

## Exercise 1

你将如何定义机器学习？

答案：

机器学习是关于构建可以从数据中学习的系统。学习意味着在给定一些绩效衡量标准的情况下，更好地完成某些任务。

## Exercise 2

你能说出四种关于机器学习的应用程序吗？

答案：

机器学习非常适合解决我们没有算法解决方案的复杂问题，取代长长的手动调整规则列表，构建适应波动环境的系统，并最终帮助人类学习（例如，数据挖掘）。

## Exercise 3

什么是带有标记的训练集？

答案：

带标签的训练集是包含每个实例所需解决方案（也称为标签）的训练集。

## Exercise 4

两种最常见的监督任务是什么？

答案：

两个最常见的监督任务是回归和分类。

## Exercise 5

你能说出四个常见的无监督任务吗？

答案：

常见的无监督任务包括聚类、可视化、降维、关联规则学习。

## Exercise 6

你会用什么类型的算法来让机器人在各种未知的地形上行走？

答案：

如果我们想让机器人学会在各种未知地形中行走，强化学习可能会表现最好，因为这通常是强化学习解决的问题类型。 可以将问题表达为监督或半监督学习问题，但这不太自然。

## Exercise 7

您将使用什么类型的算法来将客户分成多个组？

答案：

如果您不知道如何定义组，则可以使用聚类算法（无监督学习）将您的客户划分为类似客户的集群。 但是，如果您知道自己想要哪些组，那么您可以将每个组的许多示例提供给分类算法（监督学习）， 它会将您的所有客户分类到这些组中。

## Exercise 8

你会把垃圾邮件检测问题定义为监督学习问题还是无监督学习问题？

答案：

垃圾邮件检测是一个典型的监督学习问题：该算法会收到许多电子邮件及其标签（垃圾邮件或非垃圾邮件）。

## Exercise 9

什么是在线学习系统？

答案：

在线学习系统可以增量学习，而不是批量学习系统。 这使得它能够快速适应不断变化的数据和自治系统，并能够对大量数据进行训练。

## Exercise 10

什么是核外学习？

答案：

核外算法可以处理计算机主内存无法容纳的大量数据。 核外学习算法将数据分成小批量，并使用在线学习技术从这些小批量中学习。

## Exercise 11

什么类型的算法依赖于相似性度量来进行预测？

答案：

基于实例的学习系统会记住训练数据； 然后，当给定一个新实例时，它使用相似性度量来找到最相似的学习实例并使用它们进行预测。

## Exercise 12

模型参数和模型超参数之间的区别是什么？

答案：

一个模型有一个或多个模型参数，这些参数决定了它在给定一个新实例时会预测什么（例如，线性模型的斜率）。学习算法试图为这些参数找到最佳值，以便模型可以很好地泛化到新实例。

超参数是学习算法本身的参数，而不是模型的参数（例如，要应用的正则化量）。

## Exercise 13

基于模型的算法会搜索什么？他们最常用来获得成功的策略是什么？他们是如何做出预测的？

答案：

基于模型的学习算法搜索模型参数的最佳值，以便模型可以很好地泛化到新实例。

我们通常通过最小化成本函数来训练此类系统，该成本函数衡量系统在对训练数据进行预测时的糟糕程度，如果模型被正则化，还会对模型复杂度进行惩罚。

为了进行预测，我们使用学习算法找到的参数值将新实例的特征输入模型的预测函数。

## Exercise 14

你能说出机器学习中的四个主要挑战吗？

答案：

机器学习的一些主要挑战是缺乏数据、数据质量差、不具代表性的数据、无信息的特征、欠拟合训练数据的过于简单的模型，以及过拟合训练数据的过于复杂的模型。

## Exercise 15

如果你的模型在训练数据上表现很好，但对新的实例泛化不佳，那么发生了什么？你能说出三种可能的解决方案吗？

答案：

如果一个模型在训练数据上表现很好但对新实例的泛化能力很差，则该模型可能过拟合训练数据（或者我们在训练数据上非常幸运）。

过拟合的可能解决方案是获取更多数据、简化模型（选择更简单的算法、减少使用的参数或特征的数量，或正则化模型），或减少训练数据中的噪声。

## Exercise 16

什么是测试集，您为什么要使用它？

答案：

在模型投入生产之前，测试集用于估计模型将在新实例上产生的泛化误差。

## Exercise 17

验证集的目的是什么？

答案：

验证集用于比较模型。它使得选择最佳模型和调整超参数成为可能。

## Exercise 18

什么是训练开发集（**train-dev set**），你什么时候需要它，你如何使用它？

答案：

当训练数据与验证和测试数据集中使用的数据存在失配的风险时，将使用 **train-dev** 集（一旦模型投入生产，应始终尽可能接近所使用的数据）。

**train-dev** 集是保留的训练集的一部分（模型未在其上训练）。该模型在训练集的其余部分进行训练，并在训练开发集和验证集上进行评估。

如果模型在训练集上表现良好但在训练-开发集上表现不佳，则该模型可能过度拟合训练集。

如果它在训练集和训练-开发集上都表现良好，但在验证集上表现不佳，那么训练数据和验证+测试数据之间可能存在明显的数据不匹配，您应该尝试改进训练数据，使其看起来更像验证+测试数据。

## Exercise 19

如果您使用测试集调优超参数，可能会出现什么问题？

答案：

如果您使用测试集调整超参数，则存在过度拟合测试集的风险，并且您测量的泛化误差将是乐观的（您可能会启动一个性能比预期差的模型）。