一致的答案。

技术和示例问题

你可以在本篇《机器学习领域职业指南》的[第三章](#)了解成为一名机器学习工程师所必要的5项技能，简单总结如下：

1. 计算机科学基础知识和编程
2. 概率和统计学
3. 数据建模和评估

4. 应用机器学习算法和库
5. 软件工程和系统设计

而在这里，我们的重点是针对这几项技能，介绍在机器学习面试中你可能会遇到的相应问题，并向你提供一些应对方法。

计算机科学基础知识和编程

示例问题

- 如何查看链表具有循环？
- 给出二叉搜索树中的两个元素，找到它们的最低共同祖先。
- 编写一个对给定堆栈进行排序的函数。
- 任何基于比较的排序算法的时间复杂度是什么？可以证明吗？
- 如何在加权图中找到从一个节点到另一个节点的最短路径？如果某些权重是负值呢？
- 从给定字符串中找到所有回文子字符串。

*对于所有此类问题，你应该能够推理出你的方法的时间和空间复杂度（通常用大O记法），并尝试尽可能实现最低复杂度。

*广泛实践是熟悉不同类别的问题的唯一方式，这样才能快速得出有效的解决方案。[InterviewBit](#)、[LeetCode](#)、[Interview Cake](#)、[Pramp](#) 和 [interviewing.io](#) 等编程/面试准备平台在这方面非常有帮助。

概率和统计学

示例问题

- 某个人群中男人和女人的平均身高为 μ_W 和 μ_M 。整个人群的平均身高是多少？
- 最近的一项调查研究指出意大利有三分之一的汽车是法拉利，并且有一半的颜色是红色。如果你看到远处有一辆红色汽车正在向你开来，那么这辆车是法拉利的概率是多少？
- 你想寻找在你的网站上投放横幅广告的最佳位置。你可以让广告位大小（或厚度）小、中或大，并选择垂直顶部、中间或底部位置。至少需要多少总页面访问次数和广告点击次数，才能95% 确认其中一种位置比所有其他位置效果都要好？

*注意，有很多机器学习算法都有概率和统计学偏差。从概念上理解这些基础知识非常重要，但是与此同时，你必须能够将抽象的公式与现实数量联系起来。

数据建模和评估

示例问题

- 一位奶农想要弄明白影响到奶牛产量的因素。她一直在记录每日温度（通常是 30°C 到 40°C）、湿度 (60-90%)、消耗的粮食 (2000-2500 kgs)，以及产出的奶量 (500-1000 升)。
 - 你应该如何处理这些数据并建模，以便预测出每天的牛奶产量 (升)？
 - 这是什么样的机器学习问题？
- 贵公司正在构建一个面部表情编码系统，需要获得标准高清 1920×1080 像素摄像头拍摄的照片，并持续指出用户是否处于以下某种状态：没有情绪、开心、忧伤、生气或害怕。当用户的面部没有出现在相机框架中时，它应该指出特殊的状态：无。
 - 这属于什么类型的机器学习问题？
 - 如果每个像素由 3 个值构成（红色、绿色和蓝色通道），那么处理每个图片的原始输入数据复杂度是多少（维数）？有没有降低维数的方法？
 - 你会如何表示系统输出内容？解释下原因。
- 在过去一个世纪收集的天气数据得出了气温上升和下降的循环模式。你会如何对该数据（平均年度气温值序列）建模，以预测未来 5 年的平均气温？
- 你在一家在线新闻服务机构上班，工作职责是收集全球各地的文本报告，并将每个故事当做一篇文章，内容整合自不同的新闻媒体。你会如何设计此类系统？你会应用什么样的机器学习技巧？

应用机器学习算法和库

示例问题

- 我正在尝试将一个隐藏层神经网络拟合到给定的数据集，我发现迭代训练多次后，权重来回摆动（变化很大，经常在正值和负值之间摇摆）。我需要调整哪个参数，才能解决这一问题？
- 在训练支持向量机时，你会优化什么样的值？
- 套索回归使用 L1 范数系数作为惩罚项，而岭回归使用 L2 范数作为惩罚项。哪个正则化方法更有可能导致稀疏解，一个或多个系数刚好为 0？
- 在使用反向传播训练 10 个层级的神经网络时，我发现前 3 个层级的权重根本不变！后面几个层级 (4-6) 在变化，但是很慢。这是怎么回事，如何解决这一问题？
- 我发现了一些关于欧洲小麦种植区域的数据，包括年度降雨量 (R, 单位是英寸)、平均海拔 (A, 单位是米) 以及小麦产量 (O, 单位是千克/平方千米)。经过简单的分析和一些图表，我认为

输出与降雨量的平方及海拔对数有关： $O = \beta_0 + \beta_1 \times R^2 + \beta_2 \times \log_e(A)$, 我能使用线性回归将模型中的系数 (β) 拟合到数据吗？

*数据科学和机器学习挑战（例如 [Kaggle](#) 上的挑战）是了解不同类型的问题和细微差别的很好方式。尝试参加尽可能多的挑战，并应用不同的机器学习模型。

软件工程和系统设计

示例问题

- 你负责运行一个商务网站。当用户点击某个商品以打开详情页面时，你会根据商品特征以及用户的购买记录建议另外 5 个用户可能感兴趣的的商品，并在页面底部显示这些商品。对于这一行为，你会使用什么服务和数据库表格？假设可以使用这些服务及表格，请写一个查询或流程来获取要建议的 5 个商品。
- 为了衡量用户吸引力和视频热门程度，你需要从在线视频播放器（例如 YouTube）上收集什么数据？
- 一个简单的垃圾邮件检测系统的工作原理如下：它一次处理一封电子邮件，并输出其中每个唯一单词的出现次数（术语频率），然后将这些计数与之前见过的、是否标为垃圾邮件的电子邮件进行对比。为了扩大此系统的规模，以便处理大量电子邮件，你能设计一个在多台计算机上运行的映射降低机制吗？
- 你想生成一个实时可视化图表，展示用户目前正在查看和点击的网页部分，就像热图。为此，你需要在客户端和服务器端使用什么样的组件/服务/API？

结论

这一章中想要呈现的是机器学习面试体验的两个方面：背景方面和技术方面。如果有一点值得强调的话，那就是你应该避免侧重于后者，而忽略了前者。雄心勃勃的机器学习初学者很容易就沉浸在技术准备中，而很少会对面试提出“为何...”这种问题：为何有一个空缺职位，为何该公司要寻找机器学习人才（以及机器学习解决方案！），他们为何对你感兴趣？理解这些问题将使你需要解决的技术挑战变得有意义和背景信息，回答这些问题将使你脱颖而出，成为最适合推动公司向前发展的应聘者。



企业如何使用 Kaggle 来寻找最优秀的机器学习人才

机器学习呈指数级发展要归功于技术发展，因为有一批活跃的社区人士在推动机器学习领域的发展，包括专注于核心算法的研究人员，以及推动机器学习应用领域界限的实践者。还包括越来越多的具有非正常背景的机器学习爱好者，他们加入了讨论并带来多样化的经验和观点。

发现并吸引机器学习人才

寻找机器学习专业人士的企业与 Kaggle 等数据科学竞赛平台之间的共栖关系极大地影响了机器学习的发展速度。这一关系还改变了招聘市场。如今的企业面临着不断增大的创新压力，这样才能保持竞争力，他们正在寻求相对非传统的方法来发现和吸引新型人才，以便保持领先水平。市场对机器学习的需求极为庞大，促使企业打破常规地寻找人才。

处在特殊领域（例如机器学习）的传统界限之外的人士带来的贡献以前总是很稀少。以前，这一流程很长，并且很艰难，通常需要具备数学、统计学、计算机科学或相关领域的高级学位，以及多年的学术研究项目工作经验。但是这一切正在改变，Kaggle 等平台使任何人都能接触现实中的数据，并有机会解决很难的机器学习挑战。

Kaggle 竞赛

Kaggle 的模式是提供机器学习竞赛，使机器学习爱好者能够实验项目，并进一步磨炼他们的技能。从学习的角度来看，这样非常有意义，玩游戏和竞争因素增加了激励性。但是，Kaggle 取得成功的关键是这些竞赛依赖于由现实世界的企业提供的现实数据。

例如，Kaggle 目前举办的竞赛任务是识别超声图像中的神经结构。由 State Farm 发布的另一个问题是：计算机视觉能发现走神的驾驶员吗？ Facebook 想要你根据充满噪音的数据识别出正确的签到地点。作为参与者，能够处理这些现实世界的数据集和结构化问题非常宝贵；此外，一些竞赛还提供大笔的奖金。但是，企业为何会向 Kaggle 提供他们的数据呢？实际上远不止这些，他们经常还会赞助竞赛！

这需要一些解释，而 Kaggle 联合创始人兼首席技术官 Ben Hamner 是再合适不过的人选了。所以我们询问他，这些赞助企业有何目的。Ben 指出某些企业“遇到了极具挑战性的使用案例，他们希望吸引全球最顶尖的人才来解决这些问题，并以非常具有竞争性的环境来处理这些使用案例。”

机器学习众包

简而言之，他们在众包机器学习！听起来是不是很疯狂？毫无疑问，如果没有相应的培训和背景知识，肯定不能出色地完成这些任务吧？Ben 的团队发现了一个有趣的现象：“专业技能和竞赛获胜者之间实际上没有任何关联性。偶尔有些竞赛针对的是天文学，并且获胜者是天文学家，但是更多的情况是英语专业的学生赢得了生物信息学竞赛，或者物理学家赢得了大脑信号处理竞赛。”

很有趣，对不对？或许这也表明了机器学习专业技能由一组高度可迁移的核心技能组成。处理不同类型的问题可以帮助进一步发现和掌握这些技能。还可能会出现一种情况：之前没有接触过某个问题领域的参与者可能会有独特的见解，能够提出创新想法，并尝试其他人可能没有考虑过的技巧。“赢得竞赛的人士具有多样化的背景是唯一最令我惊讶的现象。”

Kaggle 竞争者与专业人士的对比效果

依然存在一个问题：Kaggle 最终只是让机器学习爱好者沉迷竞赛并大肆吹嘘的平台，还是具有任何现实价值？获胜的解决方案与前沿技术相比，结果如何？在 Ben 最喜爱的一个竞赛中，他们“拿到大约有 28,000 名中学生和高中生写的论文数据集...并提出开发自动对论文进行打分的系统这一挑战。”当顶尖的解决方案与两名教师的打分结果进行对比时，“机器的分数经常与其中一名老师的吻合，因为另一名老师和这位老师的也吻合。”

在另一项关于心脏诊断学问题的竞赛中，Ben 指出，“我们有一个具有约 500 张患者 MRI 图像（心跳视频）的数据集，我们要求参与者像心脏病科医师那样根据 MRI 图像预测出心脏的某些特性。这项任务需要花费受过训练的心脏病科医师大概 20 分钟的时间，而这是一个很浪费的过程，他们完全可以用来在临幊上进行更加宝贵的诊断操作。”令人震惊的是，“前 20 名左右的 Kaggle 团队大概达到了人类心脏病科医师的效果。”

合作与社区

我已经开始感受到正在发生令人难以置信的现象！在单项竞赛之外，Kaggle 似乎对整个机器学习领域都带来了重要的影响。“Kaggle 是一个非常出色的创新机器散布平台，因为只要有人在自己参与的竞赛排行榜上看到冠军，他们就会花费数个小时，真正地想弄明白是什么打败了自己，获胜者是如何获胜的。这意味着如果有一项技巧似乎在单项比赛中效果很不错，那么就有一千名或一万名参与者分析该技巧，并研究它的工作原理。这使得 Kaggle 成为证明该技巧真的能够很好地解决问题并促使大批人士采用这一技巧的有力平台。”

等等，这些不是竞赛吗？参与者为何会透露花费了很多精力开发的解决方案？Ben 解释道，“没错，在竞赛过程中，我们社区的顶尖人士相互之间竞争异常激烈。但是比赛结束之后，就会变成紧密合作的环境，大部分情况下，我们会发布顶级团队的获胜访谈，他们会深入描述解决方案。在很多情况下，我们实际上会公布获胜解决方案的代码。所有为研究项目开展的竞赛都要求前三名参与者对他们的解决方案进行开源。”

对于社区而言，Kaggle 正在致力于降低参与门槛，并使用 [Kaggle Scripts](#) 等工具实现更好的协作环境。“它可以让你快速进入 R、Python 或 Julia 计算环境，既链接到竞赛数据集相关联，又链接到需要使用但是安装起来很头疼的分析包。”这些脚本和相关结果会自动发布，使任何人都能分叉并加以改善。还是初学者学习并汲取经验的宝贵资源。

利用 Kaggle 奠定机器学习职业基础

积极参与 Kaggle 社区活动不仅仅可以接触大量的问题，并获得社区的支持，而且是踏入机器学习这一职业领域的重要环节。某些竞赛明确用来招聘人才，例如 Facebook 的 [签到识别问题](#)，竞赛成绩好的参与者将获得机器学习软件工程师职位面试机会。Kaggle 还通过他们的 [工作台](#)帮助联系招聘人士和应聘者。但即使不是这些情况，“雇主也会很关注你的 Kaggle 个人资料...出色的竞赛成绩可以有效地让你脱颖而出，并展现出你能够有效解决问题的能力。”

在很短的时间内，Kaggle 就从一个普通的竞赛平台发展成应用机器学习研究成果的枢纽。它成为难以解答的数据问题和渴望挑战、充满激情的参与者之间的桥梁。或许这种合作模式正是我们持续发展所需的模式。Ben 赞同机器学习依然处于发展初期阶段这一观点。“机器学习将变得和创建及使用 iPhone 应用或网络一样那样简单。因此将不再是一项神奇的技能，需要少数真正专业的人士来创建和使用。机器学习的接触门槛将越来越商品化和大众化。届时，将会推动一大批新的应用面世。”

*你可以在此处观看完整的Ben Hamner (@benhamner) 访谈。如果你想加入机器学习社区（或加深参与程度），在资深竞争者的帮助下亲身体验 Kaggle，并直接掌握这一领域的就业机会，不妨报名参加由 Kaggle 与 Udacity、Google 联合打造的“机器学习工程师纳米学位”项目》。这门课程是帮助你掌握机器学习关键技术技能的最佳方式，同时还可以在导师的带领下进入到神奇的机器学习世界，为你的未来职业发展奠定基础！



机器学习或许最吸引人的一个方面就是似乎存在无限的应用潜力。已经有无数的领域受到机器学习的影响，包括教育、金融、计算机科学等。几乎没有机器学习不适用的领域。在某些情况下，实际上急需机器学习技术。医疗保健就是个很明显的示例。机器学习技术技巧已经运用到医疗保健关键领域，影响到方方面面，包括缩小医疗保健差距的努力和医疗扫描分析。David Sontag 是纽约大学库朗数学学院和纽约大学数据科学中心的助理教授，最近发表了一场关于机器学习和医疗保健系统的演讲，他指出了“机器学习会如何潜在地改变各个医疗行业，从形成下一代电子医疗记录到从医疗保险索赔中得出人口级风险层化数据”。

毫无疑问，世界在瞬息万变，对机器学习工程师的需求将以指数级增长。世界面临着复杂的挑战，并且需要复杂的系统来解决这些挑战。机器学习工程师正在构建这些系统。如果这就是你的未来发展规划，那么现在则是掌握相关技能和培养成功所必需的思维的最佳时机。

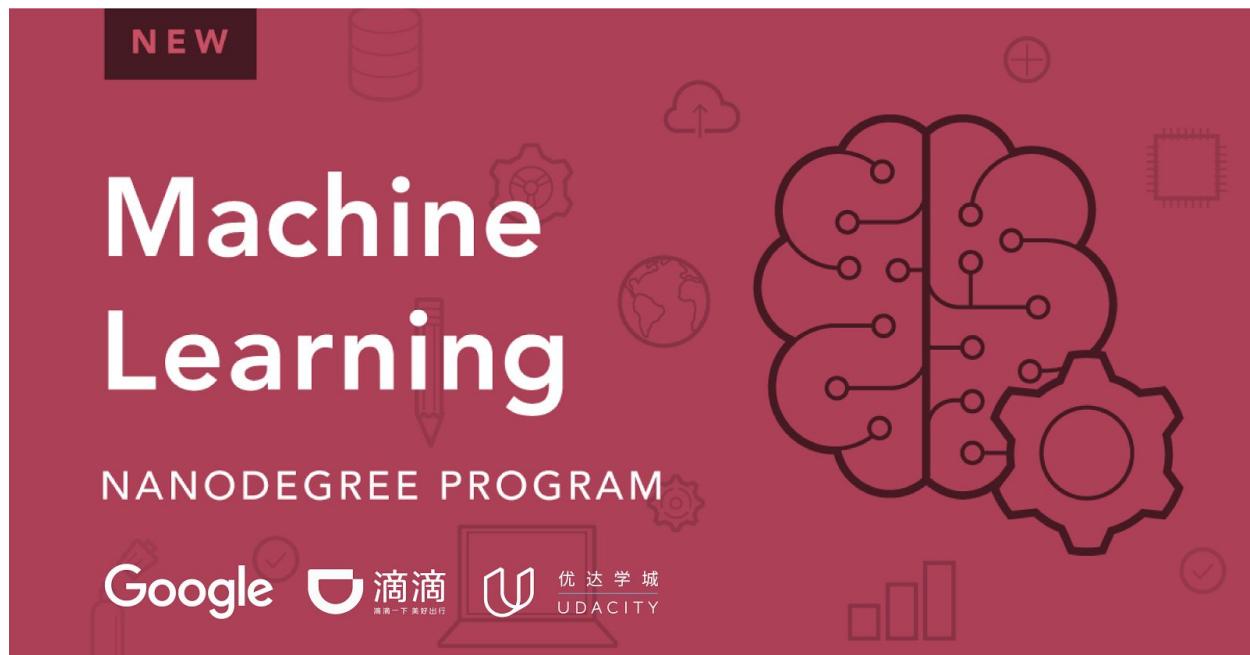
加入21世纪前沿科技浪潮，成为高薪抢手、改变世界的机器学习工程师

滴滴出行前不久宣布在硅谷成立实验室，研究以机器学习为核心的“自动驾驶”技术；百度宣布以人工智能为未来10年最重要战略；Google、Uber更是在自动驾驶、机器学习等领域内进行激烈的人才争夺。从Google、Facebook、亚马逊到BAT，到国内新兴的科技企业都在热忱地以高薪招募机器学习专家。在招聘网站上，也可以轻易搜到诸多相关职位、了解他们诱人的薪酬待遇。

据调查，中国需要500万人工智能开发人才，但现在 LinkedIn 上有资料的仅有5万。掌握稀缺的机器学习开发技术，成为人工智能、金融、大数据领域的枪手人才，获得加入名企的敲门砖。

那么问题来了，如何加入这场21世纪前沿科技的浪潮，成为高薪抢手又能改变世界的机器学习工程师？

Udacity 与 **Google**、**滴滴出行**、**Kaggle** 联合打造的 “**机器学习工程师**” 纳米学位将成为你的最佳起点，它可以让你在家就能学习全球最高水平的机器学习课程，零基础入门，掌握如何用机器学习在大数据、金融、人工智能领域进行预测分析，成为被 Google 认证的工程师，获得理想工作。



针对不同程度的学员，机器学习工程师纳米学位课程分为入门与进阶两个部分：

[**机器学习入门课程**](#)将带你从零开始，一站式搞定入门机器学习需要的编程和数学基础，掌握 Python、微积分、线性代数和统计基础知识，为成为机器学习工程师打好基础。

[**机器学习进阶课程**](#)将带你全面了解、掌握机器学习领域内的监督式学习、非监督式学习、强化学习和深度学习，并亲手挑战前沿应用项目，成为人工智能、大数据、金融领域稀缺人才。

在这一纳米学位项目中，你所接触到的实战项目来自于行业最前沿的话题，例如毕业项目的可选项中有五项来自于 Kaggle 比赛。

如果你想学习更多人工智能领域的知识与技能，毕业后你可以选择[深度学习](#)、[无人车开发](#)、[机器人开发](#)等方向继续进行深造。



正如我们在这份指南的介绍中提到的，机器学习是一个非常前沿，人才稀缺的新兴行业。在读了我们的所有介绍之后，希望你能清楚地明白机器学习领域有各种各样实现结果的方法与工具，重要的不是你“记住”了什么，而是你真正“掌握”了什么，我们的机器学习工程师纳米学位项目专门为你进入这一异常前沿的职业发展领域做好准备，教会你在快速发展的领域中运用所学，成为最优秀的机器学习工程师。

我们与 Google、Kaggle 和滴滴出行等领先公司的专家合作，开发了内容广泛的课程表，几乎涵盖了我们在此指南中提到的所有主题内容，从利用统计分析工具对观测数据建立模型，到学习如何训练决策树、SVM、神经网络等监督学习模型，用来预测已标记数据；从使用 Q-学习等强化学习算法，训练人工智能体，使它能够对周围环境做出最佳选择，到搭建一个卷积神经网络，可以识别图片中的物体。

我们的课程以项目为基础，将帮助你对机器学习产生极为深入的理解，获得实战操作的真本领，使你在面对顶尖企业机器学习工程师的面试时胸有成竹，成为企业青睐的雇员。这门课程也能为你进一步学习深度学习、无人车开发、机器人开发奠定良好的基础。

掌握机器学习技能将帮助你很快在职场上成为抢手人才，并能帮助你获得优渥的薪资如果你想选择机器学习工程师这一职业道路，那么可以阅读这份指南，我们的机器学习工程师纳米学位项目也会帮助你走向成功。

准备好加入抢手机器学习工程师的新时代人才大军了吗？马上加入 **Udacity 与 Google、Kaggle、滴滴出行** 官方联合打造的 [机器学习工程师纳米学位项目](#) 吧！