《机器学习》课程实验实践要求

准备阶段（全体成员）：

1. 在 UCI 上下载相关数据集，使用 python 对数据进行预处理，包括：归一化，缺失值，OneHot 编码等。

2. 查询算法的评价指标计算方法（大部分 python 自带），包括：

评估指标解释模型的性能

3. 查询算法实验中需要注意的问题：如，实验结果需要多次实验取平均值，同时，**多种算法的性能对比时需要固定测试集**；分类数据的类别**数据不平衡问题**时的处理方法；缺失值以及离群值处理的方法等。

4. 查询实验结果的展示内容：如，多次实验的结果**对比图形**---展示稳定性方面；估计值与真实值的对比图形；降维可视化等。

5. PPT 展示时长限定在 30 分钟以内。

实践阶段(分组完成)：

实验一：线性回归算法的比较（第 1、2 组完成，第 4 周讲解展示）：

(1)生成模拟数据,加入共线性性,对比分析不同算法（OLS, RR, LARS）的性能；

(2)使用真实数据,对比 RR 和 LARS 算法的不同正则化参数对评价指标的影响；

(3)生成模拟数据，研究 Lars 算法变量选择的稳定性。

实验二：支持向量机算法实践（第 3、4 组完成，第 4 周讲解展示）：使用真实分类与回归数据，考察不同模型参数对评价指标的影响，包括高斯核函数的尺度参数，多项式函数的阶数；正则化参数的影响。

实验三：集成算法实践 1（第 5、6 组完成，第 5 周讲解展示）：

（1）AdaBoost 算法：不同基函数的效果比较；

（2）Random Forest：在分类和回归数据集上对比实验，同时提取重要特征，与 LARS 算法对应的重要特征进行对比。

实验四：集成算法实践 2（第 7、8 组完成，第 5 周讲解展示）：真实数据上 GBDT 算法与 XGBOOST 算法的比较。

实验五：贝叶斯学习算法实践（第 9、10 组完成，第 6 周讲解展示）：

实验六：聚类算法实践（第 11、12 组完成，第 7 周讲解展示）：

（1）K-means 算法和层次聚类算法的比较以及各自参数对结果的影响；

（2）基于密度的几个聚类算法的性能比较

（3）噪声情况下对聚类的影响和解决办法探讨.

实验七：降维算法实践（第 13、14 组完成，第 8 周讲解展示）：PCA,KPCA,LLE 降维算法的探索实验。

实验八：CNN 算法实践（第 15、16 组完成，第 9 周线上讲解展示）：卷积神经网络在图像识别中的探索实验（我会提供数据），包括卷积算子的选择问题，层数的影响探索等。