# 绪论

1.1 将音乐软件的推荐问题视作一个机器学习问题，如果用户喜欢这首音乐会点击“喜欢”按钮或收藏。请根据Tom Mitchell对机器学习的定义写出该问题的E、T、P分别代表什么？

参考答案：经验E是指推荐给用户的音乐与用户的反馈等历史数据，任务T是指推荐音乐这，P是指用户对推荐的音乐喜欢与否。

1.2 简述分类任务与回归任务的区别，并给出应用举例。

参考答案：输入变量与输出变量均为连续变量的预测问题是回归问题，如判断是否为垃圾邮件；输出变量为有限个离散变量的预测问题成为分类问题，如预测房价；

1.3 假设某一数据集有50条数据、三个属性值，每个属性值分别有3、4、6种取值，请计算该样本空间的大小。

参考答案：由题意，共有（3+1）\*（4+1）\*（6+1）+1=140种。

# 模型评估与选择

2.1 简述经验误差与泛化误差的区别。

参考答案：经验误差指学习器在训练集上的误差，泛化误差指的是学习器在新样本上的误差。

2.2 简述如何通过ROC和AUC这两个指标判定学习器的优劣。

参考答案：若一个学习器的ROC 曲线被另一个学习器的的曲线完全‘包住’，则后者的性能优于前者；若两曲线交叉，则无法判断；当两个学习器的ROC 曲线交叉时，通过比较AUC 来判断两个学习器的性能优劣。AUC 值越大，则学习器性能越好。

2.3 如下表所示，表中共有20个测试样本，‘class’一栏表示每个测试样本真正的标签（p表示正样本，n表示负样本），‘score’表示某分类器中每个测试样本被预测为正样本的概率。

|  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Inst | Class | Score | Inst | Class | Score | Inst | Class | Score | Inst | Class | Score |
| 1 | N | 0.35 | 6 | N | 0.58 | 11 | P | 0.74 | 16 | P | 0.4 |
| 2 | P | 0.34 | 7 | P | 0.80 | 12 | P | 0.73 | 17 | N | 0.39 |
| 3 | N | 0.33 | 8 | N | 0.62 | 13 | N | 0.72 | 18 | P | 0.38 |
| 4 | P | 0.30 | 9 | P | 0.77 | 14 | P | 0.505 | 19 | N | 0.37 |
| 5 | N | 0.10 | 10 | P | 0.72 | 15 | N | 0.71 | 20 | N | 0.36 |

令阈值为0.7，计算该分类器的错误率、精度、查准率、查全率与F1.

参考答案：注意阈值设置为0.7，注意到这一点，计算部分相对简单。

错误率

精度

查准率

查全率

F1-Score

# 线性回归

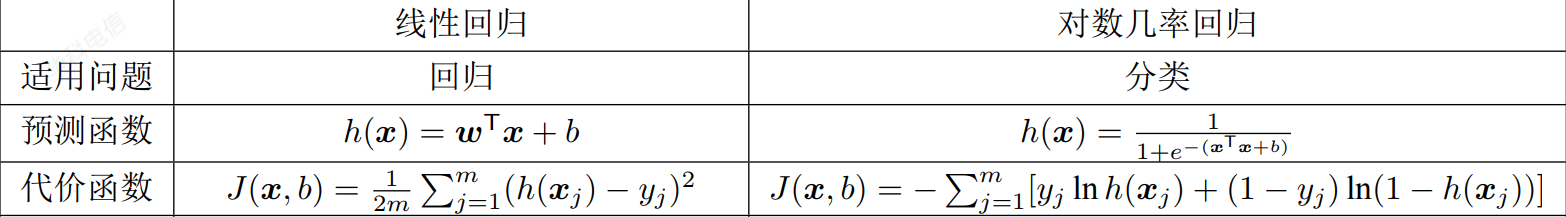
3.1 请写出sigmoid函数的表达式，大致画出函数的图像。

参考答案：



3.2 试从适用问题、预测函数、代价函数等方面比较线性回归与对数几率。

参考答案：



3.3 对于函数,给定初始出发点,利用梯度下降法求极值。

（1）设置和.

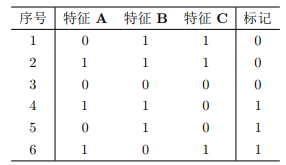
（2）设置和.

参考答案：

# 决策树

4.3 考虑下面的训练集：共计6个训练样本，每个训练样本有三个维度的特征属性和标记信

息。详细信息如表1所示。



请通过训练集中的数据训练一棵决策树，要求通过“信息增益”(information gain)为准则来选择划分属性并给出详细的计算过程并画出最终的决策树。

参考答案：

首先计算各个特征的信息增益：

由于特征A和特征C的信息增益相同且为最大，因此选取A或C为划分属性。以A为例，在A的基础上进行属性划分，得到信息增益：

因此B、C任选一个均可。

反复迭代，可得到决策树如下所示：

# 神经网络

# 支持向量机

# 贝叶斯分类器

# 集成学习

# 聚类