##### QG工作室数据挖掘小组实验报告

|  |  |  |
| --- | --- | --- |
| 实习生： 黄倬熙 | 导师： 谢光强 | 日期：2020年 8月 9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. 整理，优化所学数据挖掘算法 2. 学习，推导并实现高斯混合模型算法 3. 学习，推导并实现k近邻回归算法 4. 学习，推导并实现自适应提升算法 5. 学习，推导并实现梯度提升树算法 |
| 未完成内容：暂无 |
| 未完成原因：暂无 |
| 需要帮助：暂无 |

|  |  |
| --- | --- |
| 实验总结 | |
| 整理算法：   1. 把所有算法和分为监督学习和无监督学习 2. 所有算法接口统一，参照sklearn标准 3. 函数归类，放在Utils文件夹中，方便以后直接导入   高斯混合模型（GMM）：   1. 应用：通常用于解决同一集合下的数据包含多个不同的分布的情况（或者是同一类分布但参数不一样，或者是不同类型的分布，比如正态分布和伯努利分布）。 2. 参数估计过程：引入了一个新的 k 维变量 z （z为一个one-hot vector），用来更加方便地表示选中某个分类的概率。其中 z 的取值范围为 0 或者 1，zk = 1 表示第 k 类被选中的概率，zk = 0 表示第 k 类未被选中的概率。 3. 参数求解：EM算法 4. 求解步骤：第一轮随机给出分布的参数->根据参数计算出最大似然->根据似然重新计算参数->根据参数变动大小不断动态更新参数直到收敛 5. 似然函数求解遇到的问题：出现log求和的情况，无法求解 6. 解决办法，适用Jensen不等式。设X是一个随机变量，f(X)是一个凸函数（二阶导数大或等于0），那么有：。当且仅当X是常数的时候等号成立，如果f(X)是凹函数，不等号反向。   K近邻回归：   1. 计算目标点和所有训练集点的距离 2. 排序，找出K个最近点 3. 计算K个点的均值作为预测值 4. K值如何确定？使用交叉验证遍历k值，找出回归效果最好的一类。 5. 如何降低时间复杂度？使用KD树搜索最近的K个点   自适应提升（AdaBoost）：   1. 实现思路：初始化训练数据的权值分布。第一步假设具有相同的权值分布->使用当前分布加权的训练数据集，学习基本分类器->计算基本分类器在加权训练数据集上的分类误差率->计算基本分类器的系数->更新训练数据集的权值分布 2. 优点：泛化错误率低，容易实现，可以应用在大部分分类器上，无参数调整 3. 缺点：对离散数据点敏感 4. Boosting适用的基学习器：树模型【1）容易将样本的权重带入训练中；2）可以通过剪枝等方式灵活地调整基学习器的表达能力和泛化能力；3）对特征工程的要求少；4）可解释性强】线性模型【线性模型符合boosting对基学习器的高偏差、低方差要求】神经网络【1）可以通过调整神经元个数、层数、连接方式等灵活地调整模型的表达能力和泛化能力；2）属于不稳定的学习器，对数据分布比较敏感】   梯度提升树（GBDT）：   1. 实现思路：初始化n棵CART回归树->赋予初始值（一般为均值）->遍历n颗树，每棵树不断拟合损失函数的梯度，并不断更新预测值->预测：第一棵树的预测值减去其他树的预测值 2. 损失函数的选择：与bp类似，回归用误差平方和，分类用二元交叉熵 3. 与AdaBoost的联系与区别：gbdt和Adaboost都是基于boosting思想的集成学习。gbdt使用对数损失，通过拟合代价函数的导数提高模型性能；Adaboost使用指数损失，根据学习器输出结果的错误率调整样本、基学习器的权重来提高模型性能。除此之外，Adaboost是一个模型计算框架，理论上可以使用任意基学习器，而gbdt是以决策树为基学习器的模型。 4. 与随机森林的联系与区别：gbdt都是以决策树为基学习器的集成学习模型。gbdt基于boosting思想，每轮在上一轮模型的残差上训练得到一个二叉回归树，通过降低偏差提高模型性能，最后输出所有树的累加；随机森林基于bagging思想，用样本的M个自助抽样子集，独立地训练M个回归/分类树，通过降低方差提高模型性能，输出预测均值/多数投票结果。 5. 优点：1）能灵活处理多种数据类型，且无须进行标准化等特征工程工作；2）在相对较少的调参时间下，能得到较高的性能；3）有xgboost等优秀的开源实现，节省许多编码、部署工作 6. 缺点：1）树之间需要串行训练；2）通常不适用于高维稀疏问题 | |
| 遇到问题：   1. 尝试实现GBDT的时候，是要导入自己写的CART回归树的，但是由于接口不规范造成很大的困扰。 2. EM算法开始学习的时候难以理解 | **解决过程：**   1. **引申出想要优化整理自己所实现的所有算法，包括归类，函数分类，接口统一等。** 2. **由于没有学习概率论，有的东西确实难以理解，只能慢慢的搜知识点去理解了，但是这样也是比较耗时** |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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** |
|  |  |  |  |  |