HW2: Logistic 回归&SVM

1、 任务描述

请在 Pima Indians Diabetes Data Set（皮马印第安人糖尿病数据集）进行分类器练习。

需要提交代码文件，并给出必要的结果解释。

1) 训练数据和测试数据分割（随机选择 20%的数据作为测试集）;（10 分）
2) 适当的特征工程（及数据探索）;（10 分）
3) Logistic 回归，并选择最佳的正则函数（L1/L2）及正则参数；（30 分）
4) 线性 SVM ,并选择最佳正则参数 ,比较与 Logistic 回归的性能 ,简单说明原因。（20 分）
5) RBF 核的 SVM，并选择最佳的超参数（正则参数、RBF 核函数宽度）;（30 分）

2、 数据说明：

原始数据集地址：https://archive.ics.uci.edu/ml/datasets/Pima+Indians+Diabetes

数据集只有一个文件（**diabetes.csv**）：Pima Indians Diabetes Dataset 包括根据医疗记录的比马印第安人 5 年内糖尿病的发病情况，这是一个两类分类问题。每个类的样本数目数量不均等。一共有 768 个样本，每个样本有 8 个输入变量和 1 个输出变量。缺失值通常用零值编码。

1) 字段说明

Pregnancies： 怀孕次数

Glucose： 口服葡萄糖耐受试验中，2 小时的血浆葡萄糖浓度。

BloodPressure： 舒张压（mm Hg）

SkinThickness： 三头肌皮肤褶层厚度（mm）

Insulin：2 小时血清胰岛素含量（μU/ ml）

BMI： 体重指数（体重，kg /（身高，m）^ 2）

2) DiabetesPedigreeFunction： 糖尿病家族史

3) Age： 年龄（岁）

Outcome： 输出变了/类别标签（0 或 1，出现糖尿病为 1，否则为 0）