实验四　利用 SVM 实现分类实验

实验目标：理解 SVM 的分类原理；

能根据数据集设计合理的 SVM 分类方法；

准确评估分类器精度。

实验步骤：

一、SVM 分类原理

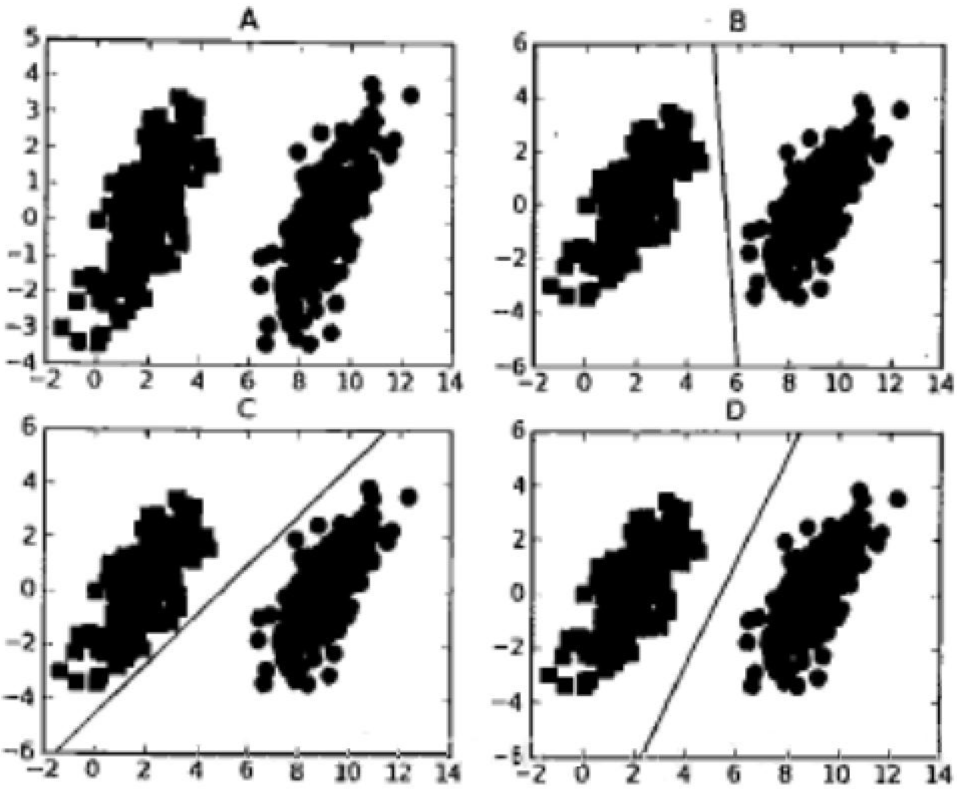
支持向量机(Support Vector Machines)是目前被认为最好的现成的算法之一。

优点：泛化错误率低，计算开销不大，结果易解释。

缺点：对参数调节和核函数的选择敏感，原始分类器不加修改仅适用于处理二类问题。

适用数据类型：数值型和标称型数据。

1. 线性可分数据的 SVM 分类



上图为线性可分数据，将数据集分隔开来的直线称为分隔超平面，也就是分类的决策边界，表示为 $\omega^T x + b$ 。分布在超平面一侧的所有数据都属于某个类别，而分布在另一侧的所有数据则属于另一

个类别。

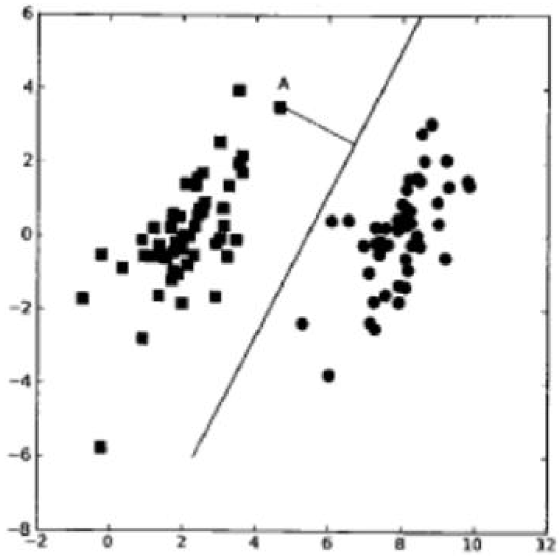
➢ **支持向量**

我们希望找到离分隔超平面最近的点，确保它们离分隔面的距离尽可能远，从而使分类器尽可能健壮。

支持向量（support.vector)就是离分隔超平面最近的那些点。接下来要试着最大化支持向量到分隔面的距离，需要找到此问题的优化求解方法。

➢ **函数间隔和几何间隔**

点 A 到分隔面的距离，即点到分隔面的法线或垂线的长度被称为间隔(margin)，值为 $\left|\omega^T A+b\right| / \|\omega\|$。



我们的**单样本的函数间隔**定义为：

$$\hat{\gamma}^{(i)}=y^{(i)}\left(\omega^T x+b\right)=y^{(i)} f(x)$$

总体的函数间隔： $\hat{\gamma}=\min_{i=1, \ldots \ldots, m} \hat{\gamma}^{(i)}$

$y^{(i)}$即 label，函数间隔就是类别标签乘上了 f(x)的值，这样保证数据点无论出于哪一类（+1 类或-1 类），$label *\left(\omega^T x+b\right)$都会是一个正数。

上述定义的函数间隔虽然可以表示分类预测的正确性和确信度，但在选择分类超平面时，只有函数间隔还远远不够，因为如果成比例的改变 w 和 b，如将他们改变为 2w 和 2b，虽然此时超平面没有改变，但函数间隔的值却变成了原来的 4 倍。

在实际中，我们定义点到超平面的距离时，通常采用几何间隔。

单样本几何间隔定义：$\gamma^{(i)} = y^{(i)}\left(\left(\frac{\omega}{\|\omega\|}\right)^T x^{(i)} + \frac{b}{\|\omega\|}\right)$

总体几何间隔：$\gamma = \min_{i=1,\ldots\ldots,m} \gamma^{(i)}$

由此可得样本总体的函数距离与几何距离的关系式： $\gamma = \frac{\hat{\gamma}}{\|\omega\|}$

- **样本总体几何距离最大化**

目标是找出分类器定义中的ω和 b，我们要先找到具有最小间隔的数据点，然后对该间隔最大化：

$$\arg\max_{\omega,b}\left\{\min_{n}(label \cdot \left(\omega^T + b\right)) \cdot \frac{1}{\|\omega\|}\right\}$$

$$\textbf{s.t.}\ \ label * \left(\omega^T + b\right) \geq 1$$

我们可以令γ=1，即将支持向量的函数间隔为 1，相当于对 ω^T 、b 进行了缩放，对超平面的确定并没有影响。

优化问题变为： $\max_{\omega,b} \frac{1}{\|\omega\|}$ $\quad s.t. \quad \left(\omega^T X^{(i)} + b\right) \geq 1, i = 1,2,3,\ldots,m$

或 $\min\limits_{\omega,b} \frac{\|\omega\|^2}{2} \quad s.t. \quad 1 - y^{(i)} \cdot \left(\omega^T X^{(i)} + b\right) \le 0, i = 1,2,3,\ldots,m$

引入拉格朗日乘子，将超平面写成数据点的形式：

$$\max_{\alpha}\left[\sum_{i=1}^{m}\alpha - \frac{1}{2}\sum_{i,j=1}^{m} label^{(i)} \cdot label^{(j)} \cdot \alpha_i \cdot \alpha_j \left\langle x^{(i)}, x^{(j)}\right\rangle\right]$$

$$s.t. \quad \alpha \ge 0, \sum_{i=1}^{m}\alpha_i \cdot label^{(i)} = 0$$

针对数据不能 100%线性可分，我们引入松弛变量 C，此时约束条件：

$$C \ge \alpha \ge 0, 且 \sum_{i=1}^{m}\alpha_i \cdot label^{(i)} = 0$$

SVM 的主要工作就是求解所有的 α_i，从而表达出分隔超平面。

- **SMO（序列最小优化）**

对偶函数最后的优化问题，SMO 算法目标在于求出一系列 α_i和b，主要工作原理是：每次循环中选择两个 α_i, α_j 进行优化处理，一旦找到一对合适的，就增大其中一个同时减小另一个。

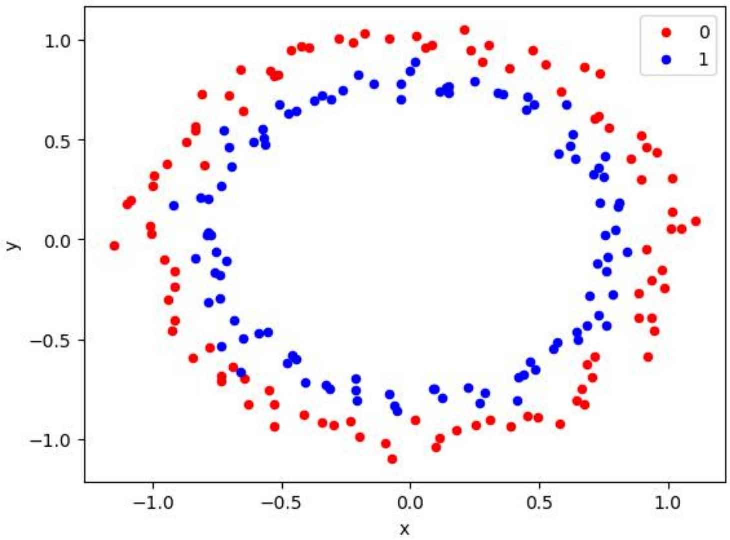
参考 SMO 的主要步骤：

Repeat till convergence {

1. Select some pair α_i and α_j to update next (using a heuristic that tries to pick the two that will allow us to make the biggest progress towards the global maximum).
2. Reoptimize $W(\alpha)$ with respect to α_i and α_j, while holding all the other α_k's $(k \neq i, j)$ fixed.

}

2. 非线性可分数据的 SVM 分类



上图数据点处于一个圆中，对于分类器而言，如果只在 x 和 y 轴构成的坐标系中插入直线进行分类的话，并不会得到理想的结果。

可以利用**核函数**，将数据从某个特征空间到另一个特征空间的映射，在新空间下，可以很容易利用已有的工具对数据进行处理。在通常情况下，这种映射会将低维特征空间映射到高维空间。

➢ **径向基函数（RBF）**

径向基函数是 SVM 中常用的一个核函数，采用向量作为自变量的函数，能基于向量距离运算输出一个标量。这个距离可以是从<0,0>向量或者其它向量开始计算的距离。径向基函数的高斯版本公式：

$$k(x,y)=\exp\left(\frac{-\|x-y\|^2}{2\sigma^2}\right)$$

σ 是用户定义的用于确定到达率（reach）或者说函数值跌落到 0 的速度参数。

二、使用 SVM 进行数据点分类

1. 利用 SMO 优化算法，对线性可分数据进行分类

    数据来源：testSet.txt

    简化版 SMO 算法：请参考文件 SMO_simp.py

2. 针对非线性分类数据，使用核技巧，选用 RBF 进行分类

    数据来源：testSetRBF.txt

RBF 分类器的设计：kernel.py

三、实验要求

1. 参考代码及数据来源如上面文件，请调试并查看结果。
2. 请写出简化 SMO 算法的伪代码或者流程图？ 调节 C 的大小，可以发现什么现象？
3. 使用核技巧，更换不同的 kl 参数，请问对应的测试错误率、训练错误率、支持向量个数随 kl 的变化情况是什么？

注意：请将实验报告以文件形式提交到 QQ 群里，统一命名为：学号+姓名+实验几.doc/rar/zip/pdf