第一章 概念理解：

人工智能：

1. 弱人工智能：让机器具备观察和感知的能力，可以做到一定程度的理解和推理；
2. 强人工智能：让机器获得自适应能力，解决一些之前没有遇到过的问题。

机器学习：一种实现人工智能的方法（使用算法来解析数据、从中学习，然后对真实世界中的事件做出决策和预测） eg. 网购的时候，浏览网上商城时，经常会出现商品推荐的信息。这是商城根据往期的购物记录和收藏清单等数据的收集，识别出这其中哪些是你真正感兴趣，并且愿意购买的产品。这样的决策模型，可以帮助商城为客户提供建议并鼓励产品消费。

深度学习：一种实现机器学习的技术---最初的深度学习是利用深度神经网络来解决特征表达的一种学习过程。为包含多个隐含层的神经网络结构。为了提高深层神经网络的训练效果，对神经元的连接方法和激活函数等方面做出相应的调整。

深度学习，存在以下问题（业界有一种错误的较为普遍的意识，即“深度学习最终可能会淘汰掉其他所有机器学习算法”）：

1. 深度学习模型需要大量的训练数据，才能展现出神奇的效果，但现实生活中往往会遇到小样本问题，此时深度学习方法无法入手，传统的机器学习方法就可以处理；

2. 有些领域，采用传统的简单的机器学习方法，可以很好地解决了，没必要非得用复杂的深度学习方法；

3. 深度学习的思想，来源于人脑的启发，但绝不是人脑的模拟.

eg. 给一个三四岁的小孩看一辆自行车之后，再见到哪怕外观完全不同的自行车，小孩也十有八九能做出那是一辆自行车的判断，也就是说，人类的学习过程往往不需要大规模的训练数据，而现在的深度学习方法显然不是对人脑的模拟。

Types of Machine Learning Systems:

**Supervised learning**

1. In supervised learning, the training set you feed to the algorithm includes the desired solutions, called labels
2. Another typical task is to predict a target numeric value, such as the price of a car, given a set of features (mileage, age, brand, etc.) called predictors.

Eg.常用监督算法：

* Linear Regression
* k-Nearest Neighbors
* Logistic Regression
* Support Vector Machines (SVMs)
* Decision Trees and Random Forests
* Neural networks2

**Unsupervised Learning**（就像是没有老师的课堂，学生通过接受的教材和练习册进行自学，最终通过考试）

Eg

• Clustering

— K-Means

— DBSCAN

— Hierarchical Cluster Analysis (HCA)

• Anomaly detection and novelty detection

— One-class SVM

— Isolation Forest

• Visualization and dimensionality reduction

— Principal Component Analysis (PCA)

— Kernel PCA

— Locally Linear Embedding (LLE)

— t-Distributed Stochastic Neighbor Embedding (t-SNE)

• Association rule learning

— Apriori

— Eclat

比如通过聚类算法进行简单分类，利用可视化算法（这些算法试图尽可能多地保留结构(例如，试图在输入空间中保持独立的集群，以免在可视化中重叠)）。

非监督学习算法的相关任务：

1. 降维，其目标是在不丢失太多信息的情况下简化数据。一种方法是将几个correlated特性合并为一个。

Eg，一辆汽车的行驶里程可能与它的年龄密切相关，因此维数缩减算法会将它们合并成一个特征来表示汽车的磨损。这叫做特征提取。

1. 异常检测，eg，检测不寻常的信用卡交易，以防止欺诈，捕捉制造缺陷，或自动删除离群值从数据集之前，将其提供给另一个学习算法。系统在训练中看到的大多是正常的例子，所以它学会了识别它们;然后，当它看到一个新的实例时，它可以判断它看起来是正常的还是可能是异常的。

\*一个非常类似的任务是新奇检测:它的目标是检测与训练集中所有实例不同的新实例。这需要有一个非常干净的训练集，不包含任何你希望算法检测的实例。例如，有成千上万的狗的照片，而这些照片中只有1%是吉娃娃狗，那么新奇检测算法就不应该把新的吉娃娃狗照片当作新奇的东西。另一方面，异常检测算法可能会认为这些狗如此罕见，与其他狗如此不同，以至于可能会将它们归类为异常。

**Semisupervised learning**（有些算法可以处理部分标注的数据。）

大多数半监督学习算法是无监督和有监督算法的结合。例如，深度信念网络(DBNs)是基于被称为限制性玻尔兹曼机器(RBMs)的未被监视的组件，这些组件堆叠在一起。RBMs以无监督的方式进行顺序训练，然后使用监督学习对整个系统进行微调。

1.Reinforcement Learning

（通过大量的数据源分析实现）

2.Batch and Online Learning

批处理学习  
在批量学习中，系统不能渐进地学习:它必须使用所有可用的数据进行训练。这通常会花费大量的时间和计算资源，所以它通常是离线完成的。首先对系统进行培训，然后投入生产，无需学习就能运行;它只是应用它学到的东西。这被称为离线学习。

在线学习  
在在线学习中，您可以通过连续地向系统提供数据实例来逐步地培训系统，这些数据实例可以是单独的，也可以是称为迷你批处理的小组。（相对廉价）

**Instance-Based Versus Model-Based Learning**

另一种分类机器学习系统的方法是通过它们的归纳。  
大多数机器学习任务都是关于预测的。这意味着给定一些训练示例，系统需要能够对以前从未见过的示例做出良好的预测(泛化)。对训练数据有一个良好的绩效衡量是好的，但还不够;真正的目标是在新实例上执行良好。

1. 基于实例的学习  
   最简单的学习方式可能就是背诵。如果想用这种方式创建一个垃圾邮件过滤器，它只会标记所有与已经被用户标记的邮件相同的邮件——这不是最坏的解决方案，但肯定不是最好的解决方案。
2. 基于模型的学习  
   另一种从一组例子中归纳出的方法是建立一个这些考试标准的模型，然后利用这个模型进行预测。

什么样的数据属于Poor-Quality Data

某些实例存在明显的异常值和特性确实，可选择手动修复或者丢弃。  
Irrelevant Features

Overfitting the Training Data 和 Underfitting the Training Data（当模型过于简单而无法了解数据的底层结构时，就会发生这种情况。）   
Eg假设你去外国旅游，出租车司机把你宰了。你可能会说那个国家的所有出租车司机都是小偷。过度概括是我们人类经常做的事情，机器也会落入同样的陷阱。在机器学习中，这被称为过拟合:这意味着模型在训练数据上表现良好，但不能很好地推广。

**Testing and Validating：**

Hyperparameter Tuning and Model Selection （关于选择合适的模型，比如线性模型还是多项式模型）

我能理解的一种选择是：同时训练两者，并比较它们使用测试集的泛化效果。

Data Mismatch

**第二章End-to-End Machine Learning Project**

1. Look at the Big Picture
2. Frame the Problem
3. **Select a Performance Measure**
4. **Get the Data**
5. **Create the Workspace**
6. **Download the Data**