数据挖掘与大数据分析

Assignment 4

1. 数据集（10 分）

- 从 UCI dataset repository 中下载以下数据集
  - （5 分）IRIS
  - （5 分）Breast Cancer Wisconsin (Diagnostic) Data Set

  下载以后，仔细阅读数据集的使用说明，理解其用途及每一列数据的含义。

2. 随机森林与 AdaBoost（50 分）

- （25 分）分类任务：分别对两个数据集按照自己设定的比例进行训练集、测试集的划分，使用训练集分别训练随机森林模型跟 AdaBoost(基分类器采用决策树模型) 分类器，并分别用测试集测试其性能；
- （25 分）回归任务：忽略两个数据集中的类别属性，从其余属性中任选一个作为回归任务的目标属性；分别对两个数据集按照自行设定的比例进行训练集、测试集的划分，使用训练集分别训练随机森林跟 AdaBoost 回归器，并分别用测试集测试其性能。

这两个模型均可直接使用 Scikit-learn 包中的实现。

3. Stacking 方法（20 分）

对 Breast Cancer 数据集按照自己设定的比例进行训练集、测试集、验证集的划分，训练一个 Stacking 模型。第一阶段的基础分类器与第二阶段的元分类器均自行选取，至少包含 3 个基础分类器，所有分类器均可直接使用 Scikit-learn 包中的实现。

4. 撰写技术报告（20 分）

以科技论文的形式撰写 assignment 的技术报告。

- 自行设计实验，达到以下目的

  - 对于分类任务，对比三个模型在同样数据集上的性能（使用合适的指标），并对结果进行分析；
  - 对于回归任务，对比随机森林与 Adaboost 在同样数据集上的性能（使用合适的指标），并对结果进行分析；
  - 对于回归任务，体现不同大小的训练集对回归器性能的影响（无需交叉验证）。
- 实验部分应对数据集进行介绍，参考文献中给出该数据集的原始出处并在报告正文中第一次出现给数据集的地方添加对文献的引用；
- 对实验结果的呈现，必须以文字形式进行阐述、解释或者说明，不能只是简单地展示结果的图，否则会减分；调整图的大小，使之清晰美观，否则会减分；
- 报告应以正规的书面语言进行客观的阐述，切勿使用口语化的表达方式或使用随意的网络用语；
- 插图应使用矢量图，图、表要添加编号与标题，并在正文中引用其编号；
- 报告中对使用的算法应引用其出处的参考文献，引用格式为用方括号括起来的上标数字形式，按引用的次序依次顺序编号，并在报告末尾添加“参考文献”一节；每一条文献条目中至少应包括作者名，文章标题，期刊名，期号，卷号，出版年月，pp：页码范围，DOI 号或官网的 URL。

5. 必须提交的材料

- 下载的数据集：各个数据集各自存入一个文件中，文件名为程序中使用该数据集时的名称；
- python 的源程序：每个源程序存入一个文件，文件名能体现其作用；
- pdf 版本的技术报告；
- 以上三部分压缩成一个压缩包，以学号 + 姓名对压缩包进行命名。