**实验题目：**基于随机森林的缺失值预测

**实验目的：**了解随机森林算法的工作原理，认识算法超参数对算法性能的影响，掌握利用随机森林算法进行缺失值预测的方法；

**实验内容：**

1. **利用随机森林进行分类**
   1. 数据集描述Breast Cancer Wisconsin (Diagnostic) Data Set，包含569个样本，30维特征，分两类，见表1。

表 1 WDBC 数据集

|  |  |  |  |
| --- | --- | --- | --- |
| 数据集名称 | 样本数量 | 特征维数 | 类别 |
| WDBC | 569 | 30 | 2类，212,357 |

Txt文件描述：每一行代表一个样本，每一列代表一维特征，具体信息：

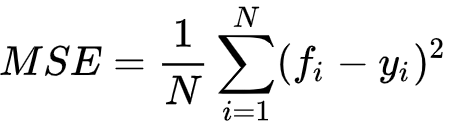
第1列：患者编号，无用信息，应删去；

第2列：类别编码，M代表恶性，B代表良性，这一列作为样本的标签；

第3-32列：30维特征。

* 1. 要求：分析决策树的数量与分类准确率的关系；分析特征采样数与分类准确率的关系。
  2. 实验方法：数据集划分，分成训练集和测试集；利用训练集训练RF模型，利用OOB评估模型性能，最后用测试集验证；分别统计OOB错误率和测试集分类错误率，绘制错误率与b）中两个超参数的关系曲线，并尝试分析之。

1. **利用随机森林进行缺失值预测**
2. 按照一定比例，随机将WDBC数据集中的特征置0，利用RF方法进行缺失值预测；
3. 缺失比例：10%，20%，30%，40%，50%；
4. 评价指标为均方误差，预测值与原始值的差的平方和的均值：



1. 绘制不同缺失比例下，RF模型的性能变化曲线，并分析之。
2. **编程软件建议：Python，MATLAB。**