各位学员下午好，欢迎来到机器学习课堂，本次课程我们将从机器学习的概念、机器学习的分类和决策树算法三个方面来介绍机器学习的基本知识。下面依次进行

在具体概念前，我们先来思考人是怎么学习的？不管是婴儿时期的学习说话、学习走路，还是长大后的在校学习，人的学习过程基本符合上图的流程。基于一定的样本资料，经过思考归纳总结，得到知识和经验。当出现类似任务的时候，就可以根据知识和经验做出反应。这就是人的学习过程。

仿照人的学习过程，我们可以给出如图机器的学习过程，输入大量学习资料，运行机器学习算法进行训练，最终得到一个模型。当有新的样本输入模型，就可以得到预测结果。这里运行机器学习算法训练的过程，就好比人的归纳总结的过程。训练的模型就好比知识和经验。这就是机器学习的过程。

早在1959年，机器学习先驱阿瑟•萨谬尔就给出了机器学习的定义，说机器学习是不显式编程地赋予计算机能力的研究领域，这里赋予计算机能力很好理解，关键在于如何理解不显式编程？

下面我们举例说明，如图我们这里有一些菊花和玫瑰的图片，任务是识别图片是菊花还是玫瑰。那么显式编程我们可能就直接告诉计算机黄色菊花，红色玫瑰，计算机也就按照这个规则去判断；而不显式编程，我们不告诉计算机如何判断，仅说这些是菊花，这些是玫瑰，计算机自己根据算法选择策略去判断，这就是不显式编程。

下面我们来讲解机器学习的分类，根据学习方式不同可以将机器学习分为监督学习，无监督学习和半监督学习。需要注意的是机器学习根据不同的标准可以有许多不同的分类方法，这里介绍的仅是最为常见的一种分类方法。

根据这张表格可以很清楚看出监督学习，非监督学习和半监督学习之间的区别！首先训练数据有无标签是他们的最大区别，监督学习训练数据全部具有标签，非监督学习全部没有标签，半监督学习部分数据具有标签。这也导致三种学习各自具有不同的特点，具体来说，监督学习根据有标签数据主要训练获得模型用于分类和回归；非监督学习主要根据无标签数据特征通过聚类等方法对数据划分类别；半监督学习结合其他两种学习的学习方式，可以先根据有标签数据建立模型，再给无标签数据标定标签，也可以先通过聚类等方法划分类别，根据有标签数据标定类别标签，再训练模型。最后一列给出了三种学习各自常见的算法，后面我们会详细介绍决策树算法，其他算法学员可以课下自行学习。

学习了机器学习的概念和分类，下面我们来介绍决策树算法。百度百科给出的定义是决策树是一种从无次序、无规则的样本数据集中推理出决策树表示形式的分类规则方法。其实决策树就是由一系列判断条件构成的树型结构。如右图就是一棵简单的决策树，树的叶子结点表示预测结果，分支节点表示判断条件，连线表示不同的决策。右图这棵决策树表示根据年龄和性别给人一个标签，小于18岁，标签小孩，大于18岁，男性标签男人，女性标签女人。决策树的优点：速度快，准确性高，缺点：忽略了属性之间的相关性。

结合机器学习过程我们给出决策树的学习过程，如图带标签的训练数据，通过决策树算法训练得到决策树模型。当测试数据输入决策树，就可以得到预测结果。这就是决策树的算法流程。

决策树算法的核心在于如何构造决策树？下面我们以CART树的gini系数为例进行讲解。gini系数可以理解为一个函数，数值越小分类效果越好。第一步从根节点出发，包含所有的训练样本。第二步判断结点标签是否属于同一个类别，如果是就将其定义为叶子结点，并取值。第三步如果不属于同一个类别，那就选取gini最小的特征将数据分为两部分，并分支为两个新节点。第四步，回到第二步递归检查新节点，直到数据都在叶子节点不再分裂，得到决策树。

下面我们给出一个决策树算法例子，还是关于菊花和玫瑰，训练数据特征有颜色、花蕊大小和花瓣长度，类别标签有菊花和玫瑰。算法的目的是通过训练一颗决策树，可以用来根据三个特征预测是玫瑰还是菊花。根据决策树构建算法，我们可以得到右图所示的决策树。当有下方的测试数据特征输入决策树，可以得到两个预测结果。大家仔细观察就会发现，在这个模型中，不论测试数据为什么颜色都不影响结论，这就是不显式编程的魅力，发现数据本身的特征优势，避免人为因素导致的低效率设计。

总结，本次课介绍了机器学习的概念，不显式编程地赋予计算机能力，机器学习的分类，监督，无监督，半监督；还有决策树算法的概念。课程到此结束，谢谢！