数据挖掘与大数据分析实验 2

1. 数据集（20 分）

- 使用正弦函数生成一个包含两个正弦周期的数据集（振幅可自行设定），从中均匀采样 20 个数据样本，对每个样本的目标变量 y_i 添加一个随机的扰动值（扰动值不要太大），形成数据集 D_1；（10 分）
- 从 UCI dataset repository 中下载一个适合用于回归分析的数据集，满足以下要求：
  - 至少包含三列以上连续的数值型数据；（5 分）
  - 至少包含 100 个以上的样本；（5 分）

  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. 数据预处理（10 分）

- 【选做】选定一种标准化处理方法，将所下载数据的每一列进行标准化处理，使所有列的数据处于同一规模；
- 对下载的数据集，根据数据集的用途及每一列的物理含义，选取其中一列作为目标变量 $\boldsymbol{y}$，选取其它至少两列作为自变量 $\boldsymbol{x_1}, \boldsymbol{x_2}, \cdots$，形成数据集 D_2；（10 分)

3. 回归分析（50 分）

- 一元多项式回归（25 分）

  变换多项式的阶数 $m(m = 1, 2, \cdots, 5)$，对每一个 m 将数据集 D_1 按照 $|D_{train}| : |D_{test}| = 80\% : 20\%$ 的比例进行划分，用训练集 D_{train} 进行训练确定回归系数，用测试集 D_{test} 进行测试，获取 MAE 和 RMSE 值。
- Ridge 回归或 Lasso 回归（25 分）
  - 选取 Ridge 回归模型或 Lasso 回归模型，将 D_2 全部用作训练集，变换正则化系数 λ 的取值，确定回归系数，获取正则化路径数据，从中确定稳定的超参数 λ 的取值；（15 分）
  - 将 D_2 按 $|D_{train}| : |D_{test}| = 80\% : 20\%$ 的比例随机进行划分，用训练集 D_{train} 对选定的模型进行训练（使用刚才确定的 λ 值），重新确定回归系数，用测试集 D_{test} 进行测试，

获得 MAE 和 RMSE 值；重复这一过程 5 次以上，获取多组 MAE 和 RMSE 值。（10 分）

4. 实验报告（20 分）

实验报告应包含实验目的、相关算法介绍（不能照抄、照搬文献、网页中的文字，需要大家对相关方法进行提炼、总结，并用书面语言重新进行描述）、实验过程论述、数据集介绍、实验结果及分析、结论与思考等内容，实验报告的最后应该列出引用的文献。

- 对于一元多项式回归的结果：画出生成数据集的正弦曲线及采样并添加扰动的数据点，并画出 m 的不同取值得到的拟合曲线；画出不同的 m 值对应的 MAE、RMSE 的条形图，m 的每一个取值对应的 MAE、RMSE 构成一组；针对这两张图，对其展示的结果进行分析、解释；
- 对于 Ridge 回归或 Lasso 回归的结果：用折线图画出正则化路径，对其进行分析，讨论如何确定超参数 λ 的取值；对选定的 λ，画出多组 MAE、RMSE 的条形图，每一次对 D_2 进行划分得到的 MAE、RMSE 构成一组，并在最后添加一组 MAE、RMSE 的平均值，并针对该图进行分析、解释；
- 实验部分应对数据集进行介绍，说明使用了什么样的方式人工合成了数据集 D_1，从哪里下载了一个数据集并进行了怎样的处理得到 D_2；
- 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，如果是用 Wps/Word 书写，则插图应该转成.emf 或者.wmf 格式，使其在缩放时不失真；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

- 生成的数据集、下载的数据集及预处理以后的数据集：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（不要使用 Jupyter notebook）：按照正弦曲线生成数据集采样并添加扰动的源程序、数据预处理的源程序、实现回归的源程序，各自存入一个文件，文件名能体现其作用；
- pdf 版本的实验报告；

- 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。