Lab1报告存在的问题

- 尽量不要准点提交实验报告，网络可能会很卡，导致提交失败
- TA会准时拉取实验报告，但发现拉取后还有同学在提交，这部分的实验报告我们是不会再拉取的
- 压缩包里只需要包含实验报告和代码，不需要上传整个工程文件
- 可以稍微注意一下实验报告的排版，不要求做的很精致，但至少看起来比较舒服
- 注意流程图的规范，不要画成流水线的形式
- 核心代码部分，不需要把所有的代码都贴上去，只需要贴关键部分，并适当写一下注释
- 实验结果分析部分，着重对比分析不同的参数设置下的不同结果，尽量通过图表形式进行展示

感知机 & 逻辑回归

主讲TA：张祖胜

2020/9/25

感知机（PLA）

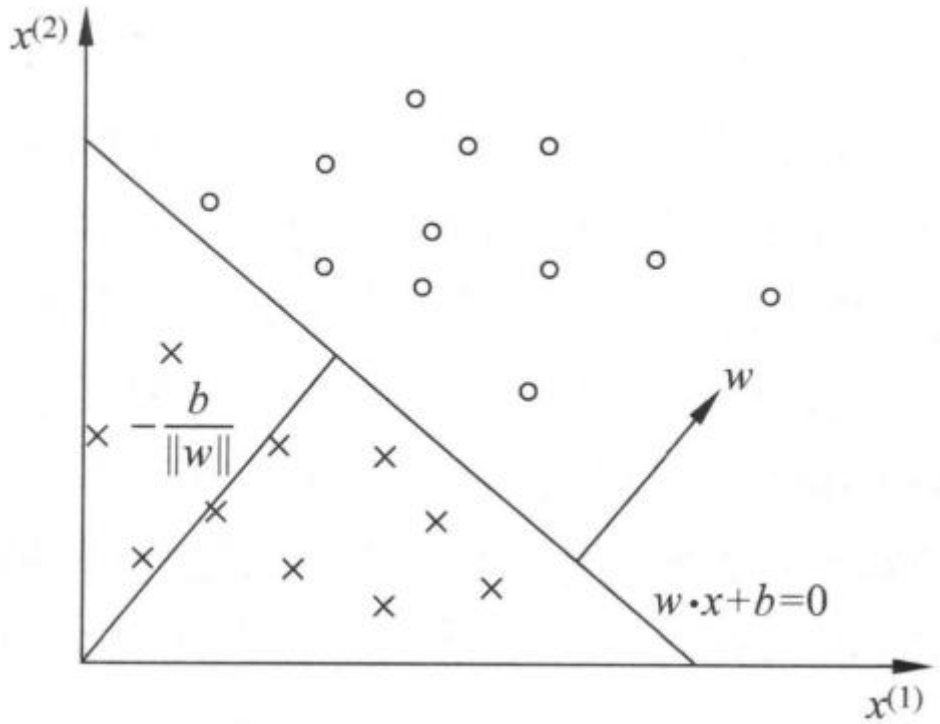
- 感知机针对二分类问题，输入是样本的特征向量$\mathbf{x} \in \mathbf{R}^n$，输出是样本的类别$\mathbf{y} \in \{+1, -1\}$

- 感知机可以表示为：$f(x) = \mathbf{sign}(w \cdot x + b)$
  - $\mathbf{sign}(x)$ 表示符号函数
  - w和 b 表示模型参数，$w \in \mathbf{R}^n$ 称为权值向量，$b \in \mathbf{R}$ 称为偏置

- 感知机的几何解释：线性方程 $w \cdot x + b = 0$ 对应于特征空间$\mathbf{R}^n$的一个分离超平面；平面将特征空间分为两部分，位于两部分的点分别对应于正、负样本

感知机（PLA）

- 感知机的几何解释



感知机（PLA）

- 损失函数：所有误分类点到分离超平面的距离之和，感知机的学习目标是令该和尽可能小，即令误分类点的数量尽可能少

- 误分类点 (x_i, y_i) 到分离超平面的距离：$\frac{1}{\|w\|}\|w \cdot x_i + b\|$

  - 由于 $-y_i(w \cdot x_i + b) > 0$，该距离可写成：$-\frac{1}{\|w\|} y_i (w \cdot x_i + b)$

  - 假设误分类点集合为 M，并且不考虑 $-\frac{1}{\|w\|}$，损失函数可写成：

$$L(w,b) = -\sum_{x_i \in M} y_i (w \cdot x_i + b)$$

感知机（PLA）

- 使用随机梯度下降对损失函数进行优化

- 损失函数的梯度为：

$$\nabla_w L(w,b) = -\sum_{x_i \in M} y_i x_i$$

$$\nabla_b L(w,b) = -\sum_{x_i \in M} y_i$$

- 选择一个误分类点 (x_i, y_i) 对参数进行更新，其中 η 表示学习率：

$$w = w + \eta y_i x_i$$

$$b = b + \eta y_i$$

感知机（PLA）

- 感知机算法步骤：
  1. 设置学习率η；随机初始化w和b，一般可初始化为0
  2. 选取一个误分类点(x_i, y_i)，即如果$y_i(w \cdot x_i + b) \leq 0$，则对参数进行更新

$$w = w + \eta y_i x_i$$

$$b = b + \eta y_i$$

  3. 重复步骤2，直到训练集中没有误分类点

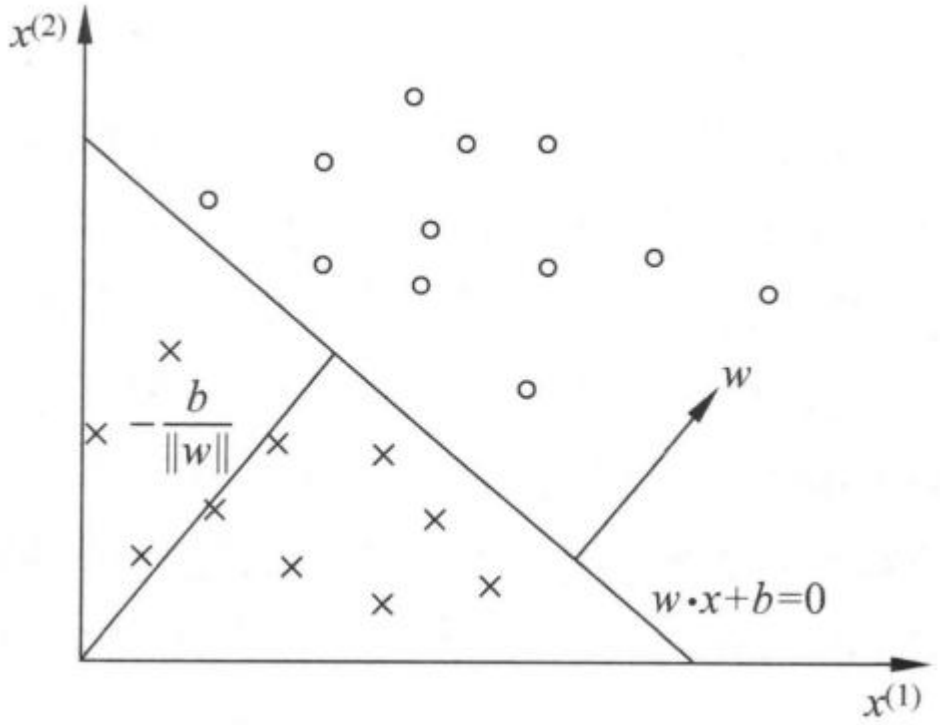
感知机（PLA）

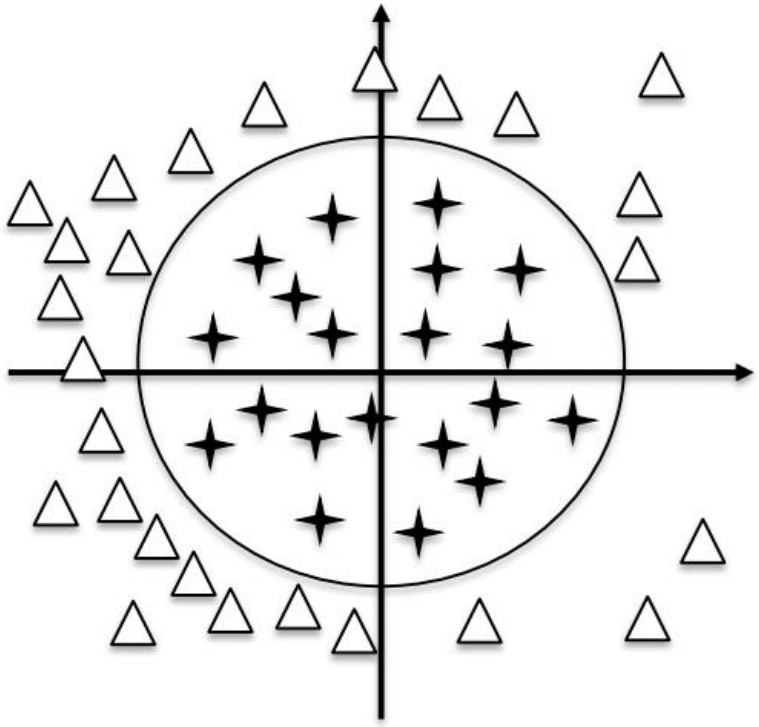
- 对于线性可分的数据集，总是可以找到一个模型将其正确划分
  - 线性可分：存在一个分离超平面，该平面能够将正、负样本完全正确地划分到平面的两侧

- 对于线性不可分的数据集，有两种解决方案：
  - 设置最大迭代次数；当迭代次数到达预设值时，停止训练
  - 找到一组参数(w,b)，使得训练集使用该组参数进行划分后，分类错误的样本最少

感知机（PLA）

- 线性可分与线性不可分





逻辑回归（LR）

- 硬分类模型：非概率模型，通常表示为函数形式，即使用一个决策函数来直接判断样本的类别，如感知机、决策树等

- 软分类模型：概率模型，通常表示为概率分布形式，即先算出每个类别的概率，然后根据概率的大小来判断样本的类别，如逻辑回归

逻辑回归（LR）

- 逻辑回归通常针对二分类问题，输入是样本的特征向量 $\mathbf{x} \in \mathbf{R}^n$，输出是样本属于某个类别 $y \in \{0,1\}$ 的概率

- 逻辑回归可以表示为：

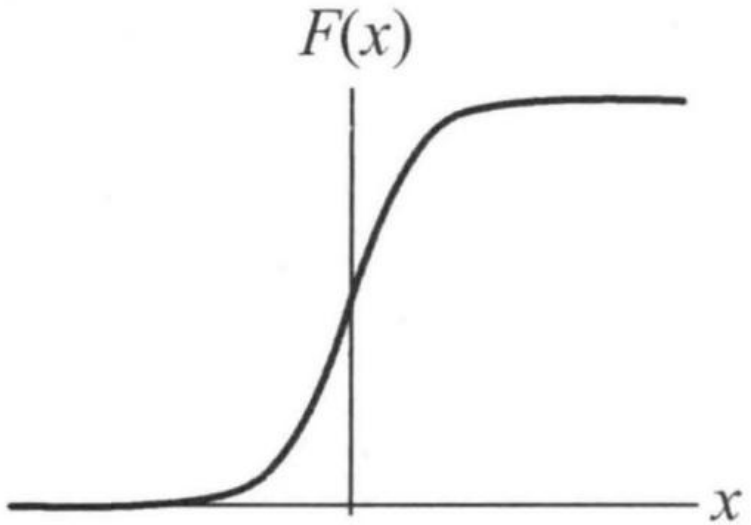
$$P(y=1|x)=\frac{\exp(w\cdot x+b)}{1+\exp(w\cdot x+b)},\quad P(y=0|x)=1-P(y=1|x)$$

- 为方便表示，将 w 表示为 $w=\left(w^{\mathrm{T}},b\right)^{\mathrm{T}}$，将 x 表示为 $x=\left(x^{\mathrm{T}},1\right)^{\mathrm{T}}$：

$$P(y=1|x)=\frac{\exp(w\cdot x)}{1+\exp(w\cdot x)}=\frac{1}{1+\exp(-w\cdot x)},\quad P(y=0|x)=1-P(y=1|x)$$

逻辑回归（LR）

- Logistic函数：$F(x)=\frac{1}{1+\exp(-x)}$
  - $F(+\infty)=1$，当加权求和结果无穷大，样本属于正类别的概率为1
  - $F(-\infty)=0$，当加权求和结果无穷小，样本属于正类别的概率为0



逻辑回归（LR）

- 令 $\pi(x)=\frac{1}{1+\exp(-w\cdot x)}$，则某个样本 x 属于某个类别 y 的概率表示为

$$f(x)=P(y\mid x)=\pi(x)^{y}(1-\pi(x))^{1-y}$$

  - 当 $y=1$，$f(x)=P(y=1\mid x)=\pi(x)$
  - 当 $y=0$，$f(x)=P(y=0\mid x)=1-\pi(x)$
  - 在某种模型下，利用给定数据 x 得到给定标签 y 的概率，称为该问题的似然；$f(x)$ 称为似然函数

逻辑回归（LR）

- 在整个训练集上考虑似然函数：

$$\prod_{i=1}^{N}[\pi(x_i)]^{y_i}[1-\pi(x_i)]^{1-y_i}$$

- 对数似然函数为（似然函数取对数）：

$$\begin{aligned} L(w) &= \sum_{i=1}^{N}[y_i \log \pi(x_i) + (1-y_i)\log(1-\pi(x_i))] \\ &= \sum_{i=1}^{N}\left[y_i \log \frac{\pi(x_i)}{1-\pi(x_i)} + \log(1-\pi(x_i))\right] \\ &= \sum_{i=1}^{N}[y_i(w \cdot x_i) - \log(1+\exp(w \cdot x_i)] \end{aligned}$$

逻辑回归（LR）

- 根据极大似然估计，对 $L(w)$ 求极大值，可得到参数 w 的估计值

- 对 $L(w)$ 取负，将 $-L(w)$ 作为逻辑回归模型的损失函数，并使用梯度下降法对损失函数进行优化

- 损失函数的梯度为：

$$-\nabla_w L(w) = -\sum_{i=1}^{N}[y_i - \pi(x_i)]x_i$$

- 对参数进行更新：

$$w = w + \eta\sum_{i=1}^{N}[y_i - \pi(x_i)]x_i$$

逻辑回归（LR）

- 逻辑回归算法步骤：
  1. 给每个样本的特征向量添加一维常数项1，即 $x=\left(x^{\mathrm{T}},1\right)^{\mathrm{T}}$
  2. 设置学习率 η；对n+1维的权值向量 $w=\left(w^{\mathrm{T}},b\right)^{\mathrm{T}}$ 进行随机初始化
  3. 计算当前梯度，并对参数进行更新：$w=w+\eta\sum_{i=1}^{N}\left[y_i-\pi\left(x_i\right)\right]x_i$
  4. 重复步骤3，直至满足一定的收敛条件

思考题

- 不同的学习率η对模型收敛有何影响？从收敛速度和是否收敛两方面来回答。

- 使用梯度的模长是否为零作为梯度下降的收敛终止条件是否合适，为什么？一般如何判断模型收敛？

实验注意事项

- 实现感知机和基于批梯度下降的逻辑回归，分别提交一份代码
- 本次的实验数据是train.csv，前40列表示特征，最后一列表示标签（0或1）
- 请自行分好训练集、测试集（在报告里说明怎么分的），评测指标为测试集上的准确率
- **验收**时使用的基准模型如下，学习率均设为1：
  1. 感知机：固定迭代次数，权重初始化为零，每次迭代按顺序从第一个样例开始找下一个错误的样例
  2. 逻辑回归：固定迭代次数，权重初始化为零，使用批梯度下降优化
- DDL：
  1. 验收：第六周周五实验课（10.9）
  2. 实验报告：第六周周四晚12:00（10.8）