# Project 1 算法实验基础

## 目标

1. 掌握常见的机器学习任务：分类任务、或回归任务
2. 掌握常用的机器学习算法，如线性模型、决策树、神经网络、支持向量机等，不仅包括算法的思想和具体过程，还包括算法的适用性、优缺点、参数（作用、敏感性、参数调节方法等）
3. 掌握数据预处理的常用方法，如离散值、连续值的转换方法，缺失值的处理等等
4. 深入了解一种机器学习算法平台，如[Weka](https://www.cs.waikato.ac.nz/ml/weka/)等；
5. 能够调用算法，掌握参数选择方法
6. 掌握机器学习实验方法的基础，包括实验方法（如留出法、交叉验证法等）、统计检验方法（如t检验、Friedman检验等）
7. 掌握算法分析和比较的方法和过程、包括结果的记录与展示，实验分析等
8. 撰写实验报告

## 参考教材

* I. H. Witten, E. Frank, and M. A. Hall. Data Mining: Practical Machine Learning Tools and Techniques, 3rd edition. Morgan Kaufmann Publishers, 2011. （3rd  Ed. in Chinese. 4rd Ed. has been published.）

## 内容

1. **算法基础**

任务：

1. 针对分类任务，选择1种学习算法（可以是已学过的任何算法）
2. 对于每种任务类型，选择一个数据集，建议UCI数据集（可以到UCI官网下载，也可以到Weka的数据集页面下载），了解数据集背景、属性、样本数等各个方面的情况
3. 对数据集进行预处理（关于数据集预处理，可以参考Weka教材）
4. 掌握算法思想与过程、推导等
5. 了解算法的参数，设置合适的参数
6. 选择合适的性能评价准则
7. 选择一种实验方法（留出法、交叉验证法、bootstraping方法等），对算法的泛化性能进行估计
8. **多数据集多算法比较实验**
9. 选取分类任务的10个数据集
10. 对比至少3个算法(可是是前6章，以及第7-11章与大作业不重复章节的算法，也可以是领域前沿算法)
11. 注意要对算法进行参数选择，或者设置合理的默认参数。  
    注意：至少要有一个算法需进行参数选择。对使用默认参数的算法，需指明默认参数的取值，以及说明合理性，必要时，需进行参数敏感性分析实验。
12. 使用t-test+sign-test假设检验对算法进行两两对比 或 使用Friedman test等假设检验方法对多个算法进行对比
13. 总结实验结果：对实验结果进行分析并得出相应的结论，比如哪个算法最优，各算法的优缺点，算法对参数的敏感性等。
14. **报告（提交时间：第15周周日1月10日24:00）**

注意：不需要将每次实验课内容全部报告。只需报告第2次实验课内容。

实验报告格式：

1. 限制在10页之内
2. 学习任务以及性能度量指标
3. 数据集，将数据集信息整理成表格形式，包括样本数，属性数，几个类别，每个类别包含多少样本等信息。不需要对数据集背景信息情况逐一详尽说明。
4. 数据集预处理
5. 对比算法及参数设定（或参数选择方式），以及代码来源（使用何种工具箱或算法包，或者是自己实现）。对于算法，如果是教材上没有的算法，要给出简要的介绍，比如多类的神经网络，它的网络结构有何要求，学习的目标函数是什么等需要介绍的内容
6. 实验方法：包括采用何种实验方法，何种假设检验方法等
7. 结果分析：使用图、表等方式。若涉及算法的参数敏感性分析，需要给出一张图。

## 评分指标

1. 任务说明/性能度量：指明是何种类型的任务（本次实验限定在分类问题，但不限二类或多类分类问题）及性能度量标准。
2. 数据集介绍： 用表格描述数据集信息
3. 数据预处理
4. 实验方法说明：10x10cv或其他方法
5. 参数设置：对每个算法进行参数说明（说明使用软件的默认参数也算）。
6. 参数选择及选择方法：对进行参数选择的算法进行参数选择说明，包括参数选择范围与步长等。对使用默认参数的算法，需指明默认参数的取值，以及说明合理性，必要时，需进行参数敏感性分析实验
7. 总体结果：使用表格给出，统计均值及方差。结果保留3位小数
8. 假设检验：使用方法及检验结果（两种方式，t-test+sign-test, Friedman test）。
9. 结果分析：结果分析的全面性、结果展示的丰富性（比如有图等）。
10. 若有参数敏感性分析实验，则包含参数敏感性分析结果及结论
11. 报告质量（表述、合适长度、简练程度）：超过页数限制的扣分。

注意：请在截止时间内提交。超期减分。