##### QG工作室数据挖掘小组实验报告

|  |  |  |
| --- | --- | --- |
| 实习生： 黄倬熙 | 导师： | 日期：2020年7月23日 |

|  |
| --- |
| 实验名称：基于周志华《机器学习》进行数据挖掘理论学习，并且尝试实现其中的算法 |
| 已完成内容：   1. 支持向量机的推导以及手动实现 2. 随机森林的手动实现 3. 复习已学习的算法（任务内的之前都学过） |
| 未完成内容：暂无 |
| 未完成原因：暂无 |
| 需要帮助：暂无 |

|  |  |
| --- | --- |
| 实验总结 | |
| 知识点总结：  支持向量机：   1. 算法目的：找到一个分离超平面，能将实例分到不同的类。间隔最大化保证了了超平面的唯一性，即不仅将正负实例点分开，对离超平面最近的点，也有足够大的确信度将它们分开。 2. 何为支持向量：训练数据集的样本中，与分离超平面距离最近的样本点实例。其他点去掉也不影响解。 3. 如何求解超平面：构建拉格朗日函数，使带约束问题转化为无约束->令拉格朗日函数取最大值，找出支持向量，再从样本点中找最大间隔->转换对偶问题->结合KKT条件，求解W\*和b\*->使用SMO算法求解λ\*。 4. 为何转换对偶问题：使问题易于求解以及引入核函数。 5. 如何处理非线性问题：把低维数据映射到高维空间，使数据线性可分。 6. 为何使用核函数：求出映射函数较为繁琐，使用核函数可简化计算，且核函数结果与代入映射函数的结果相等。 7. 如何处理噪声：引入松弛变量，将硬间隔最大化修改为软间隔最大化。   随机森林：   1. 实现：设定筛选样本数目N（N<样本总数）和筛选特征数目a（a<特征数目）->设定常数X为决策树数目，并创建X个决策树（创建时，每个决策树随机选定数目为a的条件，随即选定数目为N的样本）->创建完成后，带入测试样本，以结果少数服从多数或平均值作为最终测试结果，并与真实结果做对比，判断是否符合实际。 2. 优点：处理高维度数据切不用做特征选择；训练结束能给出哪个特征比较重要；训练速度快；实现简单 3. 为何不需要剪枝：剪枝是为了避免过拟合，随机森林随机选择变量与树的数量，已经避免了过拟合，没必要去剪枝。 4. 为何要有放回的抽样：保证样本集间有重叠，若不放回，每个训练样本集及其分布都不一样，可能导致训练的各决策树差异性很大，最终多数表决无法 “求同”，即最终多数表决相当于“求同”过程。 | |
| 遇到问题：   1. 推导支持向量机的过程繁琐复杂，需要消耗大量时间完成 2. 手动实现算法过程中比较难理解SMO算法 | **解决过程：**   1. 结合《机器学习》《统计学习方法》以及网上的博客，B站白板推导视频，以及坚持不懈的面对它。 2. 使用Python中的cvxopt库，可轻易解决凸优化问题，但是静心理解还是能读懂SMO算法的。 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 导师评价 | | | | |
| 实验分数 | **知识掌握情况** | **代码编写能力** | **建议** | **评价日期** |
|  |  |  |  |  |