**数据挖掘课程作业——分类问题**

**一、数据集：**

在UCI（http://archive.ics.uci.edu/ml/datasets.php）中自选数据集，解决一个分类（Classification）问题。例如：

UCI葡萄酒数据集（https://archive.ics.uci.edu/ml/datasets/Wine+Quality）：

Data Set Information:

The two datasets are related to red and white variants of the Portuguese "Vinho Verde" wine. For more details, consult: [Web Link] or the reference [Cortez et al., 2009]. Due to privacy and logistic issues, only physicochemical (inputs) and sensory (the output) variables are available (e.g. there is no data about grape types, wine brand, wine selling price, etc.).

These datasets can be viewed as classification or regression tasks. The classes are ordered and not balanced (e.g. there are many more normal wines than excellent or poor ones). Outlier detection algorithms could be used to detect the few excellent or poor wines. Also, we are not sure if all input variables are relevant. So it could be interesting to test feature selection methods.

Attribute Information:

For more information, read [Cortez et al., 2009].

**Input variables (based on physicochemical tests):**

1 - fixed acidity

2 - volatile acidity

3 - citric acid

4 - residual sugar

5 - chlorides

6 - free sulfur dioxide

7 - total sulfur dioxide

8 - density

9 - pH

10 - sulphates

11 - alcohol

**Output variable (based on sensory da**ta):

12 - quality (score between 0 and 10)

**二、课程作业要求：**

1. 从以下方法中任选一种进行实现：决策树（ID3或C4.5），SVM，朴素贝叶斯。

2. 内容包括：

* 数据集的描述
* 数据集的预处理过程
* 训练数据集和测试数据集的划分
* 分类器的构建
* 代码实现与结果可视化
* 结论与展望
* 源代码（作为附录）

3. 撰写报告，在12月24日之前提交。