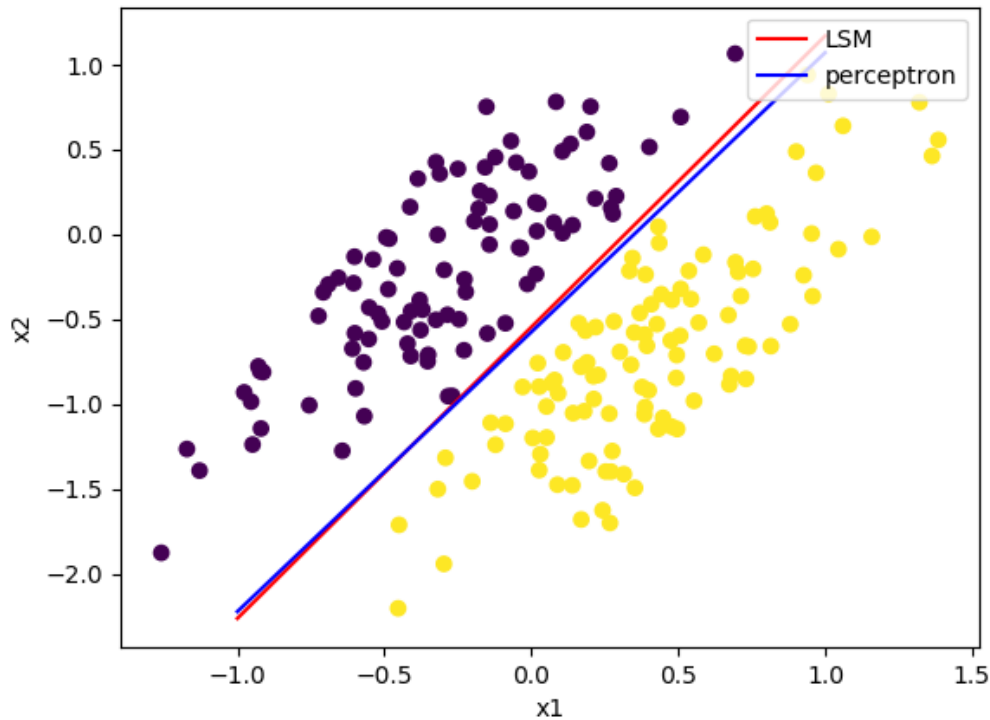


Assignment 2 Brief Report

1. 最小平方误差模型和感知器模型

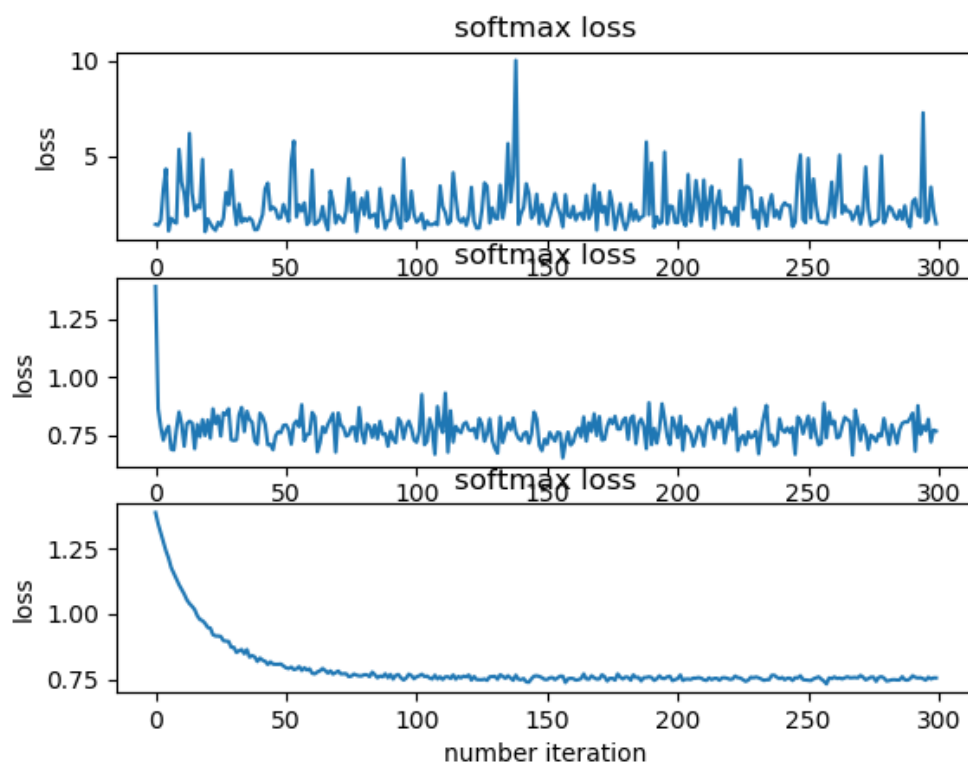
对于线性可分的模型，最小平方误差模型和感知器模型的表现都很好，在测试集上的表现都达到了 100%，下图为两个模型决策边界的可视化。



2. Text classification

- ① 文本分类过程较为简单，先遍历数据集建立词典，将每个文档转变为 multi-hot 的稀疏向量表示并以此作为特征使用一个线性模型后接上 SVM 分类器即可。训练端到端，同时检验不同 batch_size 带来的影响
- ② 梯度用链式法则即可求得
 - a. Bias 项不用正则化惩罚，因为对 bias 项惩罚，说明希望 bias 项尽可能接近 0，而 bias 项只控制模型的位置而不控制模型的形状，因此不存在说过拟合的情况，对 bias 项加正则化容易导致欠拟合。
 - b. 用数值梯度计算方法，在某个点左右加减小一个范围 ϵ (取 10^{-6})，计算 $F(x + \epsilon) - F(x - \epsilon) / 2\epsilon$ ，检查计算结果和解析法梯度算出来的结果是否相同
- ③ 在下图中我们放上三种训练方式的 loss 曲线。
 - a. Learning_rate 超参数，说实话一般根据经验来判断，适合的 learning_rate 需要在训练的开始就可以迅速的得到 loss 值的下降，并且在训练后期不会因为过大而是 loss 值上升。当然可以使用 adam 等方式自动调节 learning rate。
 - b. 一般是经过一定数量 (e.g. 100) 的梯度下降后，如果 loss 值趋于稳定，这时候就可以停止训练过程，通过在测试机上评测模型，判断模型的效果。
- ④ 对于三种不同的训练方式：SGD、Mini-batch GD 和 Full-batch GD 的 loss 曲线，可以发现 Full batch 最稳定，但是花费时间较长；SGD 训练速度很快，但是起伏很大，最终结果也不是很好。相对来说，Mini-batch GD 是两者的一个权衡，这里我们选

择的 `batch_size = 32`。



- ⑤ Full batch 的准确率最高，达到了 92.18%；Mini-batch 差一点，达到了 84.04%，SGD 的表现最差，只有 34.39%。这里使用的 epoch 数量为 300，可以明显的看到 mini-batch 的情况下 loss 函数下降的最快，到达极值点的速度也是最快的，但是后期波动较大可能是因为 `learning_rate` 太大，需要使用 RMSprop 或者 adam 等高级优化器自适应的调节学习率。