

实验任务书

时间：第 7、8 周周二上午 1-4 节

上机地点：B3-138

一、实验任务一：基于插值的数据扩充及不平衡分类问题。

二、实验目的：了解不平衡分类 (Unbalanced Classification) 问题及用插值方法对少数类样本进行扩充的基本原理。

三、具体任务：

- 1、了解数据挖掘 python 库 scikit-learn 的使用，参见 <https://scikit-learn.org/>。
- 2、从 UCI 机器学习数据库中 (<http://archive.ics.uci.edu/ml/>) 下载任务为 **Classification** 的类严重不平衡的数据集（至少 2 个数据集）。
- 3、用线性插值法对少数类样本进行扩充（参见插值法 PPT 中对 SMOTE 的简介，其详细介绍可以参见论文 SMOTE: Synthetic Minority Over-sampling Technique），然后使用决策树 (Decision Trees) 及贝叶斯分类器 (Naive Bayes) 对比扩充前后的分类准确率（参数可以多使用几个不同的值，评价指标使用 F1-measure）（用图表列出）。

一、实验任务二：基于多变量线性回归模型的数据拟合。

二、实验目的：掌握用多变量线性回归模型及最小二乘法进行数据拟合的基本原理。

三、数学原理回顾：

1、多维线性回归模型及最小二乘解

设多维线性模型 $y = a_0 + a_1x_1 + a_2x_2 + \dots + a_mx_m$ ，其中常数项 a_0 为偏置(bias)。

给定 n 个观测样本 $(\mathbf{x}^{(1)}, y_1), (\mathbf{x}^{(2)}, y_2), \dots, (\mathbf{x}^{(n)}, y_n)$ ，我们可以写成如下矩阵形式 $\mathbf{X}\boldsymbol{\theta} = \mathbf{y}$ ：

$$\begin{bmatrix} 1 & x_1^{(1)} & x_2^{(1)} & \dots & x_m^{(1)} \\ 1 & x_1^{(2)} & x_2^{(2)} & \dots & x_m^{(2)} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_1^{(n)} & x_2^{(n)} & \dots & x_m^{(n)} \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ \vdots \\ a_m \end{bmatrix} = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$$

其中：

\mathbf{X} ： $n \times (m+1)$ 维数据矩阵，每一行 \mathbf{x} 表示一个样本的数据(共 n 个样本)，每个样本有 m 个特征，但线性模型还带有 1 个常数偏置项 a_0 。

\mathbf{y} ： n 维列向量，表示 n 个观测目标值。

$\boldsymbol{\theta}$ ： $m+1$ 维列向量，表示线性模型的 $m+1$ 个系数。

(1) 模型参数的最小二乘解为：

$$\boldsymbol{\theta} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{y}$$

(2) 为了防止过拟合，常常使用一些正则化项，如 L_2 正则化以控制参数 $\boldsymbol{\theta}$ ，相应的最小二乘解如下，其中 λ 为标量超参数 (hyperparameter)， \mathbf{I} 为单位矩阵：

$$\boldsymbol{\theta} = (\mathbf{X}^T \mathbf{X} + \lambda \mathbf{I})^{-1} \mathbf{X}^T \mathbf{y}$$

四、具体任务：

1、从 UCI 机器学习数据库中(UCI 机器学习数据库：<http://archive.ics.uci.edu/ml/>) 下载任务为回归 (Regression) 的数据集 (至少 2 个)。

2、用多变量线性模型拟合数据，并计算测试集的平方误差和的平均值（用 5 折交叉验证）（用图表列出）。

3、用多变量线性模型拟合数据，但使用 L_2 正则项，重复上项实验。关于 λ 参数要进行多个参数的设置实验，并就有无正则项比较平方误差和的平均值（用图表列出）。

五、提交时间：第 10 周周一以前，实验占总评成绩的 10%，逾期未缴，此部分成绩为 0 分。

六、提交内容及方式：实验报告、源码的纸质版及电子版，其中电子版以学号+姓名+使用的数据集名命名，由班长统一提交（由班长压缩后发到邮箱 jbwang@scut.edu.cn）；如果太大（超过 50M）可能被学校邮箱拒收，可以通过网盘发送。