第一次小作业：SVM

在本次作业中，需要手动实现一个简单的 SVM 二分类器，并使用给定数据集评估分类器的性能。有关 SVM 的基础知识可参考课上 PPT 文件及讲义。

作业要求：

1. 提交截止日期：2019 年 11 月 26 日 23:00，命名格式："学号_姓名_第一次作业.zip"，提交至课程中心。
2. 建议语言：Python 3.5 或以上版本。
3. 软件包要求：模型训练和优化部分不允许调用现有的软件包，但允许使用 numpy/scipy 等类似工具包进行数学运算。对于数据读入、划分、预处理等则不做要求。
4. 算法要求：使用序列最小优化算法（SMO）作为参数训练算法。
5. 模型评估：课程提供训练数据集与评价指标。数据集包含特征和标签，可自行使用交叉验证等方法对模型预测性能进行评估与调优，注意防止过拟合。
6. 模型保存：由于作业包含课上测试部分，实验课上将公布测试集进行现场测试，建议实现模型保存和读取部分的代码，并将训练完成的模型参数提前保存，以免实验课上现场训练出现耗时过长的情况。
7. 提交内容：包含两部分，即相关代码文件和说明文档。说明文档需包含两个部分 i）阐明如何通过运行提交的代码文件，完成从数据输入到预测结果输出的过程，并尽量保证结果的可复现性；ii）对代码中各部分的作用进行简要介绍。

数据集和评价指标

本次作业采用的数据集为某分类数据集，原始数据包含 14 个变量，变量的组成如下：

- id：数据的唯一识别码，int 型
- label：待预测变量（1：正类，-1：负类）
- 12 个与调查对象相关的统计特征，命名为 x1-x12，对应变量类型为
  - 类别属性：
    - 标称类型 Nominal：x2, x5, x6, x7, x8, x9
    - 序数类型 Ordinal：x4
  - 数值属性：
    - 比率类型 Ratio：x1, x3, x10, x11, x12

模型预测性能的评价指标为 F1 分数。记 label 为 1 的样本为正样本，label 为-1 的样本为负样本，则 F1 的计算方法如下：

$$F1 = \frac{2 \times precision \times recall}{precision + recall}$$

其中$precision = \frac{TP}{TP+FP}, recall = \frac{TP}{TP+FN}$，分别表示精确率和召回率；$TP$、$FP$、$TN$、$FN$分别对应真正例（预测为正样本且实际为正样本）、假正例（预测为正样本但实际为负样本）、真负例（预测为负样本且实际为负样本）和假负例（预测为负样本但实际为正样本）的预测样例数目。

本次作业要求使用 SMO 作为优化算法，以下提供了一些相关文献及博客内容，以供参考

1. 《支持向量机（五）SMO 算法》，这是一篇中文博客，比较详细地分析了 SMO 算法的运行过程。
   https://www.cnblogs.com/jerrylead/archive/2011/03/18/1988419.html
2. http://pages.cs.wisc.edu/~dpage/cs760/SMOlecture.pdf
3. Sequential minimal optimization: A fast algorithm for training support vector machines. SMO 算法的原始论文，提供了算法原理和性能分析等更加详细的资料。
   https://www.microsoft.com/en-us/research/publication/sequential-minimal-optimization-a-fast-algorithm-for-training-support-vector-machines/