**机器学习**部分考试复习重点：

（一）试卷题型（闭卷考试）

单项选择题 30分，填空题15分，简答题25分，

计算题20分，程序分析题10分。（满分100分）

（二）复习重点

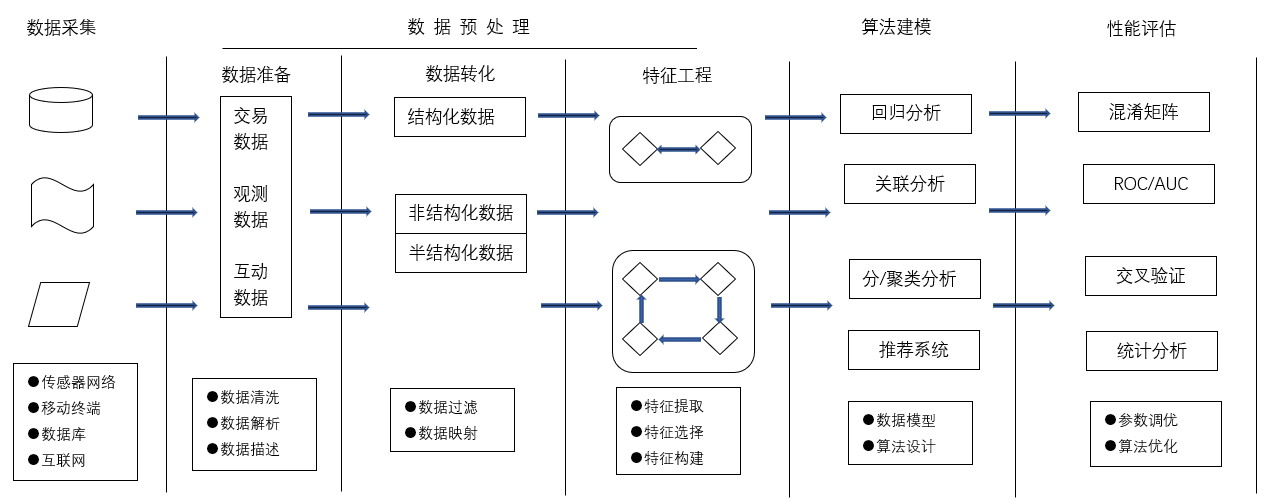
1. 机器学习这门课程的名称（中英文）

**机器学习(Machine Learning ML)**

1. 机器学习的概念。

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

1. 机器学习的基本流程。



1. 有监督和无监督算法的意义。

**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

1. 机器学习中的数据集的划分和划分的意义。

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。

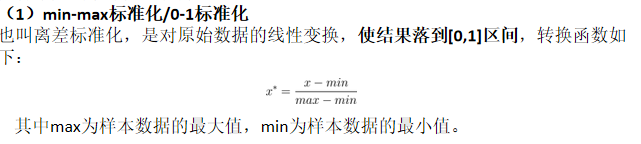
训练集（Train set）——用于训练模型，拟合的数据样本。

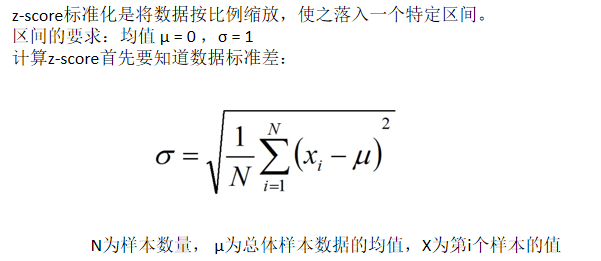
验证集（Development set）——是模型训练过程中单独留出的样本集，它可以用于调整模型的超参数和优化算法。

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. 数据标准化的意义和两种不同的方法。

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归一化处理，即将数据统一映射到[0,1]或者[-1,1]区间上。





1. 交叉验证三种方式。

**交叉验证(Cross-Validation)来避免模型过拟合。**

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。

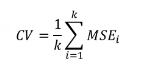
2.**LOOCV**方法（Leave-one-out cross-validation**留一交叉验证**）

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. 3.将3次的MSE取平均得到最后的MSE。



1. 模型拟合的三种方式。

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。**拟合合适（Just right）**

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. 线性回归算法的基本步骤。

通常在回归问题中，我们可以总结回归算法的的常规步骤为：

（1）寻找假设函数（Hypothesis函数），即y\_t；

（2）构造J函数,即损失函数；

（3）求解使得J函数最小并求得回归参数(ω∗,b∗)



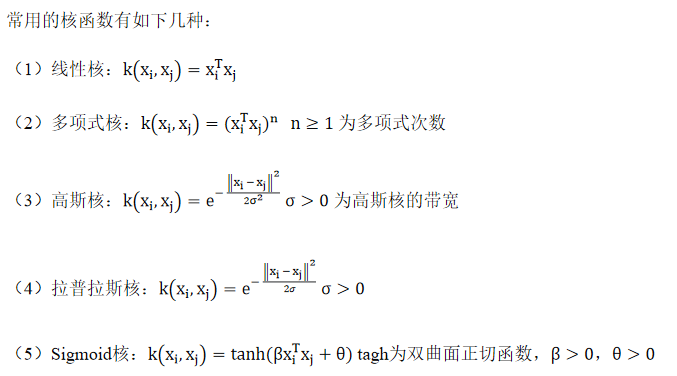
1. 决策树算法中生成决策树的三种算法。

决策树的构建算法主要有**ID3、C4.5、CART**三种，其中ID3和C4.5是**分类树**，CART是**分类回归树**。ID3是决策树最基本的构建算法，C4.5和CART是在ID3的基础上进行的优化算法。

1. 随机森林算法的基础算法。

**随机森林**是一种基于Bagging类型的集成学习算法。随机森林的基础是决策树，其通过随机的方式将多颗相互之间没有关联的决策树进行组合，构成森林。

1. 支持向量机算法中的几种核函数。



1. KNN算法的名称。

**KNN(k-nearest neighbor classification 最近K邻居)**

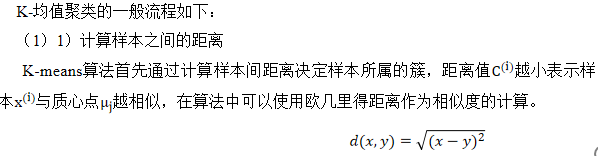
1. 目前常用的两种数据降维的算法。

**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 (Orthogonal Transformation)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)** 是一种有监督的（Supervised）**线性降维算法**。PCA算法一般只能用作数据降维，LDA是一种有监督的降维算法，其既可以用于数据降维，也可以用于数据分类。

与PCA保持数据信息不同，LDA的思想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. 欧几里得距离的计算。

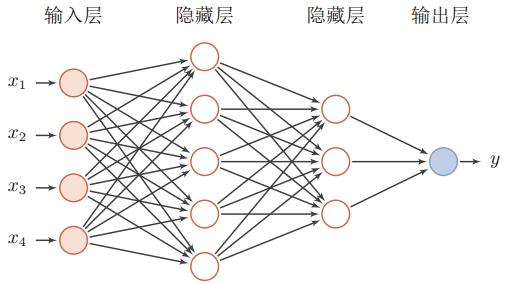


**a(2,3) b(4,5)**



1. 神经网络的基本结构。

在神经网络结构中，如图所示，**输入层**只从外部环境接收信息，是由输入单元组成，而这些输入单元可接收样本中各种不同的特征信息。每个圆圈，都代表一个神经元，也叫节点(Node)，只为下一层传递信息；**隐藏层**介于输入层和输出层之间，这些层完全用于分析，其函数联系输入层变量和输出层变量，使其更配适数据。最后，**输出层**生成最终结果，每个输出单元会对应到某一种特定的分类，为网络送给外部系统的结果值，整个网络由调整链接强度的程序来达成学习的目的。



1. 朴素贝叶斯算法的基本要素。

贝叶斯的数学公式十分简单，主要包含**先验概率P(A)、似然性P(B|A)**，以及最终得到**后验概率P(A|B)**，这三者是构成贝叶斯统计的**三要素**。

●先验概率是指根据常识、生活经验所观测到的“原因”的概率。

●后验概率是在知道“结果”之后，去推测“原因”的概率。

●似然函数是根据已知结果去推测固有性质的可能性（likelihood），是对固有性质的拟合程度。

1. 层次聚类的2种类别。

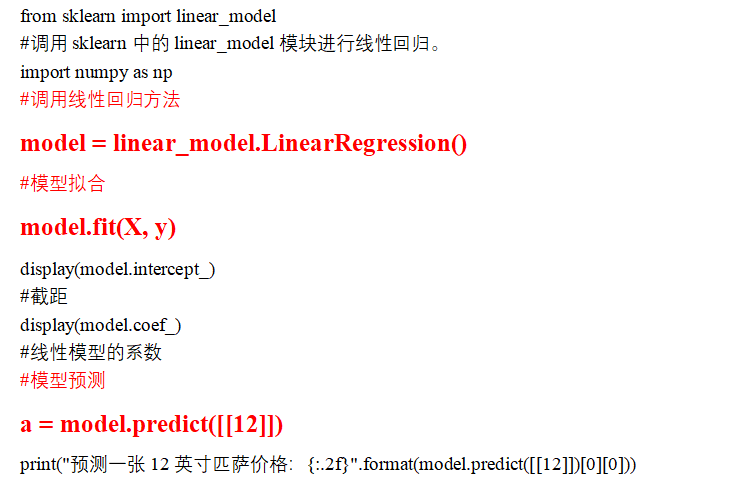
**层次聚类(Hierarchical Clustering)**是最简单的一类聚类算法，通过计算不同类别数据点间的相似度来创建一棵有层次的嵌套聚类树。在聚类树中，不同类别的原始数据点是树的最底层，树的顶层是一个聚类的根节点。创建聚类树有自下而上合并和自上而下分裂两种方法。层次聚类算法根据层次分解的顺序分为：**自下向上（凝聚型层次聚类 agglomerative）**和**自上向下（分裂型层次聚类divisive）**。自下而上法就是一开始每个个体（object）都是一个类，然后根据链接关系(linkage)寻找同类，最后形成一“类”。自上而下法正好相反，开始时所有个体都属于一“类”，然后根据链接关系进行分类，最后每个个体都成为一“类”。

1. DBscan算法应用的场合。

我们前面介绍的层次聚类和k-means聚类算法只能处理球形的簇，也就是一个聚成实心的团（这是因为算法本身计算相似度距离，如欧几里得距离的局限）。但往往现实中还会有各种形状，比如下面两张图，**环形和不规则形**，这个时候，那些传统的聚类算法显然就混乱了。于是就出现了新的聚类算法，基于密度的聚类算法，代表性的就是DBSCAN聚类算法。

1. 理解线性回归和决策树算法程序，重点理解哪些语句是进行算法模型训练（拟合），哪些语句是进行数据预测的。

**线性回归**



**决策树**

