experiment record and report: ex1

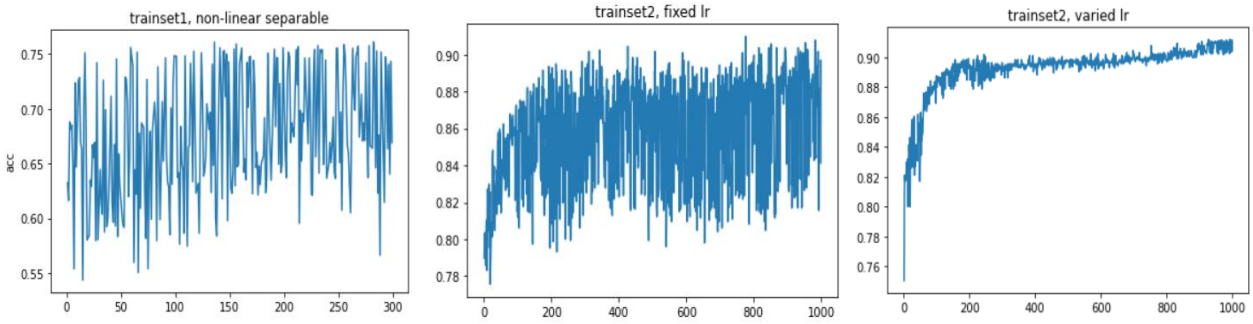
王敏行 id: 2018012386 wangmx18@mails.tsinghua.edu.cn

本实验中，我构建了 FLD、感知机和逻辑斯蒂回归的训练和预测模型，代码均由本人完成，第三方包仅用于计算、评估等目的。

分别用 FLD 方法对 trainset1 和 trianset2 进行分类，并判断准确率。发现，前者的分类准确率为 0.7944，后者准确率为 1.0。据 FLD 的结果，trainset2 是线性可分的，trainset1 则不是。

将在 trainset1 上训练的超平面用于 testset1 的数据分类，准确率为`0.7757520510483136`，即错误率为 `0.22424794895168643`。其准确率相较于 trainset1 较低，也许是两个数据集本身的波动导致的。

自行构建的感知机可以设定学习效率、迭代次数，也可以选择随着训练进程降低学习效率。首先，通过绘制不同训练进度下的分类准确率，可以看出感知机模型在 trainset1 上是不收敛的（最终 acc 约 0.67），在 trainset2 上是收敛的（acc 可优化到 0.9，迭代次数增加肯定可以更高）。这与两个模型的线性可分性一致。训练过程中的准确率变化如下图所示。



从 trainset2 中的趋势可以看出振荡，说明 learning rate 过大，一般可以采用 momentum 或 learning rate decay 的方法避免振荡。这里采用后者，结果如右图所示。可见通过采取变化学习率的策略，可以加快模型的收敛速度，并提到模型最终的准确率。

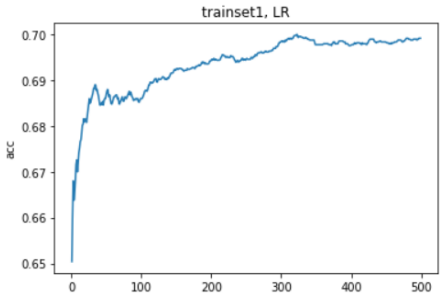
分别在 trainset1、2 上测试在 trainset1、2 上训练的感知机，错误率如下表。

err rate	trainset1	trainset2
trainset itself	`0.259`	`0.100`
testset1	`0.262`	`0.235`
testset2	`0.251`	`0.193`

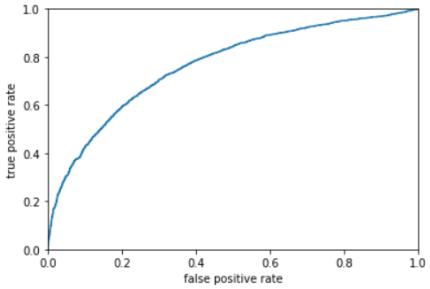
显然，trainset2 上训练的模型过拟合情况较为明显，这是由于 trainset2 本身是线性可分的，且感知机本身没有 margin 的概念，因此模型的可推广性较差，即体现为过拟合情况严重。

逻辑斯蒂回归的部分分为模型训练、数据打分和类别数据类别预测三个部分。

在 trainset1 上采用 learning rate decay 的策略训练分类器，发现收敛迅速，最终训练集准确率在 0.7 左右。测试集准确率约为 0.719(test err=0.18)，没有迹象指向过拟合。



利用 matplotlib 包绘制 ROC，并计算 AUC。AUC=0.766。



采取删除单一维度的特征，对剩下的特征进行分类并考察模型的准确率。发现即使删除单一维度，预测效果仍然不错。也许接下来需要尝试两个维度一起删除。