实验任务

一、实验题目：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 $y = a_0 + a_1x_1 + a_2x_2 + \ldots + a_mx_m$，其中常数项 a_0 为偏置(bias)。给定 n 个观测样本$(\mathbf{x}^{(1)}, y_1)$, $(\mathbf{x}^{(2)}, y_2)$, …, $(\mathbf{x}^{(n)}, y_n)$, 我们可以写成如下矩阵形式 $\mathbf{X\theta} = \mathbf{y}$：

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

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：

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

（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（hyperparamter）：

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

2、矩阵逆的计算：

矩阵行操作：交换两行；k 乘以某一行；某一行乘以 k 加到另一行。

当把矩阵 A 仅使用上述行操作转换为单位矩 I 时，则单位矩阵将转换为 A^{-1}：$[A\ I] \rightarrow [I\ A^{-1}]$。

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（数据集为几百条比较适宜，不要太小如只有几十个样本；但也不要太大，如几千上万条，内存可能溢出）。

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）、运行时间及内存统计（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误差（用图表列出）。

注：

（1）不能使用库函数求矩阵的逆，而是要自己编程实现。

（2）独立完成，若发现抄袭，则抄袭和被抄袭者同计 0 分。

五、提交时间：纸质版 12 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，其中电子班以学号+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）；如果太大可能被学校邮箱拒收，可以通过网盘发送。