第二次大作业 最小二乘方法应用

最小二乘方法在函数逼近、回归分析、数据拟合等邻域应用广泛，是最重要的方法之一。狭义上的最小二乘是指在给定二次误差函数$L(\mathbf{y},\hat{\mathbf{y}})=||\mathbf{y}-\hat{\mathbf{y}}||_2^2$下，利用在样本点上预测值与样本值间误差和$L=\sum_{i=1}^{N}L(\mathbf{y}_i,\hat{\mathbf{y}}(\mathbf{x}_i))$最小确定数学模型$\hat{\mathbf{y}}(\mathbf{x})$。最小二乘方法的使用前提是给定数学模型$\hat{\mathbf{y}}(\mathbf{x})$的函数形式$f(\mathbf{x},\theta)$，当此函数相对于模型参数$\theta$是非线性函数时，此问题就是非线性最小二乘问题，否则就是具有闭式解的线性最小二乘问题。近年以来，最小二乘问题有很多扩充，一类是针对模型参数θ进行约束或者限制，比如等式约束、稀疏性约束、能量最小化约束等，也有对拟合模型$\hat{\mathbf{y}}(\mathbf{x})$进行约束，比如光滑性约束等，这类问题可以转化为普通最小二乘问题或者利用迭代重加权最小二乘方法求解，另一类是优化参数具有多重线性，可以运用交替最小二乘方法求解。

本次大作业所要解决的问题是所谓协同滤波问题，这个问题的经典背景为：假设有 1000 个用户和 200 部电影，用户i对电影j的打分为M_{ij}，因此形成了一个1000×200的评分矩阵$\mathbf{M}$，由于每个用户不一定对所有电影都打过分，因此矩阵M不会被完全填充。所要解决的问题就是填充矩阵。

一般可以将所要填充的矩阵表示为：$\mathbf{M}=\mathbf{U}\mathbf{V}^T$，其中$M_{ij}=\mathbf{U}(i,:)\mathbf{V}(j,:)^T$，表示用户$i$对电影$j$的打分是用户$i$的隐变量$\mathbf{U}(i,:)$与电影$j$的隐变量$\mathbf{V}(j,:)$的内积。考虑到用户类型和电影类型有限，因此有理由相信评分矩阵$\mathbf{M}$是低秩的。由于矩阵的秩是矩阵奇异值向量的 0 范数，可以将其放缩到 1 范数，即矩阵的核范数。由于矩阵的核范数满足：

$$||\mathbf{X}||_*=\min_{\mathbf{U},\mathbf{V}|\mathbf{X}=\mathbf{U}\mathbf{V}^T}\frac{1}{2}(||\mathbf{U}||_F^2+||\mathbf{V}||_F^2)$$

因此上述问题可以通过优化下述目标函数进行求解：

$$\min_{\mathbf{U},\mathbf{V}}||\mathbf{W}*(\mathbf{M}-\mathbf{U}\mathbf{V}^T)||_F^2+\frac{\lambda}{2}(||\mathbf{U}||_F^2+||\mathbf{V}||_F^2)$$

其中λ是控制矩阵低秩程度的超参数。$\mathbf{W}$是标志矩阵，$W(i,j)=1$表示用户i对电影j已经打过分了，$W(i,j)=0$表示未打分，$*$表示矩阵对应元素相乘。上述优化问题

的基本想法是交替最小二乘，即固定$\mathbf{U}$以$\mathbf{V}$作为优化变量，固定$\mathbf{V}$以$\mathbf{U}$作为优化变量，交替地进行此过程求解。

另一种思路将矩阵的秩放缩到光滑 Schatten-p 函数，即：

$$f_p(\mathbf{X}) = \mathrm{tr}(\mathbf{X}^T\mathbf{X} + \gamma\mathbf{I})^{p/2}$$

其中$0 \leq p \leq 1$，γ为近似光滑性参数，一般取值很小，上式中先求$(\mathbf{X}^T\mathbf{X} + \gamma\mathbf{I})^{p/2}$（由矩阵函数定义）再进行求迹。当 $\gamma = 0$时，$f_1(\mathbf{X}) = ||\mathbf{X}||_*$，当$\gamma = 0$且$p \to 0$时，$f_p(\mathbf{X}) \to \mathrm{rank}(\mathbf{X})$。这样就可以采用下面的目标函数：

$$\min_{\mathbf{X}} f_p(\mathbf{X}) + \lambda||\mathbf{W} * (\mathbf{M} - \mathbf{X})||_F^2$$

此问题可以采用迭代重加权最小二乘方法求解，迭代重加权最小二乘方法基本思路是将函数$f_p(\mathbf{X})$在当前值下放缩到二次型函数，然后每步求解都是利用普通最小二乘方法的闭式解。

本次作业中给出里的数据尺寸是943×1682，其中有 90000 个数据已经给出，这 90000 个数据拆分为两部分 80000+10000，利用 80000 点作为训练数据，另外 10000 作为测试数据测试算法效果。在此之外还有 10000 个数据作为作业测试数据(未给出，助教以此评估算法)。测试指标是均方误差，即：

$$MSE = \frac{1}{|S|}\sum_{(i,j)\in S} ||M(i,j) - X(i,j)||^2$$

其中S为测试样本集合。

作业要求：

(1).作业报告完整，格式规范，对算法描述清楚。

(2).此问题有多种解决方法，也可以采用其他方法，但上述两种方法中必须择其一实现。

(3).报告中需要关注超参数选择、测试集上的均方误差、算法收敛时间等。

(4).最终版本需要提交报告和估计矩阵X(存储为 mat 格式文件)。

(5).考虑到作业提交时间已在考试周，希望同学们尽早着手，保证按时按质完成。

(6).此次作业每人不分组，每人独立完成。