实验四： Perceptron Learning Algorithm ——感知机算法

PPT制作：李彦良，林东定
出题人　：詹雪莹，王耀威

PLA分类

- 用来解决二元分类问题（+1和-1两类）
- 样本x={$x_1, x_2, x_3 \ldots\ldots x_d$}
- 权重向量w={$w_1, w_2, w_3 \ldots\ldots w_d$}
- 阈值threshold
- if $\sum_{i=1}^{d} w_i x_i$ > threshold, predict +1
- if $\sum_{i=1}^{d} w_i x_i$ < threshold, predict -1
- if $\sum_{i=1}^{d} w_i x_i$ == threshold, predict 0(both ok)
- 用一个符号函数表示y=sign($\sum_{i=1}^{d} w_i x_i$-threshold)
- int sign(int x) {return x > 0 ? +1 : -1; }

PLA分类

- 为简便计算
- y = sign($\sum_{i=1}^{d} w_i x_i$-threshold)
- = sign($\sum_{i=1}^{d} w_i x_i$+(-threshold)*(+1))
- = sign($\sum_{i=0}^{d} w_i x_i$)
- = sign($w^T x$)
- 样本x={+1, $x_1, x_2, x_3 \ldots\ldots x_d$}
- 权重向量w={$w_0, w_1, w_2, w_3 \ldots\ldots w_d$}

PLA分类

训练1：$x_1 = \{x_{11}, x_{12}, x_{13} \ldots\ldots x_{1d}\}$ label = y_1

训练2：$x_2 = \{x_{21}, x_{22}, x_{23} \ldots\ldots x_{2d}\}$ label = y_2

训练3：$x_3 = \{x_{31}, x_{32}, x_{33} \ldots\ldots x_{3d}\}$ label = y_3

步骤1：给每一个样本前加常数项1

训练1：$x_1 = \{1, x_{11}, x_{12}, x_{13} \ldots\ldots x_{1d}\}$ label = y_1

训练2：$x_2 = \{1, x_{21}, x_{22}, x_{23} \ldots\ldots x_{2d}\}$ label = y_2

训练3：$x_3 = \{1, x_{31}, x_{32}, x_{33} \ldots\ldots x_{3d}\}$ label = y_3

PLA分类

训练1：$x_1 = \{1, x_{11}, x_{12}, x_{13} \ldots\ldots x_{1d}\}$ label = y_1

训练2：$x_2 = \{1, x_{21}, x_{22}, x_{23} \ldots\ldots x_{2d}\}$ label = y_2

训练3：$x_3 = \{1, x_{31}, x_{32}, x_{33} \ldots\ldots x_{3d}\}$ label = y_3

步骤2：初始化权重向量w_0=0或者其他值

步骤3：找到一个预测错误的样本（点），

即sign($w_t^T x_{n(t)}$)≠$y_{n(t)}$，

更新$w_{t+1} \leftarrow w_t + y_{n(t)} x_{n(t)}$，

重复步骤3直至全部预测正确

PLA分类

训练1：$x_1 = \{1, x_{11}, x_{12}, x_{13} \ldots\ldots x_{1d}\}$ label = y_1

训练2：$x_2 = \{1, x_{21}, x_{22}, x_{23} \ldots\ldots x_{2d}\}$ label = y_2

训练3：$x_3 = \{1, x_{31}, x_{32}, x_{33} \ldots\ldots x_{3d}\}$ label = y_3

步骤5：此时得到的w就是我们要求的值

步骤6：用此w来预测测试集的label

以上步骤全部基于数据集线性可分（即能够有一条线将它们完全区分）的假设，如果数据集并非线性可分，则需要将结束条件“全部预测正确”改为“超过设定的迭代次数”，下一张PPT会介绍，请大家注意。

存在的问题？

感知器不适用非线性的问题，很多时候w无法满足全部点，这时候有两种方法：

1. 设置迭代次数，迭代到一定程度就返回此时的w而不管它到底满不满足所有训练集。
2. 找一个w使得在训练集里以此w来分割错误的样本最少，即相当于有一个口袋，把算到的w跟口袋里的w比对，优胜劣汰，放入比较好的一个w，这种算法又被称为口袋（pocket）算法。

口袋算法

步骤1：给每一个样本前加常数项1

步骤2：初始化权重向量w_0=0或者其他值，

初始化全局权重向量w

步骤3：找到一个预测错误的样本（点），

即sign($w_t^T x_{n(t)}$)≠$y_{n(t)}$，

更新w_{t+1}←w_t+$y_{n(t)}x_{n(t)}$，

若w_{t+1}错误率小于w，w←w_{t+1}，

重复步骤3直至达到指定迭代次数

步骤5：此时得到的w就是我们要求的值

步骤6：用此w来预测测试集的label

简单的栗子

编号	特征1	特征2	标签
train1	1	-1	+1
train2	3	3	-1
test1	-2	3	?

步骤1：样本数据加常数项

train1: x_1={1, 1, -1} y_1 = +1

train2: x_2={1, 3, 3} y_2 = -1

Test1 : x_3={1, -2, 3} y_3 = ?

简单的栗子

train1: x_1={1, 1, -1} y_1=+1

train2: x_2={1, 3, 3} y_2 = -1

test1: x_3={1,- 2, 3} y_3 = ?

步骤2：初始化向量w={0, 0, 0}

步骤3：计算sign($w^T x_1$)=0≠y_1，

更新w得w=w+$y_1 x_1$={1, 1, -1}

步骤4：计算sign($w^T x_2$)=+1≠y_2，

更新w=w+$y_2 x_2$={0, -2, -4}

简单的栗子

train1: x_1={1, 1, -1} y_1=+1

train2: x_2={1, 3, 3} y_2 = -1

test1: x_3={1,- 2, 3} y_3 = ?

步骤5：计算sign($w^T x_1$)=+1≠y_1，

计算sign($w^T x_2$)=-1≠y_2，

预测全对，停止学习

步骤6：计算sign($w^T x_3$)=-1，

所以test1的预测标签为-1

数据集与评测指标

本次数据集为癌症诊断，根据一些因素来判断是否有癌症，每一维具体意义不给出，患病为1，不患病为-1

train_data.txt:	100个病人样例，每一个样例为10000维的特征向量
test_data.txt:	100个病人样例，每一个样例为10000维的特征向量
train_labels.txt:	100个训练集的标签
test_labels.txt:	100个测试集的标签

数据集与评测指标

本次实验有4个指标

Accuracy, Precision, Recall, F1

对于二元分类，结果只有以下4种情况

TP：本来+1，预测为+1

FN：本来+1，预测为-1

FP：本来-1，预测为+1

TN：本来-1，预测为-1

T: Ture F: False N: negative P: positive

数据集与评测指标

对于二元分类，结果只有以下4种情况

TP：本来+1，预测为+1

FN：本来+1，预测为-1

FP：本来-1，预测为+1

TN：本来-1，预测为-1

$$\text{Accuracy}=\frac{TP+TN}{TP+FP+TN+FN} \qquad \text{Recall}=\frac{TP}{TP+FN}$$

$$\text{Precision}=\frac{TP}{TP+FP} \qquad F1=\frac{2*\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}}$$

实验任务与提交要求

任务：

1. 在提供的数据集上实现PLA原始算法和口袋算法两种 （PS：请尽量）
2. 采用4种评测指标评价你的实验结果
   （PS：请尽量将你的评测指标计算封装成函数，因为将不止这一次会使用到它们）
3. 优化与分析

实验任务与提交要求

提交要求：

报告篇

1. 请按照这次实验给出的模板进行报告撰写！
2. 4种评测指标的数据分析
3. 实验思路（推荐伪代码）
4. 详细描述创新点（如果有）及优化前后对比
5. 回答实验报告模板里的3个问题
6. 命名：学号_拼音名字.pdf

实验任务与提交要求

提交要求：

代码篇

1. 提交原始PLA版本和pocket PLA两个版本算法，做过优化的话，请选择你认为最优的。
2. 命名：将PLA_initial_学号.xxx 和PLA_pocket_学号.xxx 两个文件压缩成 学号_拼音名字.zip 提交

实验任务与提交要求

提交要求：

注意事项

1. 截止日期：2016年10月30日23点59分59秒，超过则视为迟交
2. FTP地址：ftp://my.ss.sysu.edu.cn/~ryh
3. 抄袭，双方均0分