同学们下/上午好,欢迎来到机器学习课堂,本次课程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介绍机器学习的基本知识,使同学们对机器学习有个初步认识,下面依次进行:

首先我们学习机器学习的概念：在讲解具体概念前,我们先来思考下人是如何学习的？不管是婴儿时期的学习走路、学习说话,还是长大以后在校学习，人的学习过程基本都符合上图的流程，基于一定的学习资料,经过思考学习整理归纳等过程，形成知识和经验。人们根据知识和经验在遇到类似任务的时候就可以做出相应的反应，这就是人的学习过程。

那么机器该如何学习呢？仿照人的学习过程，我可以给出如下机器的学习过程,输入大量学习资料，运行机器学习算法进行学习，最后得到一个训练模型。当新的样例输入模型，就可以得到预测结果。这里运行机器学习算法的过程就好比人归纳整理思考的过程，得到的模型就好比人学习到的知识和经验，这就是机器学习的过程。

早在1959年机器学习先驱阿瑟·萨缪尔就给出了机器学习的定义：机器学习是**不显式编程**地赋予计算机能力的研究领域。赋予计算机能力很好理解就是我们无数程序员每天做的事情，让计算机能够执行某项任务以提高我们生活工作的效率。关键在于如何理解不显式编程？

下面我们给出一个具体的例子带同学们理解一下究竟什么是不显式编程。如图我这里有一些菊花和一些玫瑰花的图片，任务是识别图片是菊花还是玫瑰花，如果显式编程，那么我就可能直接告诉计算机黄色是菊花，红色是玫瑰花，计算机就直接根据颜色是黄色还是红色判断图片是菊花还是玫瑰花。不显式编程就是不直接告诉计算机如何判断，仅仅告诉他这些图片是菊花，另一些图片是玫瑰花，然后让计算机跟据算法自己学习如何区分菊花和玫瑰花。这样也许计算机还是按照颜色来判断，但也有可能根据花瓣的长度，花蕊的大小等去判断，也有可能是多个特征的组合去判断，具体就看训练样本的选择。综上可见，不显式编程的特点就是：在模型训练结束前，我们不知道计算机最终会以什么判断依据来识别样本。机器学习的关键也就在如何实现这个不显式编程算法。

下面我们来讲解机器学习的分类：根据学习方式的不同可以把机器学习分为监督学习、非监督学习、半监督学习。它们的不同之处就在于训练数据是否都具有标签。比如我们前面提到的菊花和玫瑰的例子，如果我仅给了一堆菊花和玫瑰的图片，并不标注哪些是菊花，哪些是玫瑰，那么这些就是无标签数据，反之每个图片都带有菊花或者玫瑰的标签就是有标签数据。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，我们这里讲解的仅是其中最常见的一种分类方法。

通过这个表格我们可以很清楚知道监督学习、非监督学习和半监督学习之间的区别。监督学习要求训练数据全部具有特征和标签，通过训练得到的模型为函数，主要用于分类和回归。非监督学习训练数据只有特征，模型根据特征自行选择类别，常见就是聚类。半监督学习部分数据具有特征和标签，部分数据只有特征，结合两类数据生成函数。

还拿菊花和玫瑰来举例，带标签的菊花和玫瑰的图片训练得到模型后，得到的模型函数可以用来判断新的无标签图片是菊花还是玫瑰，这就是监督学习。没有标签的菊花和玫瑰的一堆图片，通过聚类可以划分出不同的类别，给与标签，这就是无监督学习。半监督学习结合了监督学习和无监督学习，你可以根据有标签的数据先建立模型，给无标签数据标定标签，也可以先通过聚类划分出不同类别，再根据有标签数据标定类别标签，这就是半监督学习。

学习了机器学习的概念和分类，下面我们介绍一个具体的机器学习算法——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。其实决策树就是有一系列判断条件构成的用于决策的树形结构。如右图所示，叶子节点表示标签，或者说是预测结果，分支节点表示判断条件，连线表示不同的决策。这里我们输入任何一个人的年龄和性别就可以得到一个标签，年龄小于18岁，不论男女标签都是小孩，年龄大于18岁如果性别男，那么就给男人标签，如果性别女，就给女人标签，这就是一棵简单的决策树模型。决策树的有点速度快，准确性高，缺点是忽略了属性之间的相关性。优点等后面学习了更多的机器学习算法就可以比较，缺点也很明显，决策树节点之间没有必然的关系，不能体现数据属性相关性的信息。

理解了决策树定义后，结合我们介绍的机器学习过程可以得出决策树的算法流程，如图所示，带标签的训练数据，通过决策树算法训练，就可以得到决策树模型。当测试数据特征输入，决策树就能预测出测试数据的标签。这就是决策树算法的流程。决策树模型我们通过定义已经很好理解，这里的关键点就是如何实现决策树算法？或者说如何构建一棵决策树？

下面我们给出一种决策树的构建方法，由不同的指标选择其实有多种决策树的构建方法，这里以CART树的gini系数为例讲解。Gini系数我们目前就可以理解为一个函数，这个函数值越小标明分类效果越好。从根节点出发，根节点包含所有的训练样本。如果节点包含的标签同属于同一个类别，那么就将该节点定义为叶子节点，取数据的类别作为叶子结点的取值。如果节点包含的标签不属于同一个类别，那么选择GINI最小的特征将数据分为两部分，并分支为两个新节点。回到 2）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。经过这么一个过程我们就可以得到一棵决策树，这就是决策树算法的核心内容。

下面我们给出一个具体的例子，还是关于菊花和玫瑰，这里不是图片了,而是以颜色、花蕊大小、花瓣长度构成的数据集，标签是菊花和玫瑰。根据决策树的构建方法，我们首先找出最小的gini系数的特征为花蕊大小，所以根据花蕊大小把数据划分为两部分，又因为左侧数据属于同一类别所以我们指定为叶子节点，值为菊花；右侧数据我们重新选择gini系数最小的特征为花瓣长度，同样的过程我们得到了两个叶子结点。至此所有的节点都递归完成，得到决策树如右图所示。下面我们给出两个测试数据，通过决策树，很轻松就得到两个预测值，这里预测结果都为菊花。大家仔细观察这个模型会发现，原来不论测试数据为什么颜色都不影响结论，这在我们看到菊花和玫瑰图片的时候可是完全没有想到的，甚至给出左侧的训练数据的时候我们都不知道会是这样的结果。这就是不显式编程的魅力，发现数据本身的特征优势，选择最优策略，避免人为因素导致的低效率设计。

简单总结下，我们今天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，有监督，无监督，半监督；以及详细介绍了机器学习算法决策树算法。本次课程到此结束，谢谢！