第三章 经典算法

1.支持向量机

在空间上线性可分的两类点，分别向SVM分类的超平面上做投影，这些点在超平面上的投影仍然是线性可分的吗？

投影后不是线性可分的。对于任意线性可分的两组点，它们在SVM分类的超平面上的投影都是线性不可分的

是否存在一组参数使VM训练误差为0？

当所有样本的类别都被正确预测，训练误差为0。

训练误差为0的SVM分类器一定存在吗？

是

加入松弛变量的SVM的训练误差可以为0吗？

不一定

2.逻辑回归

逻辑回归和线性回归，有何异同？

（1）不同：逻辑回归处理分类问题，线性回归处理的是回归问题

（2）相同：都使用了极大似然估计来对训练样本进行建模（线性回归使用最小二乘法，实际上就是在自变量x与超参数θ确定，因变量y服从正太分布的假设下，使用极大似然估计的一个化简）

（3）相同：在求解超参数的过程中，都可以使用梯度下降法

使用逻辑回归处理多标签的分类问题时，有哪些常见做法，分别应用于哪些采场景，它们之间有怎样的关系？

首先，如果一个样本只对应一个标签的话，可以假设每个样本属于不同标签的概率服从几何分布，使用多项逻辑回归来进行分类（Softmax Regression）

如果样本可能属于多个表亲的话，可以训练k个二分类的逻辑回归分类器，第i个分类器用以区分每个样本是否可以归为第i类，训练该分类器时，需要把标签重新整理为“第i个类标签”与“非第i个类标签”两种情况。

3.决策树

决策树作为最基础、最常见的有监督学习模型，常被用于**回归**和**分类**问题。将决策树应用集成学习的思想可以得到随机森林、梯度提升决策树等模型。
决策树的生成包含了特征选择、树的构造、树的剪枝三个过程。

决策树有哪些常用的启发函数?
ID3、C4.5、CART

ID3-最大信息增益（混乱减少的程度），会倾向于选择取值较多的特征，只能处理离散型变量，对样本缺失值敏感，只可用于分类
C4.5-最大信息增益比，通过引入信息增益比，一定程度上对取值较多的特征进行惩罚，可以处理连续型变量，可以对缺失值进行处理，只可用于分类
CART-最大基尼系数，可以处理连续型变量，可以对缺失值进行处理，可以用于回归和分类

ID3和C4.5可以在每个节点上产生多叉分支，且每个特征在层级之间不会复用。CART每个节点只会产生两个分支，因此最后会形成一棵二叉树，且每个特征可以被重复使用。

如何对决策树进行剪枝?
1.预剪枝，即在生成决策树的过程中提前停止树的生长：
（1）当树到达一定深度的时候，停止树的生长。
（2）当到达当前节点的样本数量小于某个阈值的时候，停止树的生长，不再继续扩展。
（3）计算每次分裂对测试集的准确度的提升，当小于某个阈值的时候，不再继续扩展。
算法简单、效率高，但有欠拟合的风险，虽然在当前的划分会导致测试集准确率下降，但在之后的划分中，准确率可能会显著上升。

2.后剪枝，让算法生成一颗完全生长的决策树，然后从最底层向上计算是否剪枝：
著名的是CART的CCP方法 参考书P.68