《机器学习与模式识别》上机实验指导书

实验1：实验环境与基本概念熟悉

**实验目的：**

* 了解Python基础知识，并安装Anaconda环境用于运行Python程序；
* 熟悉KNN分类原理，了解影响KNN分类结果的主要因素；
* 熟悉模型拟合中的欠拟合和过拟合现象，并可以通过调整参数进行改进。

**实验内容：**

**（1）K-近邻算法（KNN）实验**

根据实验提供的KNN代码文件（KNN.ipynb）开展实验：

* 1. 使用jupyter或其他环境运行KNN代码；
  2. 合理组合以下几种参数的不同值：数据集（例如iris, wine, breast cancer）、采样率（sample rate = 1, 0.5, 0.2等）和近邻范围K（K=1, 3, 5, 10等），实验并观察各种参数组合下KNN分类算法的精度变化情况，试分析产生相关结果的原因；
  3. （选做题）改进现有代码，提高KNN算法的实现性能。例如，提高大规模训练数据下的分类速度，在保证精度的前提下增强算法的鲁棒性，提供自适应的参数选择策略等。
  4. 实验报告：记录不同参数组合下的分类精度，绘制分类精度随参数变化的效果图，并结合实验结果给出参数选取的实验建议。对于选做题，改进内容不限，但需要进行必要的文字说明和实验效果验证

**（2）基于高阶多项式回归的欠拟合和过拟合分析实验**

根据实验提供的Regression代码文件（Regression.ipynb）开展实验：

1. 使用jupyter或其他环境；代码功能为：使用一元M次多项式回归拟合正弦函数（最小二乘法）。数据点数固定为200个。
2. 合理组合以下几种参数的不同值：信号频率（1、2、4、8），噪声方差0.1、0.2、0.3、0.5，每趟实验同时观察4种阶数的多项式的拟合情况M=10、30、50、180。观察各种参数组合下多项式回归对信号与噪声的拟合程度，分析产生欠拟合和过拟合现象的一些条件；
3. 实验报告：记录有明显拟合效果差异的参数组合，记录参数与对应的拟合图形，说明拟合效果与条件的关系，并就如何避免模型欠拟合和过拟合提出想法。

**参考资料：**

**scikit-learn user Guide, Release 0.23.1（May 19, 2020）.pdf**

**补充说明：**

**（1）代码中对应代码处有注释，请根据注释开展实验**

**（2）使用最小二乘法拟合曲线**

