# 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

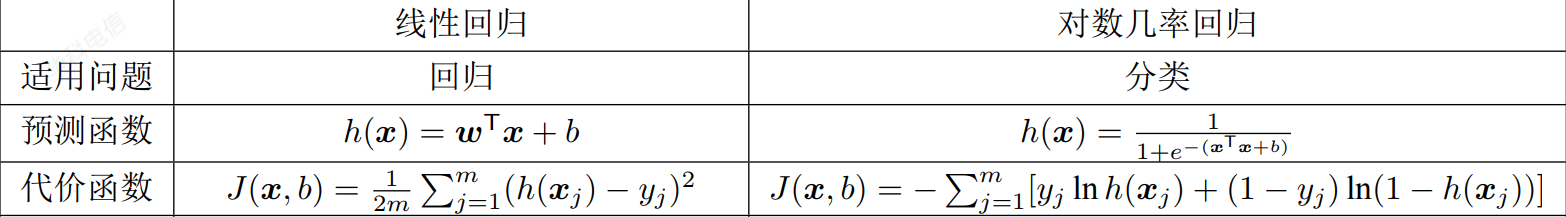
3.1 请写出sigmoid函数的表达式，大致画出函数的图像。

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

4.1 根据数据构建决策树后为什么要进行剪纸处理？

参考答案:

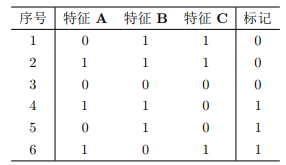
1. 决策树充分考虑了所有的数据点，有可能出现过拟合的情况，决策树越复杂，过拟合的程度会越高。
2. 剪枝修剪分裂前后分类误差相差不大的子树，能够降低决策树的复杂度，降低过拟合出现的概率。

4.2 如何判断决策树泛化性能是否提升？

参考答案：采用留出法，预留一部分数据用作“验证集”进行评估。以西瓜数据集为例，我们将西瓜数据集划分为训练集和验证集, 在接下来的例子中，我们采用训练集训练模型，用验证集评估模型的性能。如果通过某种方式(如属性划分) 提升了模型在验证集上的性能（如验证集准确率），我们则认为该方法使模型的泛化性能得到了提升。

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(只画其中一种)。

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，分析发下当选取A为根节点时，B、C的信息增益总是相同，因此只需BC即可完成决策树的构建。即

1 0

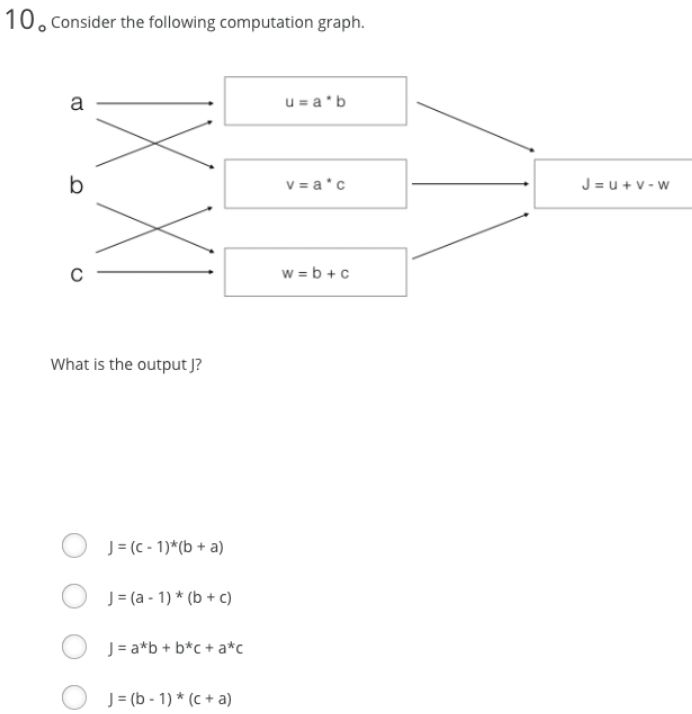
1 0

1 0

0 1 1 0

# 神经网络

5.1 观察如下所示的框图，写出输出J的表达式。



参考答案：J=u+v-w=a\*b+a\*c-（b+c）=a\*（b+c）-（b+c）=（a-1）\*（b+c）

5.2 对BP多层前馈神经网络的不足进行简要概括。

参考答案：

1） 训练过程易陷入局部极小值，从而得不到全局最优解；

2） 计算量大，训练次数多，使得学习效率低，收敛速度慢；

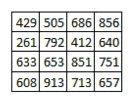
3） 对于隐含层个数和隐含层神经元节点的个数选择，至今还没有一个具体的定论，缺乏理论指导；

4） 训练时，学习新样本有遗忘旧样本的趋势。

5.3 完成下列卷积计算



参考答案：

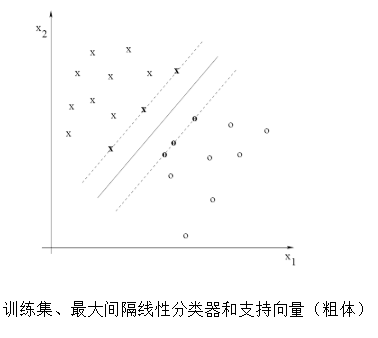


# 支持向量机

6.1 现有一个点能被正确分类且远离决策边界。如果将该点加入到训练集,为什么SVM的决策边界不受其影响,而已经学好的logistic回归会受影响?

参考答案：因为SVM采用的是hinge loss,当样本点被正确分类且远离决策边界时,SVM给该样本的权重为0，所以加入该样本决策边界不受影响。而logistic回归采用的是log损失，仍然会给该样本一个小小的权重。

6.2 根据下图所示，若采用留一交叉验证得到最大间隔分类器的预测误差估计是多少？（用样本数表示即可）



参考答案：从图中可以看出，去除任何点都不影响SVM的分界面。而保留所有样本时，所有的样本都能被正确分类，因此LOOCV的误差估计为0。

# 贝叶斯分类器

7.1 简要概述先验概率、条件概率、后验概率的含义。

参考答案：

先验概率：是指根据历史的资料或主观判断所确定的各种事件发生的概率，该概率没有经过实验证实，属于检验前的概率。

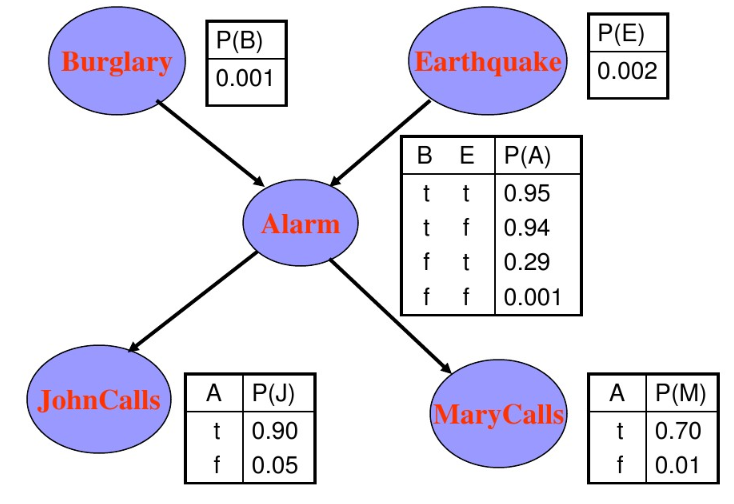
条件概率：一个事件发生后另一个事件发生的概率。一般的形式为P(x|y)表示y发生的条件下x发生的概率

后验概率：事件发生后求的反向条件概率；或者说，基于先验概率求得的反向条件概率。概率形式与条件概率相同。

7.2 概述引入拉普拉斯修正的原因及作用。

参考答案：拉普拉斯修正避免了因训练集样本不充分而导致概率估值为零的问题。当训练集变大时，修正过程所引入的先验的影响也会逐渐变得可忽略，使得估值逐渐趋于实际概率值。

7.3 某贝叶斯网络如图所示



求当MaryCall发生时，Alarm发生的概率。

参考答案：由题意，

# 集成学习

8.1 解释集成学习中基学习器与组件学习器的联系与区别。

参考答案：基学习器与组件学习器都是集成学习中的个体学习器。每个个体学习器采用相同的学习算法，称之为基学习器；个体学习器可以采用不同的学习算法，此时每个个体学习器称之为组件学习器。

8.2 简述bagging采用的自助采样法的优点与缺点。

参考答案：

优点：适用于数据集较小、较难划分训练/测试集；能从初始数据集中产生不同的训练集，对集成学习等方法有很多好处。

缺点：产生的训练集改变了初始数据集的分布，引入估计偏差。

# 聚类

9.1 对于聚类算法的性能度量指标，分为外部指标与内部指标，阐述两类指标分别的含义。

参考答案：

􀽌􄜞外部指标：将聚类结果与某个“参考模型”进行比较。

􀞻􄜞内部指标：直接考察聚类结果而不利用任何参考模型。

9.2 阐述K均值算法的主要思想。

参考答案：把n个点划分到k个聚类中，使得每个点都属于离它最近的聚类中心。