1 实验要求

1. 在 wine.data 数据集中，第一列为 class label，第二列至第十四列为 feature。数据集含有三个类，本实验需要移除其中一个类，保留另两个类以生成新的数据集，并使用感知机进行二元分类。
2. 数据集需分为 70%的训练集和 30%的测试集，分别使用 BGD 和 SGD 对模型进行训练，并对测试集的数据进行预测，评估训练效果。

2 实验过程与现象

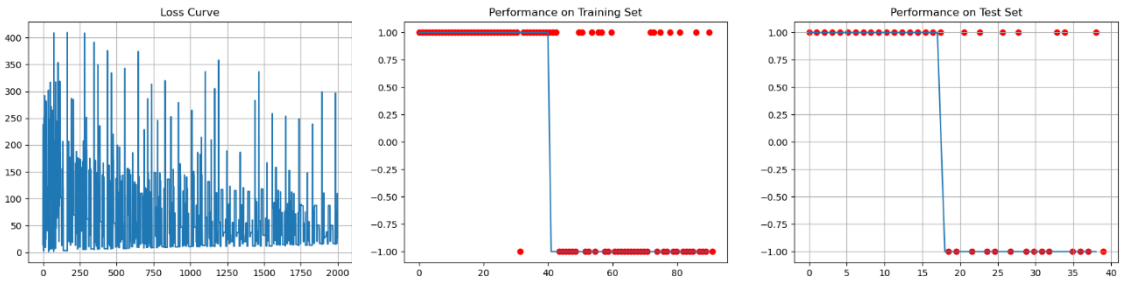
移除数据集中 class label=3 的数据，并将 class label=2 的数据重新标记为-1，按比例对 class1 和 class2 进行训练集和测试集的划分，即：

	Label	训练集元素数	测试集元素数	总数
Class 1	1	41	18	59
Class 2	-1	50	21	71
总数	\	91	39	130

2.1 第一次尝试

首先，在不使用归一化和标准化的情形下，BGD 和 SGD 的训练效果均较差，具体表现为：

1. Loss 曲线持续震荡（在 SGD 中表现非常明显）；
2. 模型在训练集和测试集上均有较多数据点不能拟合。



SGD (normalization not applied)

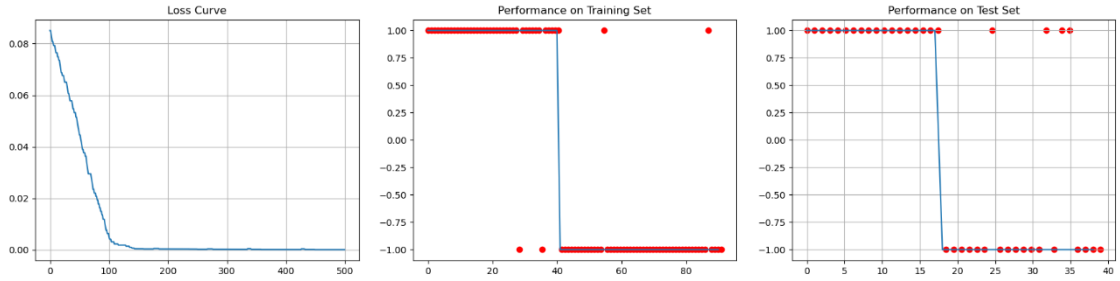
在有限的实验时间内，发现仅调节学习率和训练轮数无法解决以上问题，这说明模型存在缺陷，参数难以收敛。

2.2 第二次尝试

使用归一化或者标准化，两种梯度下降的训练效果均大幅提升：

2.2.1 SGD + Normalization

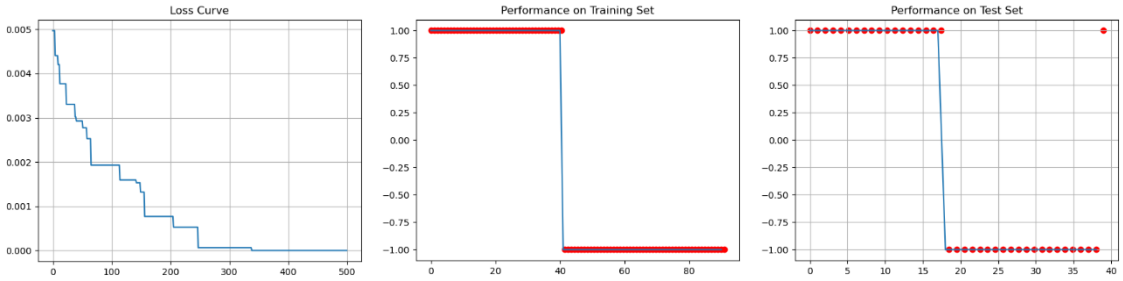
epochs	lr
500	1e-3



Accuracy	Recall	Precision	F1
0.90	1.00	0.82	0.90

2.2.2 SGD + Standardization

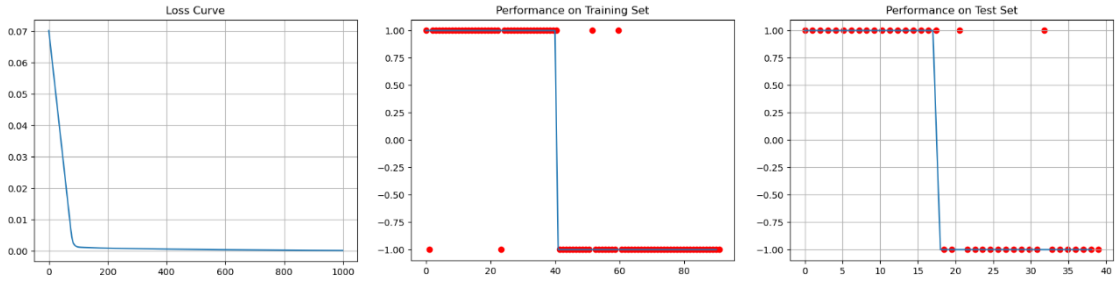
epochs	lr
500	1e-3



Accuracy	Recall	Precision	F1
0.97	1.00	0.95	0.97

2.2.3 BGD + Normolization

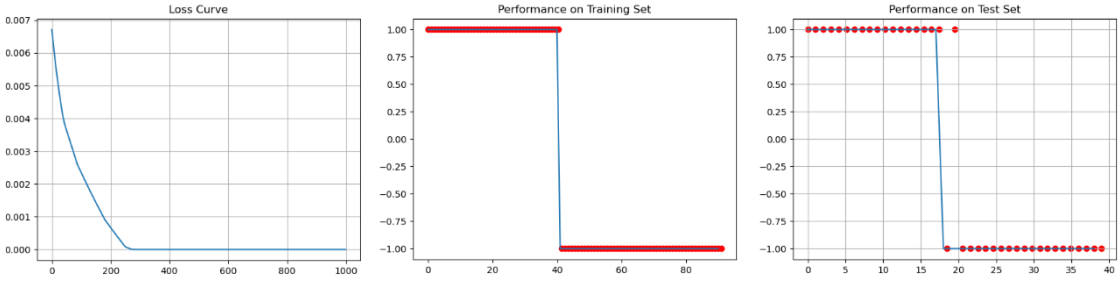
epochs	lr
1000	1e-3



Accuracy	Recall	Precision	F1
0.95	1.00	0.90	0.95

2.2.4 BGD + Standardization

epochs	lr
1000	1e-3



Accuracy	Recall	Precision	F1
0.97	1.00	0.95	0.97

3 实验结论

wine.data 数据集中，各 feature 取值相差较大，归一化和标准化使各 feature 处于同一数量级别，加快了梯度下降求最优解的速度。在该实验条件下，标准化的综合表现优于归一化。总而言之，在处理该类数据集时，归一化和标准化是非常有效的手段。

用 Perceptron 进行二元分类时，SGD 和 BGD 仍然具有曾经在线性回归中表现的特点，即：SGD 更易受到噪声的干扰；BGD 每一步更新更趋近于最优解的方向，且 BGD 的速度会慢于 SGD。

4 附：代码及作图说明：

wine.data 和 python 文件放在同一目录下。

在“performance on xxx”图中，蓝色曲线代表真实值，红色圆点代表模型预测值。红点在蓝线上表示预测正确，否则预测失败。

运行代码后会弹出对话，输入“sgd”会进行随机梯度下降，输入“bgd”会进行批量梯度下降。