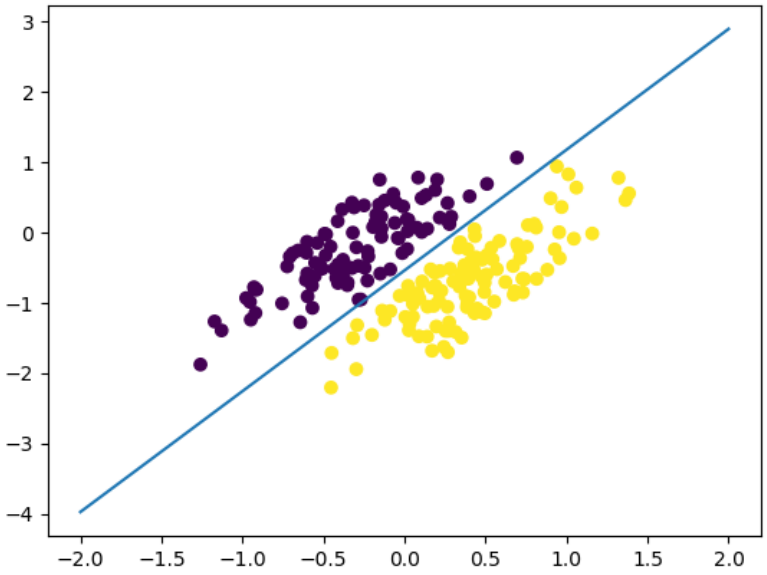
PRML Assignment 2 报告

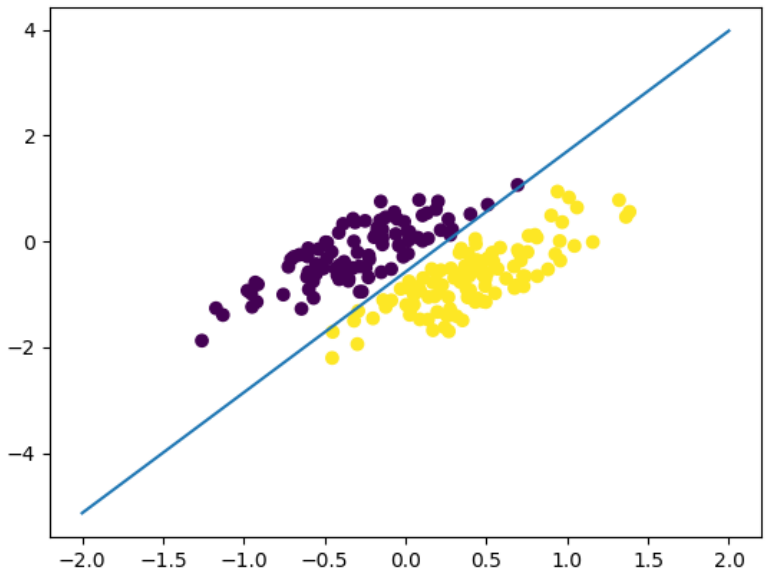
16307130076 赵伟丞

Part I

对于最小二乘法，直接使用解析解$\boldsymbol{W}=(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T\boldsymbol{y}$计算，其中令$x_0=1$使得$\boldsymbol{b}$直接在$\boldsymbol{W}$中完成计算。报告的准确率为 100%，划分线如图所示。



对于感知机模型，采用$\boldsymbol{w}=\boldsymbol{w}+\boldsymbol{\eta}\cdot\boldsymbol{y}\cdot\boldsymbol{x},\boldsymbol{b}=\boldsymbol{b}+\boldsymbol{\eta}\cdot\boldsymbol{y}$($\boldsymbol{\eta}$为学习率)进行梯度下降，当采用 0.01 的学习率时，在第 7 步收敛，报告准确率 100%，划分线如下图所示。



Part II

1、 数据的预处理。

将所有的文字读入后，将所有的 `string.punctuation` 替换为''，同时利用 `split` 函数以所有空白字符为标记进行分割，然后将结果使用 `Counter` 数据结构进行计数，计数完成后剔除所有计数值小于 `min_count` 的记录，然后提取所有键值并按字母表排序，获得字典。对于每个经处理的文本，如果其中单词出现在字典中就置该单词在字典中的序号位为 `1`，最终得到最后的 multi-hot 向量。

对于类型数据，直接读取其中的数并将对应为置为 1 即可得到所需 one-hot 向量。

2、 梯度下降的计算。

（1）

$$\begin{aligned}\boldsymbol{L} &= -\boldsymbol{y}^{\mathrm{T}}\ln\frac{e^{\boldsymbol{z}}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{e}^{\boldsymbol{z}}}\\ &= -\boldsymbol{y}^{\mathrm{T}}(\boldsymbol{z}-\ln(\mathbf{1}^{\mathrm{T}}\boldsymbol{e}^{\boldsymbol{z}}))\\ &= \ln(\mathbf{1}^{\mathrm{T}}\boldsymbol{e}^{\boldsymbol{z}})-\boldsymbol{y}^{\mathrm{T}}\boldsymbol{z}\end{aligned}$$

又

$$\mathrm{d}(\ln(\mathbf{1}^{\mathrm{T}}\boldsymbol{e}^{\boldsymbol{z}})) = \frac{1}{\mathbf{1}^{\mathrm{T}}\boldsymbol{e}^{\boldsymbol{z}}}\odot d(\mathbf{1}^{\mathrm{T}}\boldsymbol{e}^{\boldsymbol{z}})$$

$$d(\mathbf{1}^{\mathrm{T}}\boldsymbol{e}^{\boldsymbol{z}}) = \mathbf{1}^{\mathrm{T}}\mathrm{d}(\boldsymbol{e}^{\boldsymbol{z}}) = \mathbf{1}^{\mathrm{T}}(\boldsymbol{e}^{\boldsymbol{z}}\odot\mathrm{d}\boldsymbol{z})$$

则

$$\mathrm{d}L = \frac{\mathbf{1}^{\mathrm{T}}(\boldsymbol{e}^{\boldsymbol{z}}\odot\mathrm{d}\boldsymbol{z})}{\mathbf{1}^{\mathrm{T}}\boldsymbol{e}^{\boldsymbol{z}}}-\boldsymbol{y}^{\mathrm{T}}\mathrm{d}\boldsymbol{z}$$

两边同时计算迹

$$\begin{aligned}\mathrm{d}L &= \mathrm{tr}\left(\frac{(\mathbf{1}\odot\boldsymbol{e}^{\boldsymbol{z}})^{\mathrm{T}}\mathrm{d}\boldsymbol{z}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{e}^{\boldsymbol{z}}}\right)-\mathrm{tr}(\boldsymbol{y}^{\mathrm{T}}\mathrm{d}\boldsymbol{z})\\ &= \mathrm{tr}\left(\left(\frac{(\boldsymbol{e}^{\boldsymbol{z}})^{\mathrm{T}}}{\mathbf{1}^{\mathrm{T}}\boldsymbol{e}^{\boldsymbol{z}}}-\boldsymbol{y}^{\mathrm{T}}\right)\mathrm{d}\boldsymbol{z}\right)\\ &= \mathrm{tr}((\hat{\boldsymbol{y}}-\boldsymbol{y})^{\mathrm{T}}\mathrm{d}\boldsymbol{z})\\ &= \mathrm{tr}\left((\frac{\partial L}{\partial\boldsymbol{z}})^{\mathrm{T}}\mathrm{d}\boldsymbol{z}\right)\end{aligned}$$

于是有

$$\frac{\partial L}{\partial\boldsymbol{z}} = \hat{\boldsymbol{y}}-\boldsymbol{y}$$

又

$$\frac{\partial\boldsymbol{z}}{\partial\boldsymbol{w}^{\mathrm{T}}} = \boldsymbol{x}, \frac{\partial\boldsymbol{z}}{\partial\boldsymbol{b}} = \mathbf{1}$$

因此

$$\frac{\partial L}{\partial\boldsymbol{w}^{\mathrm{T}}} = -\boldsymbol{x}(\hat{\boldsymbol{y}}-\boldsymbol{y})^{\mathrm{T}}, \frac{\partial L}{\partial\boldsymbol{b}} = -\mathbf{1}(\hat{\boldsymbol{y}}-\boldsymbol{y})^{\mathrm{T}}$$

（2）在 $\boldsymbol{W}$ 上加入了 L2 正则化后，偏置上无需加上正则化，因为在 $\boldsymbol{W}$ 上加上 L2z 正则化惩罚大参数后，偏置项不会过大，因此并无添加正则化的必要。

（3）使用梯度检验，即使用下面的公式:

$$g(\boldsymbol{w}) \approx \frac{L(\boldsymbol{w}+\varepsilon)-L(\boldsymbol{w}-\varepsilon)}{2\varepsilon}$$

采用 1e-4 的ε，对于若干点进行测试，观察结果是否一致。

3、 模型的训练

（1） 学习率的确定，采用从一个较大的学习率（0.01）开始，逐步减小支准确率达到一个相当的水平。

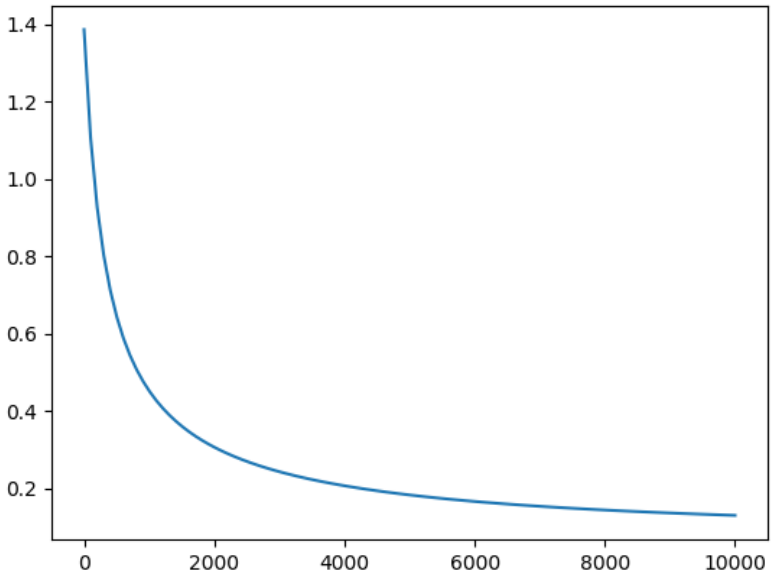
（2） 终止条件：Loss 低于阈值或迭代次数达到给定值（主要采用后者）。

4、 随机梯度下降与批量梯度下降

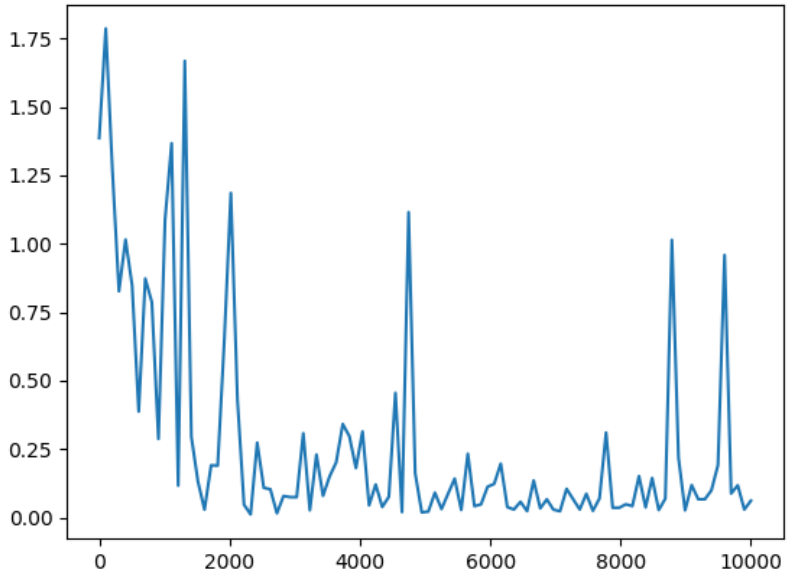
（1） 随机梯度下降的损失函数下降不稳定有波动，收敛速度也不稳定。批量梯度下降时，随着每批数据点的增多，不稳定性降低。

（2） 全梯度下降：优势：所有数据都参与，不宜陷入局部最优，收敛速度稳定；劣势，计算开销大（时间，空间）。随机梯度下降：优势：计算速度快，空间占用小；劣势：容易陷入局部最优，收敛速度不稳定。批量梯度下降：优势：介于以上两者之间；劣势：增加了一个超参数（每批数据量）。

5、 全梯度下降：使用参数 c=0.001, learning_rate=0.01, max_iteration=10000，在 10000 次迭代时终止。在训练集上报告准确率 99.47%，测试集上报告准确率 91.11%。损失曲线如下图。



随机梯度下降：使用参数 c=0.001, learning_rate=0.01, max_iteration=10000，在 10000 次迭代时终止。在训练集上报告准确率 99.19%，测试集上报告准确率 91.11%。损失曲线如下图。



随机梯度下降：使用参数 c=0.001，learning_rate=0.01，max_iteration=10000，batch_size=4，在 10000 次迭代时终止。在训练集上报告准确率 99.33%，测试集上报告准确率 91.44%。损失曲线如下图。