第一次编程作业

刘子源 无研223 2022310709

一.数据预处理

分析数据，共有10000个样本，每个样本有12个变量，其中churn是取值为0和1的标签，其余11个变量为特征，我们要根据这11个变量来预测churn。

	customer_id	credit_score	country	gender	age	tenure	balance	products_number	credit_card
0	15634602	619	France	Female	42	2	0.00	1	1
1	15647311	608	Spain	Female	41	1	83807.86	1	0
2	15619304	502	France	Female	42	8	159660.80	3	1
3	15701354	699	France	Female	39	1	0.00	2	0
4	15737888	850	Spain	Female	43	2	125510.82	1	1
...	...	...	...	...	...	...	...	...	...
9995	15606229	771	France	Male	39	5	0.00	2	1
9996	15569892	516	France	Male	35	10	57369.61	1	1
9997	15584532	709	France	Female	36	7	0.00	1	0
9998	15682355	772	Germany	Male	42	3	75075.31	2	1
9999	15628319	792	France	Female	28	4	130142.79	1	1

观察发现，customer_id是无用的变量，需要将其去除。非数值的变量有国籍和性别。可知一共有3个国家：France、Spain和Germany，其样本量分别为5014、2477和2509；一共有2种性别：Male和Female，起样本量分别为5457、4543。对于balance等特征，不同国家的群体的均指和方差有较大的差异；对于estimated_salary等特征，男性群体和女性群体的均值方差也有差异。因此我们认为性别和国家是有信息量的特征，不应该去除，所以我将其进行编码，转化为数值变量。在数据归一化中，我尝试了均匀归一化、Z-score和不做处理三种方法，Z-score的方法得到的结果最好，后续实验均采用此方法。

	credit_score	country	gender	age	tenure	balance	products_number	credit_card	a
0	-0.003375	-1.086832	-2.200969	0.027985	-0.360181	-1.964483e-05	-1.567147	1.417293	
1	-0.004552	0.363244	-2.200969	0.018894	-0.479732	1.880593e-06	-1.567147	-3.395246	
2	-0.015899	-1.086832	-2.200969	0.027985	0.357121	2.136285e-05	4.344385	1.417293	
3	0.005189	-1.086832	-2.200969	0.000711	-0.479732	-1.964483e-05	1.388619	-3.395246	
4	0.021352	0.363244	-2.200969	0.037077	-0.360181	1.259169e-05	-1.567147	1.417293	
...	...	...	...	...	...	...	...	...	
9995	0.012896	-1.086832	1.832325	0.000711	-0.001530	-1.964483e-05	1.388619	1.417293	
9996	-0.014401	-1.086832	1.832325	-0.035655	0.596222	-4.909874e-06	-1.567147	1.417293	

二、公式推导

对逻辑回归公式推导如下：

定义σ为sigmoid函数

$$\sigma(x) = \frac{1}{1+e^{-x}} \tag{1}$$

记训练样本集为$\{x_n, t_n\}_{n=1}^{N}$，则有负对数似然函数

$$E(\omega) = -\sum_{n=1}^{N}\{t_n ln y_n + (1-t_n) ln(1-y_n)\} \tag{2}$$

加入正则项后得到损失函数

$$L = -\sum_{n=1}^{N}\{t_n ln y_n + (1-t_n) ln(1-y_n)\} + \frac{\lambda}{2}||\omega||^2 \tag{3}$$

对参数ω求导可得

$$\frac{\partial L}{\partial \omega} = \sum_{n=1}^{N}(y_n - t_n)\phi_n + \lambda\omega = \sum_{n=1}^{N}(\sigma(\omega^T x) - t_n)x_n + \lambda\omega \tag{4}$$

记学习率为α，则随机梯度算法为$\omega^{(k+1)} = \omega^{(k)} - \alpha\frac{\partial L}{\partial \omega^k}$

对MSE回归公式推导如下：

使用MSE作为损失函数，有

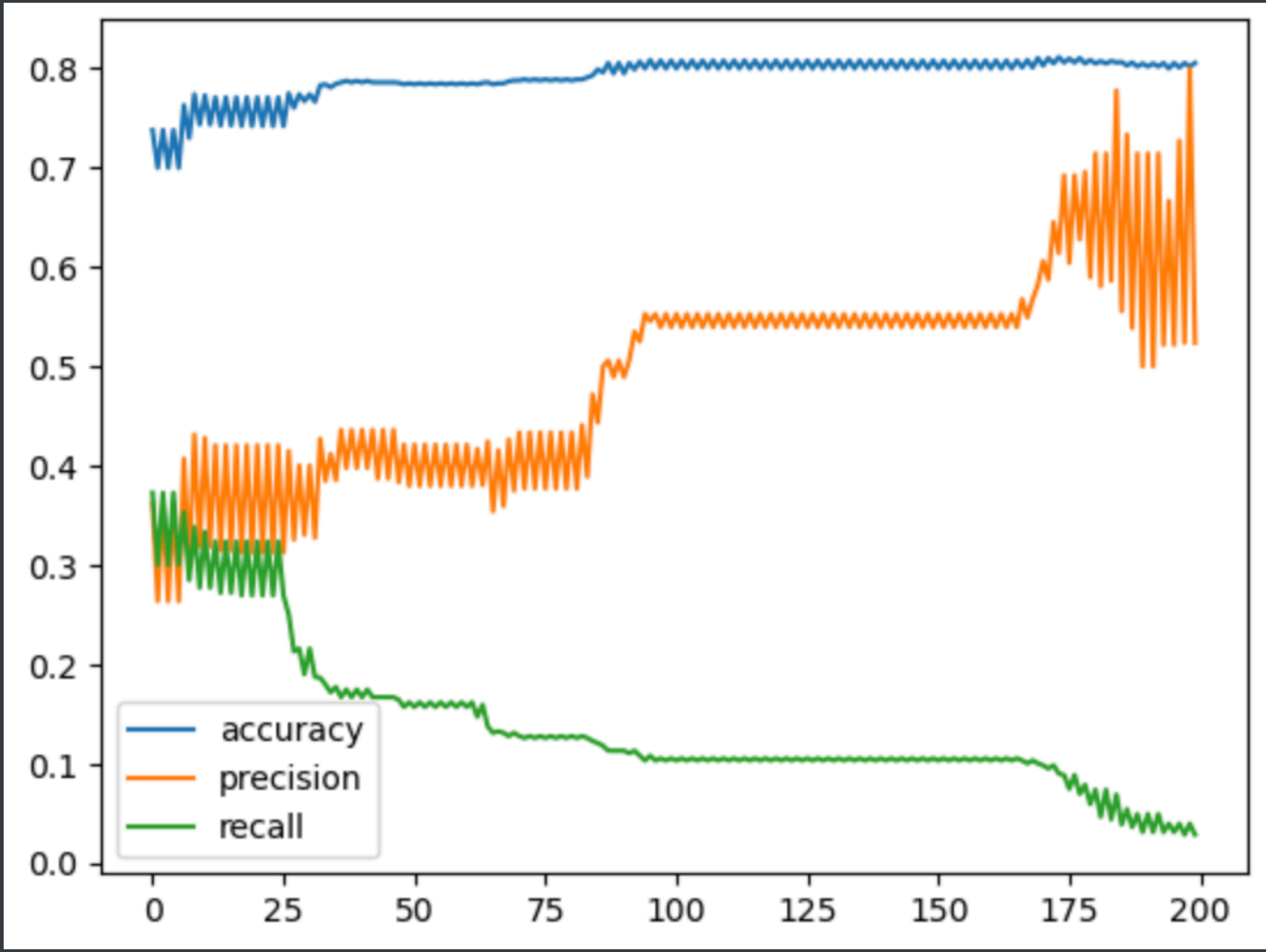
$$L = \frac{1}{2}\{\sum_{i=1}^{N}(\omega^T x_i - t_i)^2 + \lambda||\omega||^2\} \tag{5}$$

$$\frac{\partial L}{\partial \omega} = \sum_{i=1}^{N}(\omega^T x_i - t_i)x_i + \lambda\omega \tag{6}$$

三、模型训练及结果

设置训练集、验证集和测试集，分别分配8000、1000、1000个数据，验证集用于选取参数λ。用上述推导公式进行前向传播和梯度反传训练，并进行了一系列测试。

最初选择每次传入batch进行训练，实验发现当batch取到2000以上时模型才会稳定，否则会出现如图所示的震荡现象

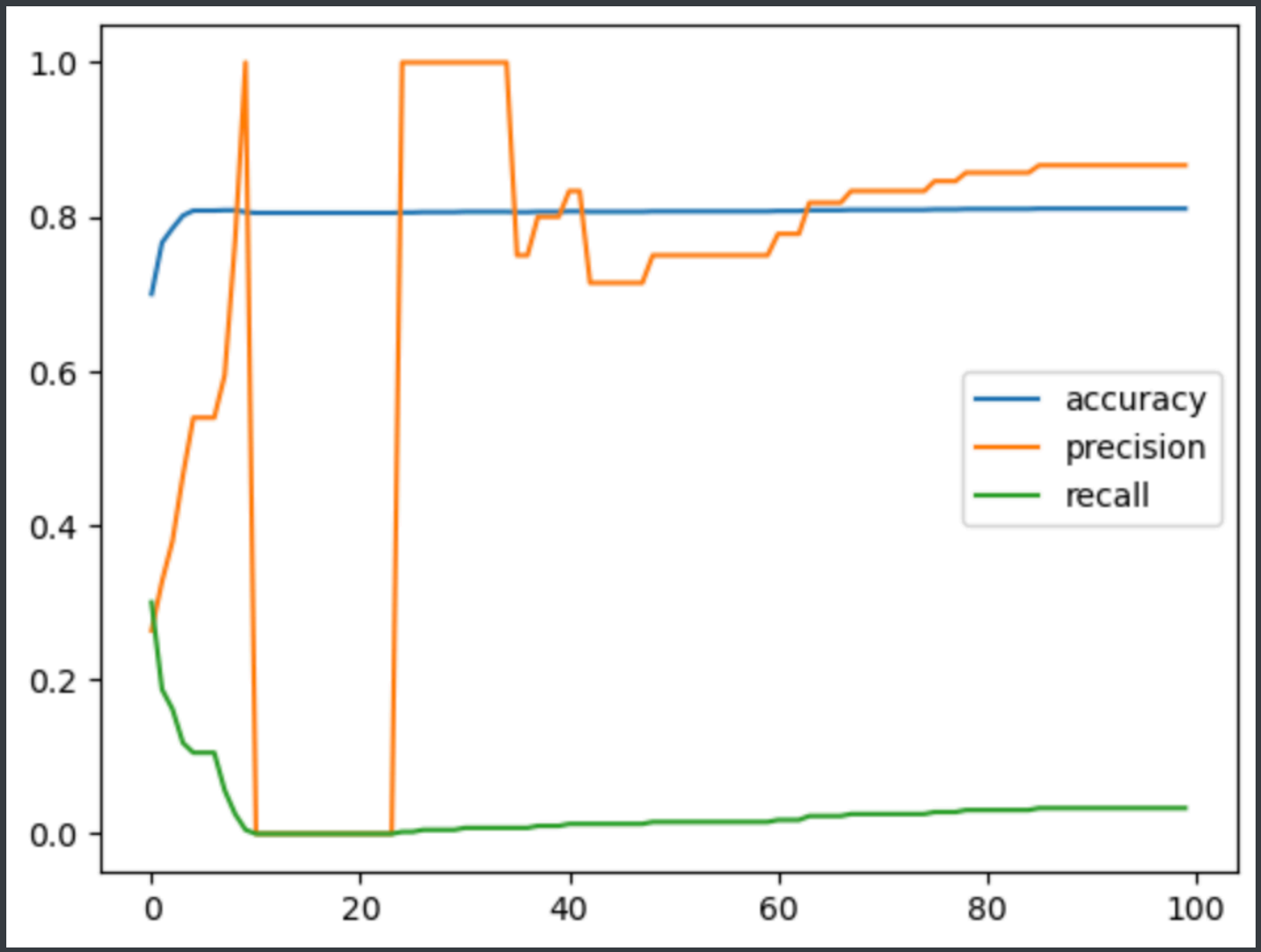


考虑到样本量并不大，故直接每次传入所有样本计算。为防止过拟合现象，另学习率随迭代次数的增加而减小，即

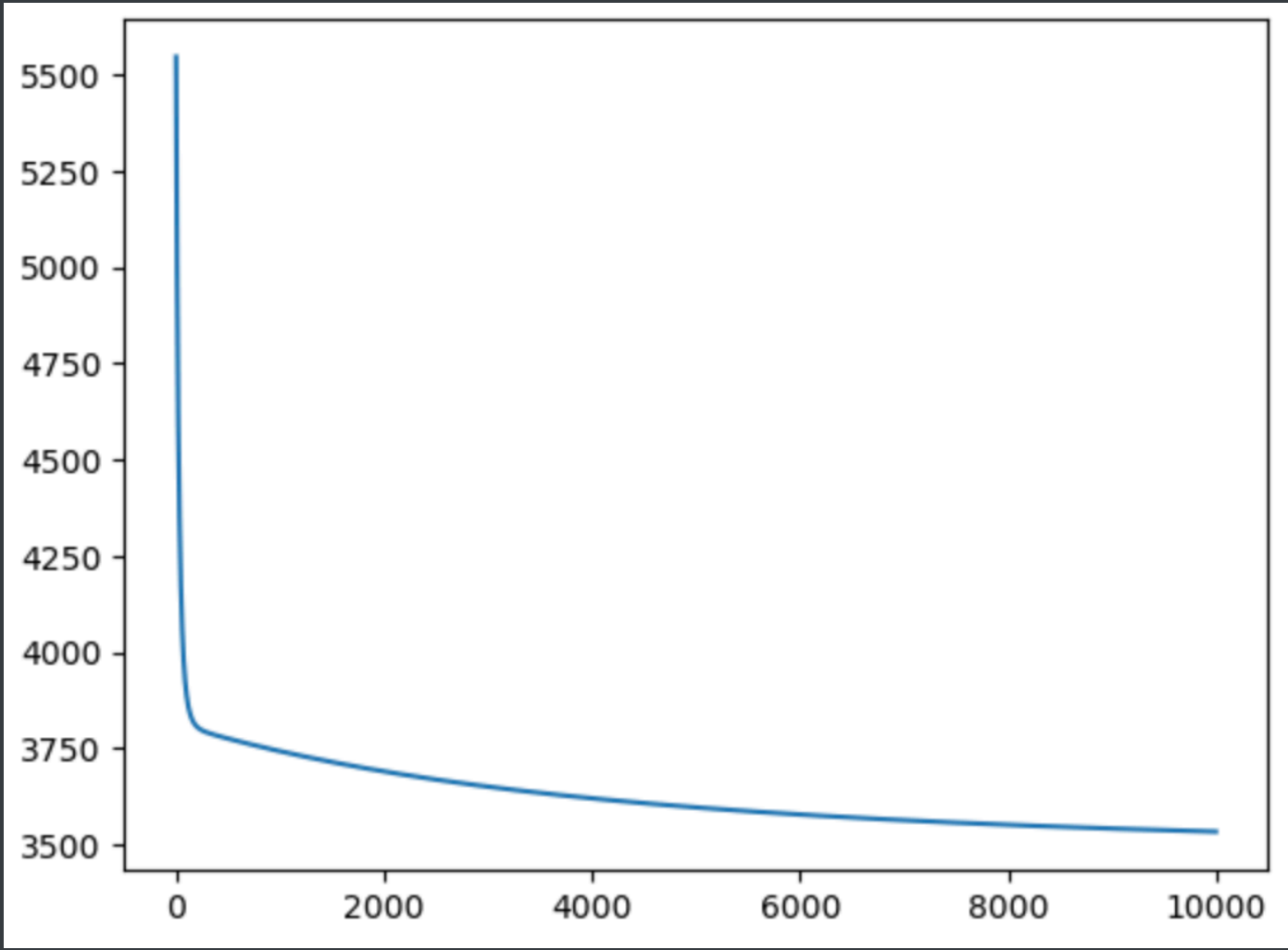
$$\alpha = \alpha_0 e^{-T \times epoch} \tag{7}$$

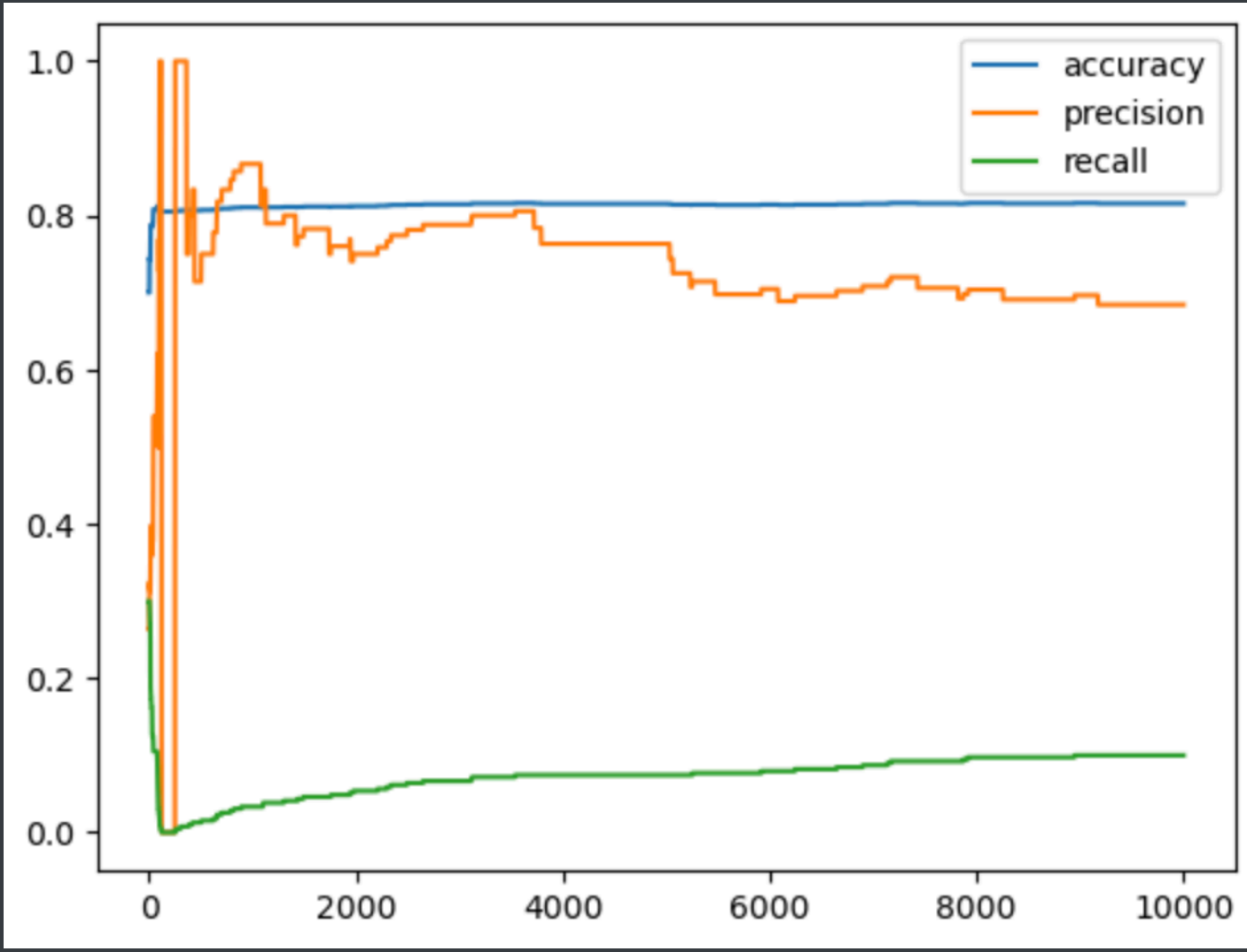
其中$\alpha = 0.0001$，T为下降速率，可选取0.0001，实验证明此方法可有效阻止过拟合现象。

对于逻辑回归模型，epoch为1000时，模型欠拟合



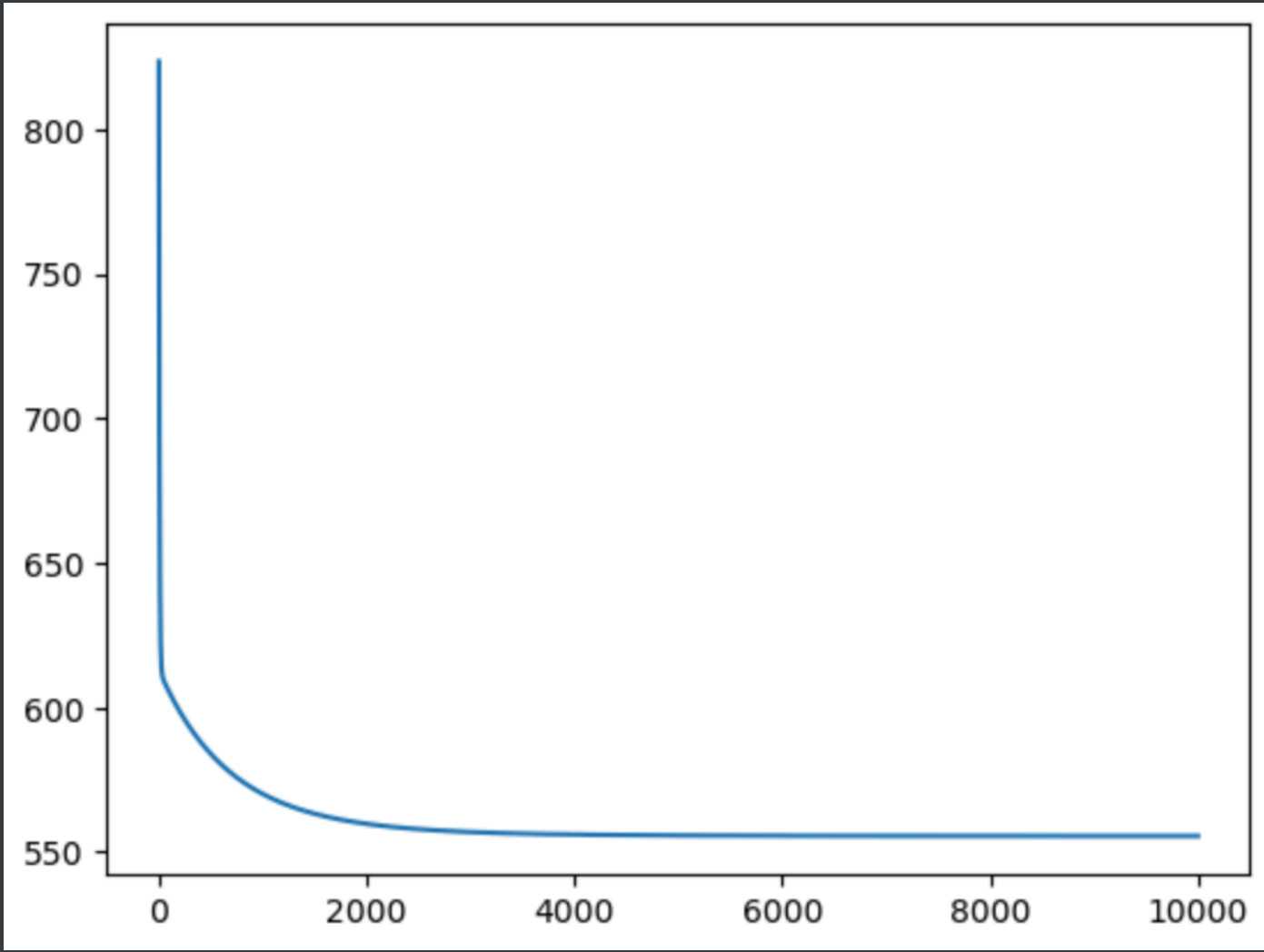
设置epoch为10000，模型可有效收敛，下图为loss和accuracy、recall、precision的变化：

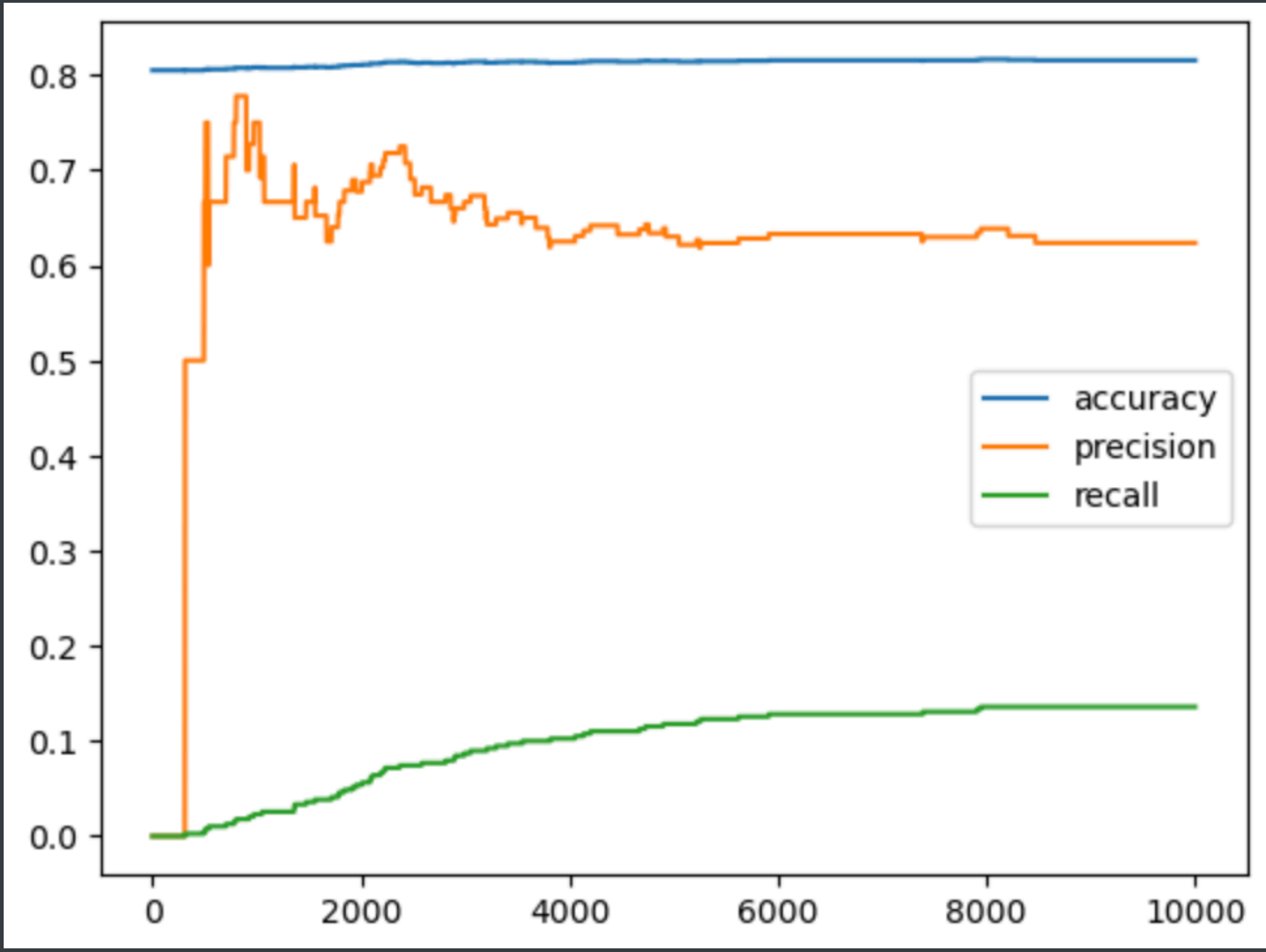

5500
5250
5000
4750
4500
4250
4000
3750
3500
0
2000
4000
6000
8000
10000



accuracy很快达到了0.8以上，recall和precision也可以逐渐收敛，未出现过拟合现象，这是因为我们在目标函数中设置了正则项，以及逐渐减小的学习率。

对MSE回归模型，超参数设置与上述相同，设置epoch为10000，模型可有效收敛，下图为loss和accuracy、recall、precision的变化：


800
750
700
650
600
550
0
2000
4000
6000
8000
10000



其整体的变化趋势与逻辑回归相似，但是曲线变化看起来更加稳定一些，可能是因为对于分类问题，回归中没有sigmoid激活函数，其相对而言更容易训练一些，当然泛化性自然会差一些，可通过增加二次项提高泛化能力。