LinearSVC算法介绍

支持向量机（support vector machines, SVM）是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；SVM还包括核技巧，这使它成为实质上的非线性分类器。SVM的的学习策略就是间隔最大化，说的通俗一点就是就是在特征空间里面用某条线或某块面将训练数据集分成两类，而依据的原则就是间隔最大化，这里的间隔最大化是指特征空间里面距离分离线或面最近的点到这条线或面的间隔(距离)最大。

支持向量机（SVM）是个非常强大并且有多种功能的机器学习模型，能够做线性或者非线性的分类，回归，甚至异常值检测。特别适合应用于复杂但中小规模数据集的分类问题。

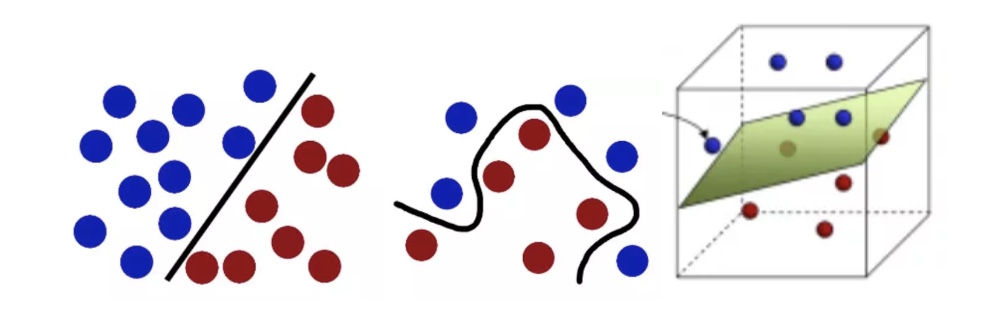
LinearSVC算法是是一种线性分类支持向量机，它是开源机器学习库Sklearn中的SVM算法的一种实现，主要用于线性可分的数据集。

1、SVM算法原理

支持向量机是一种二分类模型，它的基本模型是定义在特征空间上的间隔最大的线性分类器。

说的通俗一点就是就是在特征空间里面用某条线或某块面将训练数据集分成两类，而依据的原则就是间隔最大化，这里的间隔最大化是指特征空间里面距离分离线或面最近的点到这条线或面的间隔(距离)最大。

看下面的图来感受一下，SVM的目的就是要找打能把红色点和蓝色点准确分开的线或面。



**间隔最大化的直观解释：**对训练数据集找到几何间隔最大的超平面意味着以充分大的确信度对训练数据进行分类。也就是说，不仅将正负实例点分开，而且对最难分的实例点(即离平面最近的点)也有足够大的确信度将他们分开。这样的超平面对未知的新实例有很好的分类预测能力。

2、几个基本概念

2.1、线性/非线性

线性是指量与量之间按比例，成直线关系，在数学上可理解为一阶导数为常数的函数；而非线性是指不按比例，不成直线关系，一阶导数不为常数的函数。

2.2、线性可分/线性不可分

对于二分类问题，有那么**一条直线**可以把正负实例点**完全分开，**这些数据就是线性可分的；而线性不可分就是找不到一条直线可以把正负实例点完全分开。

2.3、超平面

其实就是实例点从二维空间转移到三维甚至多维空间中，这个时候不能再用直线划分，而需要用平面去划分数据集，这个平面就称为超平面。

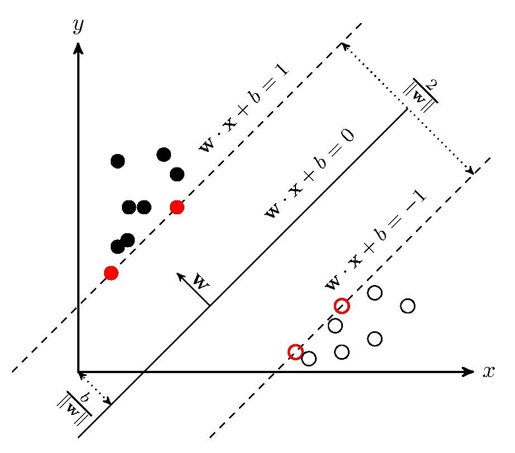
2.4、支持向量

在线性可分的情况下，训练数据集的样本点与分离超平面距离最近的样本点称为支持向量，而支持向量机的目的就是求取距离这个点最远的分离超平面，这个点在确定分离超平面时起着决定性作用，所以把这种分类模型称为支持向量机。

3、线性可分支持向量机

线性可分支持向量机是有一条直线(平面)可以将训练数据集完全分离开来，而线性可分支持向量机的学习目的就是找到那一条直线（平面）。

SVM学习的基本想法是求解能够正确划分训练数据集并且几何间隔最大的分离超平面。如下图所示，即为分离超平面，对于线性可分的数据集来说，这样的超平面有无穷多个（即感知机），但是几何间隔最大的分离超平面却是唯一的。



在推导之前，先给出一些定义。假设给定一个特征空间上的训练数据集

其中，,为第i个特征向量，为类标记，当它等于+1时为正例,为-1时为负例,再假设训练数据集是线性可分的。

几何间隔：对于给定的数据集T和超平面,定义超平面关于样本点的几何间隔为

超平面关于所有样本点的几何间隔的最小值为

实际上这个距离就是我们所谓的支持向量到超平面的距离。

根据以上定义，SVM模型的求解最大分割超平面问题可以表示为以下约束最优化问题

,

将约束条件两边同时除以，得到

因为都是标量，所以为了表达式简洁起见，令

得到

又因为最大化 ，等价于最大化, 也就是等价于最小化 ，（），因此SVM模型的求解最大分割超平面问题又可以表示为以下约束最优化问题

这是一个含有不等式约束的凸二次规划问题，可以对其使用拉格朗日乘子法得到其对偶问题（dual problem）。

首先，我们将有约束的原始目标函数转换为无约束的新构造的拉格朗日目标函数

其中为拉格朗日乘子，且。现在我们令

当样本点不满足约束条件时，即在可行解区域外：

此时，将设置为无穷大，则也为无穷大。

当满本点满足约束条件时，即在可行解区域内：

此时，为原函数本身。于是，将两种情况合并起来就可以得到我们新的目标函数

于是原约束问题就等价于

看一下我们的新目标函数，先求最大值，再求最小值。这样的话，我们首先就要面对带有需要求解的参数的方程，而又是不等式约束，这个求解过程不好做。所以，我们需要使用拉格朗日函数**对偶性**，将最小和最大的位置交换一下，这样就变成了：

要有, 需要满足两个条件：

1、优化问题是凸优化问题

2、 满足KKT条件

首先，本优化问题显然是一个凸优化问题，所以条件一满足，而要满足条件二，即要求

为了得到求解对偶问题的具体形式，令对的偏导为0，可得

将以上两个等式带入拉格朗日目标函数，消去,得

即

求对的极大，即是对偶问题

把目标式子加一个负号，将求解极大转换为求解极小

现在我们的优化问题变成了如上的形式。

我们通过这个优化算法能得到,再根据，我们就可以求解出，进而求得我们最初的目的：找到超平面，即”决策平面”。

前面的推导都是假设满足KKT条件下成立的，KKT条件如下

另外，根据前面的推导，还有下面两个式子成立

由此可知在中，至少存在一个, 反证法可以证明，若全为0，则,矛盾,对此有

因此可以得到

对于任意训练样本,总有，或者

若，则该样本不会在最后求解模型参数的式子中出现。若，则必有

所对应的样本点位于最大间隔边界上，是一个支持向量。这显示出支持向量机的一个重要性质：训练完成后，大部分的训练样本都不需要保留，最终模型仅与支持向量有关。

到这里都是基于训练集数据线性可分的假设下进行的，但是实际情况下几乎不存在完全线性可分的数据，为了解决这个问题，引入了“软间隔”的概念，即允许某些点不满足约束

采用hinge损失，将原优化问题改写为

其中为“松弛变量”，, 即一个hinge损失函数。每一个样本都有一个对应的松弛变量，表征该样本不满足约束的程度。C>0称为惩罚参数，C值越大，对分类的惩罚越大。跟线性可分求解的思路一致，同样这里先用拉格朗日乘子法得到拉格朗日函数，再求其对偶问题。

综合以上讨论，我们可以得到**线性支持向量机学习算法**如下：

**输入**：训练数据集其中，

**输出**：分离超平面和分类决策函数

（1）选择惩罚参数C > 0, 构造并求解凸二次规划问题

得到最优解

(2)计算

选择的一个分量满足条件，计算

（3）求分离超平面

分离决策函数：