1.你会怎么定义机器学习?

2.机器学习在哪些问题上表现突出，你能提出四种类型吗?

3.什么是被标记的训练数据集?

4.最常见的两种监督式学习任务是什么?

5.你能举出四种常见的无监督式学习任务吗?

6.要让一个机器人在各种未知的地形中行走，你会使用什么类型的机器学习算法?

7.要将顾客分成多个组，你会使用什么类型的算法?

8.你会将垃圾邮件检测的问题列为监督式学习还是无监督式学习?

9.什么是在线学习系统?

10.什么是核外学习?

11.什么类型的学习算法依赖相似度来做出预测?

12.模型参数与学习算法的超参数之间有什么区别?

13.基于模型的学习算法搜索的是什么？它们最常使用的策略是什么？它们如何做出预测?

14.你能提出机器学习中的四个主要挑战吗?

15.如果你的模型在训练数据上表现很好，但是应用到新的实例上的泛化结果却很糟糕，是怎么回事？能提出三种可能的解决方案吗?

16.什么是测试集，为什么要使用测试集?

17.验证集的目的是什么?

18.如果使用测试集调整超参数会出现什么问题?

19.什么是交叉验证？它为什么比验证集更好?

1.机器学习是一门能够让系统从数据中学习的计算机科学。

2.机器学习非常利于：不存在已知算法解决方案的复杂问题，需要大量手动调整或是规则列表超长的问题，创建可以适应环境波动的系统，以及帮助人类学习（比如数据挖掘）。

3.被标记的训练集是指包含每个实例所期望的解决方案的训练集。

4.最常见的两个监督式任务是回归和分类。

5.常见的无监督式任务包括聚类、可视化、降维和关联规则学习。

6.如果想让机器人学会如何在各种未知地形上行走，强化学习可能表现最好，因为这正是一个典型的强化学习擅长解决的问题。将这个问题表达为监督式或半监督式学习问题也可以，但还是有点不太自然。

7.如果你不知道如何定义分组，那么可以使用聚类算法（无监督式学习）将相似的顾客分为一组。但是，如果你知道想要的是什么样的群组，那么可以将每个组的多个示例反馈给分类算法（监督式学习），它就可以将所有的顾客归类到这些组中。

8.垃圾邮件检测是个典型的监督式学习问题：将邮件和它们的标签（垃圾邮件或非垃圾邮件）一起提供给算法。

9.在线学习系统可以进行增量学习，与批量学习系统正好相反。这使得它能够快速适应不断变化的数据和自动化系统，并且能够在大量的数据上进行训练。

10.核外算法可以处理计算机主内存无法应对的大量数据。它将数据分割成小批量，然后使用在线学习技术从这些小批量中学习。

11.基于实例的学习系统通过死记硬背来学习训练数据，当给定一个新实例时，它会使用相似度度量来找到与之最相似的实例，并用它们进行预测。

12.模型有一个或多个参数，这些参数决定了模型对新的给定实例会做出怎样的预测（比如，线性模型的斜率）。学习算法试图找到这些参数的最佳值，使得该模型能够很好地泛化至新实例。超参数是学习算法本身的参数，不是模型的参数（比如，要应用的正则化数量）。

13.基于模型的学习算法搜索使模型泛化最佳的模型参数值。通常通过使成本函数最小化来训练这样的系统，成本函数衡量的是系统对训练数据的预测有多坏，如果模型有正则化，则再加上一个对模型复杂度的惩罚。学习算法最后找到的参数值就是最终得到的预测函数，只需要将实例特征提供给这个预测函数即可进行预测。

14.机器学习面临的一些主要挑战是：数据缺乏、数据质量差、数据不具代表性、特征不具信息量、模型过于简单对训练数据拟合不足，以及模型过于复杂对训练数据过度拟合。

15.如果模型在训练数据上表现很好，但是对新实例的泛化能力很差，那么该模型很可能过度拟合训练数据（或者在训练数据上运气太好）。可能的解决方案是：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数量、对模型正则化），或者是减少训练数据中的噪声。

16.在模型启动至生产环境之前，使用测试集来估算模型在新实例上的泛化误差。

17.验证集用来比较不同模型。它可以用来选择最佳模型和调整超参数。

18.如果使用测试集来调整超参数，会有过度拟合测试集的风险，最后测量的泛化误差会过于乐观（最后启动的模型性能比预期的要差）。

19.通过交叉验证技术，可以不需要单独的验证集实现模型比较（用于模型选择和调整超参数），这节省了宝贵的训练数据。