机器学习课程复习要点：

1. 机器学习的定义。

机器学习(Machine Learning ML)是研究通过已有的经验可以进行自动改进的计算机算法，是人工智能的一个子集。机器学习算法基于样本数据(即已有的“训练数据”)建立数学模型，以便做出预测或决策，而无需经过明确的编程。机器学习算法被广泛应用于各种应用中，如电子邮件过滤、计算机视觉以及推荐系统等领域

对于某类任务T和性能度量P, 如果一个计算机程序在T上以P衡量性能随着经验E而自我完善，那么称该计算机程序在从经验E学习。Tom Mitchell, 《Machine Learning》

1. 有监督、无监督、半监督算法的各代表什么含义。

**有监督：**在样本标签已知的情况下，可以统计出各类训练样本不同的表述量，如其概率分布，或在特征空间分布的区域等，利用这些参数进行分类器设计，称为有监督学习。典型方法：SVM，ANN等。分类如下：

* 分类问题

标签信息是连续值SVM, ANN,RandomForest

* 回归问题

**无监督：**标签信息是离散的SVM，NN，RandomForest

实际应用中，不少情况下无法预先知道样本的标签。学习算法根据外部数据的统计规律来调节参数，以使输出能反映数据的某些内在结构。有监督学习与无监督学习的区别：训练集是否有标注。典型方法：Kmeans等聚类算法。分类如下：

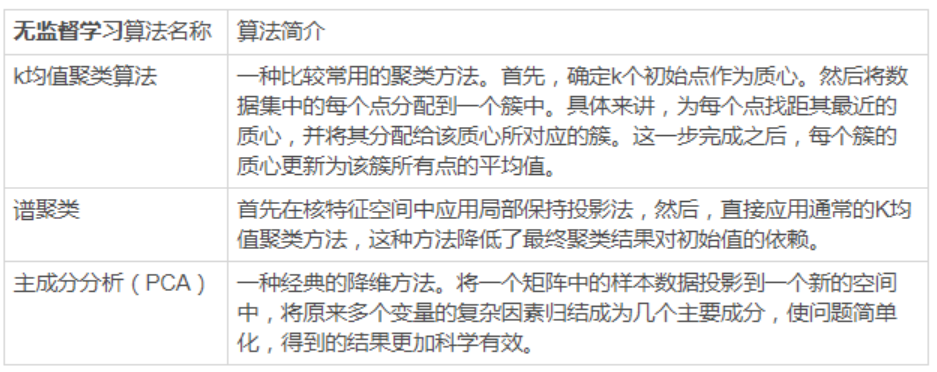
* 强化学习
* 半监督学习

**半监督算法：**介于有监督学习与无监督学习之间的机器学习方式，主要考虑如何利用少量的标注样本和大量未标注的样本进行训练和分类的问题。

应用场景类似于有监督学习，包括分类与回归。

半监督学习主要用于处理人工合成数据，无噪声干扰的数据样本是当前大部分半监督学习方法使用的数据，而在实际生活中用到的数据大部分是受噪声干扰的，通常难以得到纯样本数据。

1. 无监督算法中有哪些经典算法。



1. 本课程实验中采用的是什么算法库。

Scikit-learn

1. 什么是模型拟合？有哪些拟合方式。

对于机器学习模型而言，我们不仅要求它对训练数据集有很好的拟合（训练误差），同时也希望它可以对未知数据集（测试集）有很好的拟合结果（泛化能力），所产生的测试误差被称为泛化误差。度量泛化能力的好坏，最直观的表现就是模型的过拟合（Overfitting）和欠拟合（Underfitting），过拟合和欠拟合是用于描述模型在训练过程中的两种状态。一般来说，训练过程会是如图所示的一个曲线图。

过拟合是指训练误差和测试误差之间的差距太大，模型复杂度高于实际问题，模型在训练集上表现很好，但在测试集上却表现很差。模型对训练集"死记硬背"（记住了不适用于测试集

的训练集性质或特点），没有理解数据背后的规律，泛化能力差。

相对过拟合欠拟合容易理解。欠拟合一般是由于训练样本被提取的特征比较少，样本数理不足，导致训练出来的模型不能很好地匹配算法模型，数据预测表现得很差，甚至样本本身都无法高效的识别。

1. 机器学习数据集的划分方式以及为什么要划分不同的数据集。

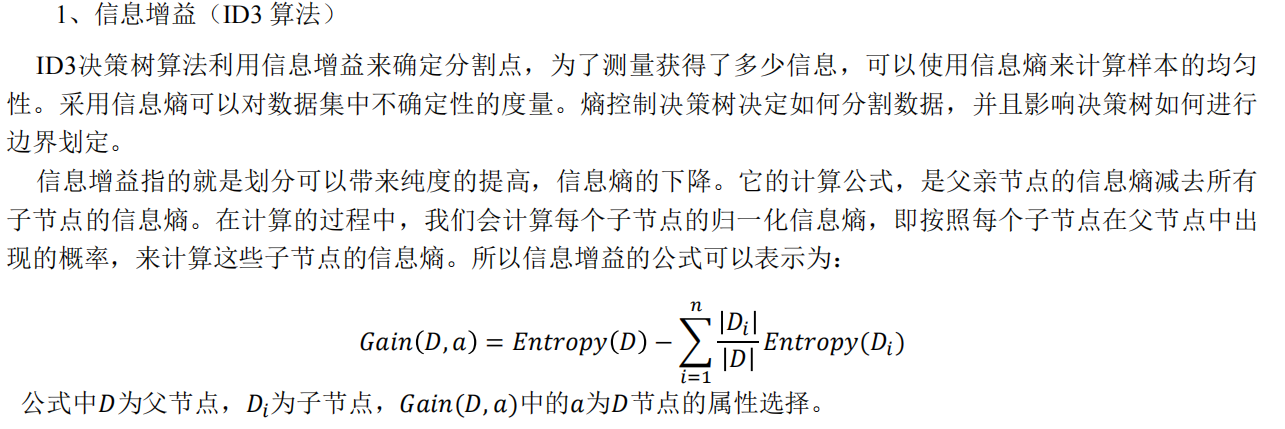
在机器学习中，我们通常将一个完整的数据集人为划分为训练集、验证集和测试集三个部分。三个数据集分别作用于机器学习算法进行训练、预测和评估的不同阶段。

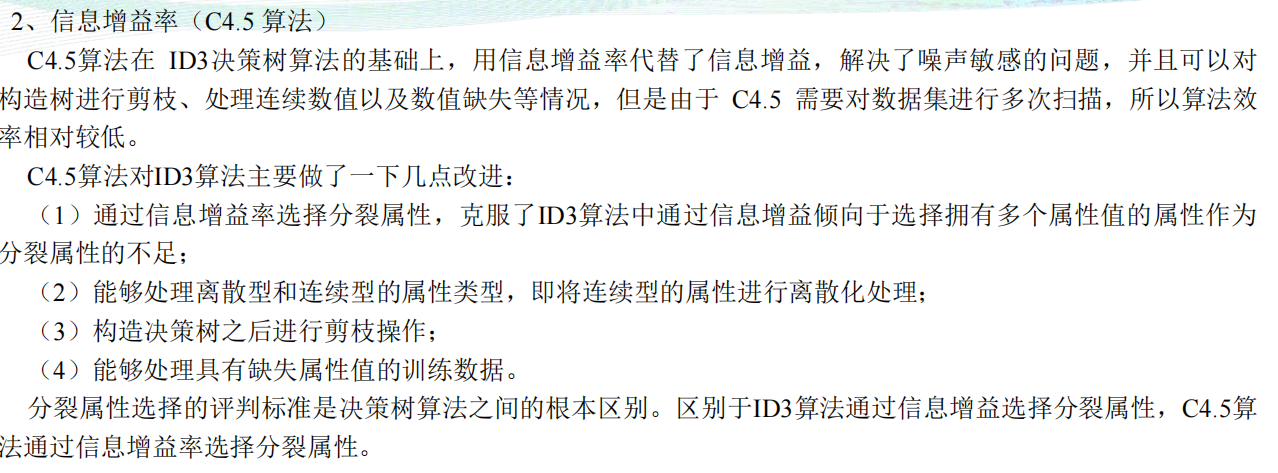
训练集（Train set）——用于训练模型，拟合的数据样本。

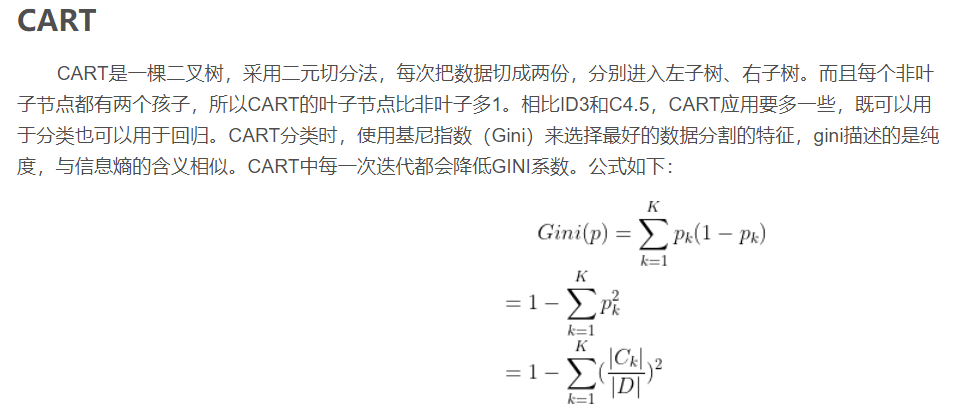
验证集（Development set）——是模型训练过程中单独留出的样本集，它可以用于调整模型的超参数和优化算法。

测试集（Test set）—— 对于已经构建和训练好的机器学习模型进行各方面的性能评估。

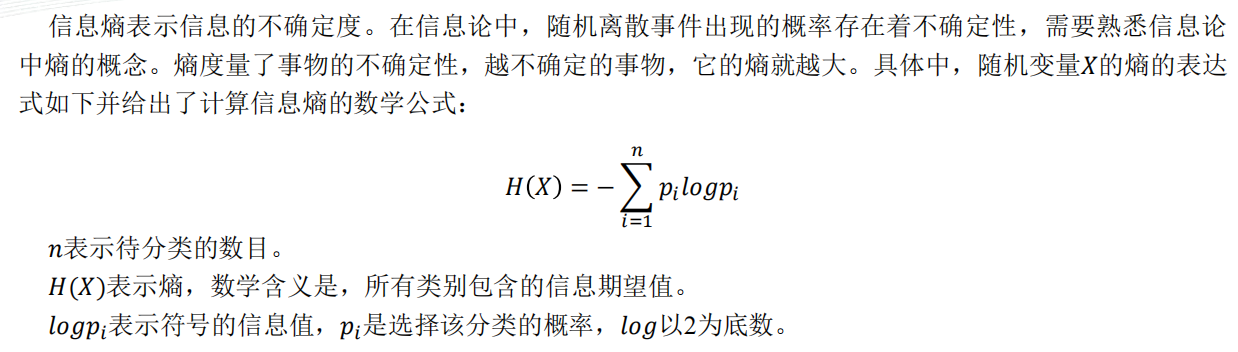
1. 构建决策树的三种不同方法。

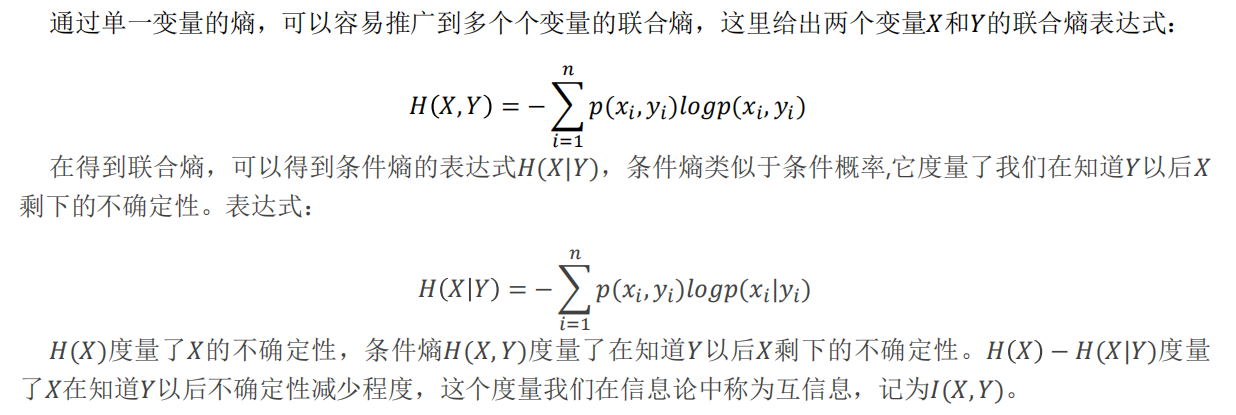


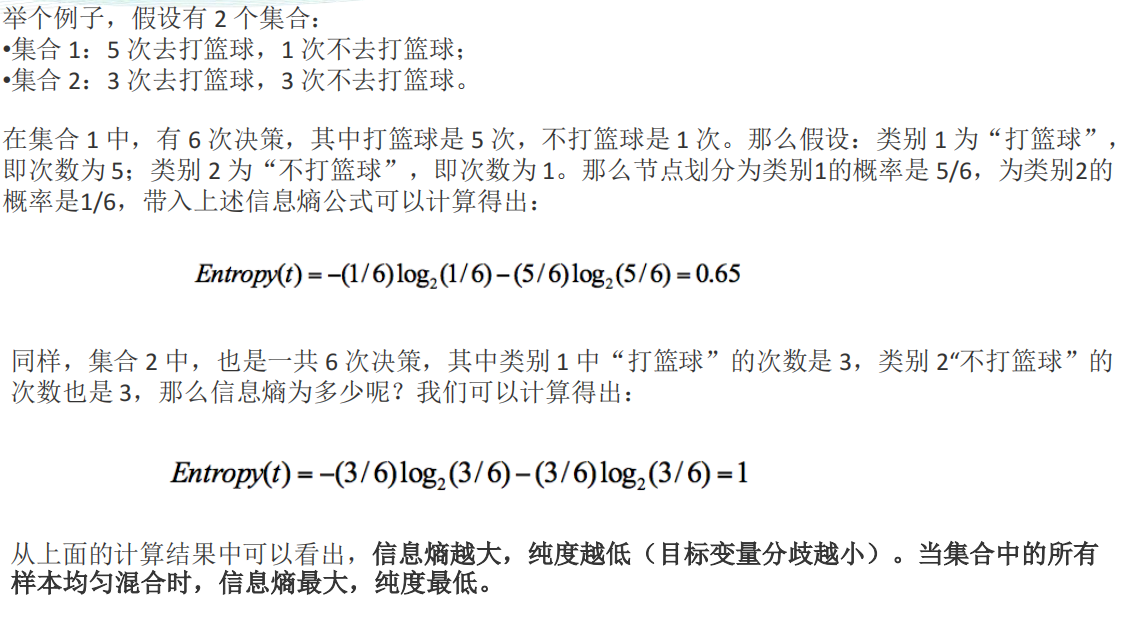




1. 如何计算信息熵，会使用信息熵计算的基本公式。







1. 数据降维的目的，常用的两种数据降维方法是哪两种。

**数据降维目的：**在高维数据空间中，往往包含有冗余信息以及噪音信息，在实际应用中例如图像识别中造成了误差，降低了分析的准确率。在机器学习问题中，有时需要在高维特征空间(每个特征都能够取一系列可能值)的有限数据样本中学习一种“自然状态”（可能是无穷分布），要求有相当数量的训练数据含有一些样本组合，但是数据维度并不是越高越好，通常在给定固定数量的训练样本，其预测能力随着数据维度的增加而减小，这就是机器学习中的“维度灾难”。而通过数据降维技术,我们希望减少高维数据中冗余信息所造成的误差，提高识别（或其他应用）的准确度，同时希望通过降维算法来寻找数据内部的本质结构特征，方便我们对数据进行进一步的处理与分析。

**主成分分析（Principal components analysis，PCA）**是目前最重要的降维方法之一。在数据压缩消除冗余和数据噪音消除等领域都有广泛的应用。主成分分析是一种数学降维方法, 利用正交变换 (orthogonaltransformation)把一系列可能线性相关的变量转换为一组线性不相关的新变量，也称为主成分，从而利用新变量在更小的维度下展示数据的特征。

**线性判别式分析 (Linear Discriminant Analysis, LDA) ，**也叫做Fisher线性判别 (Fisher LinearDiscriminant ,FLD)，是模式识别的经典算法，它是在1996年由Belhumeur引入模式识别和人工智能领域的。LDA是一种有监督的（Supervised）线性降维算法。与PCA保持数据信息不同，LDA是为了使得降维后的数据点尽可能地容易被区分，使得样本之间在新空间中投影后的类内方差最小，类间方差最大。

1. 层次聚类两种不同聚类类型。

层次聚类是最简单和较为直观的一类聚类方法。层次聚类方法是将数据对象组成一个聚类树。根据层次分解顺序分为自顶向上或是自顶向下完成。

1.凝聚层次聚类：自底向上

首先将每个对象作为一个簇，然后合并这些原子簇为越来越大的簇，直到某个终结条件被满足。

2.分裂层次聚类：自顶向下

首先将所有对象置于一个簇中，然后逐渐细分为越来越小的簇，直到达到了某个终结条件。

1. 交叉验证的三种方式。

1.如果我们只是对数据做一个初步的模型建立，不是要做深入分析的话，简单交叉验证就可以了。例如我们前面实验中，从所有数据样本中随机选取80%作为训练集，剩下的20%的数据作为测试集。

2.LOOCV方法（Leave-one-out cross-validation留一交叉验证）

LOOCV方法包含将数据集分为训练集和测试集，只用一个数据作为测试集，其他的数据都作为训练集，并将此步骤重复N次（N为数据集的数据数量）。

3.K折交叉验证(K-fold Cross Validation)

和LOOCV的不同在于，我们每次的测试集将不再只包含一个数据，而是多个，具体数目将根据K的选取决定。比如，如果K=3，那么我们利用三折交叉验证的步骤就是：1.将所有数据集分成3份； 2.不重复地每次取其中一份做测试集，用其他两份做训练集训练模型，之后计算该模型在测试集上的MSE； 3.将3次的MSE取平均得到最后的MSE。

1. 数据标准化的意义。

数据的标准化（normalization）是将数据按比例缩放，使之落入一个小的特定区间。在某些比较和评价的指标处理中经常会用到，去除数据的单位限制，将其转化为无量纲的纯数值，便于不同单位或量级的指标能够进行比较和加权。其中最典型的就是数据的标准化处理，即将数据统一映射到[0,1]或者[-1,1]区间上。数据标准化的目的就是让不同维度之间的特征在数值上有一定比较性，可以大大提高分类器的准确性

1. 如何计算两个点之间的欧几里得距离，会使用欧几里得距离的基本公式。

