课程5：
支持向量机
（SVM：Support Vector Machine）

支持向量机应用

分类、回归、密度估计

- **手写字符识别**
- **文本自动分类**
- **人脸识别**
- **时间序列预测**
- **蛋白质识别**
- **DNA 排列分析**

支持向量机定义

所谓支持向量机，顾名思义，分为两个部分了解：

一，什么是支持向量（简单来说，就是支持或支撑平面上把两类类别划分开来的超平面的向量点）

二，这里的“机（machine，机器）”便是一个算法。在机器学习领域，常把一些算法看做是一个机器，如分类机（当然，也叫做分类器），而支持向量机本身便是一种监督式学习的方法，它广泛的应用于统计分类以及回归分析中。

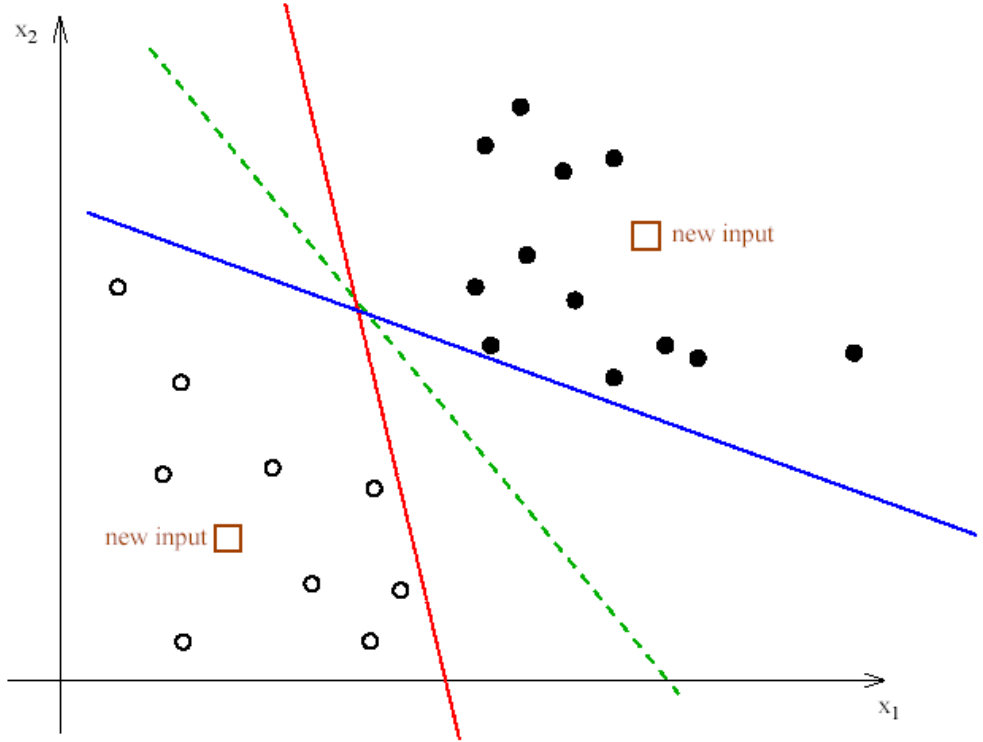
SVM的描述

目标：找到一个超平面，使得它能够尽可能多的将两类数据点正确的分开，同时使分开的两类数据点距离分类面最远。

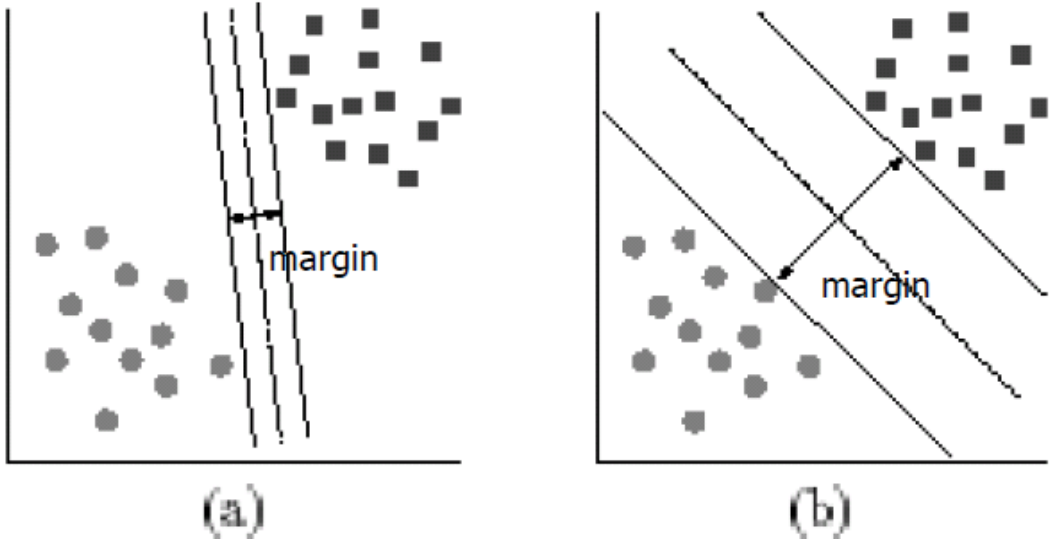
解决方法：构造一个在约束条件下的优化问题，具体的说是一个约束二次规划问题(constrained quadratic programing)，求解该问题，得到分类器。

1. 线性可分情况

线性可分情况



最大边缘超平面



从超平面到其边缘的侧面的最短距离等于到其边缘的另一个侧面的最短距离，边缘侧面平行于超平面

数学表达：

分类面与边界距离（margin）的数学表示：

$(x_1, y_1), (x_2, y_2), \cdots (x_n, y_n) \qquad x_i \in \Re^d, \quad y_i \in \{-1, 1\}$

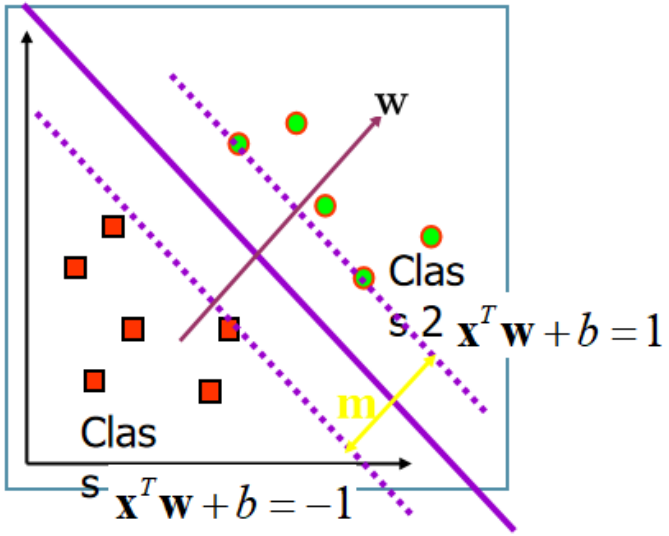
$y_i = 1$ 表示 $x_i \in \omega_1$；$y_i = -1$ 表示 $x_i \in \omega_2$

分类超平面表示为：

超平面方程：$\mathbf{x}^T\mathbf{w} + b = 0$

任意一点到超平面的距离：$r = \dfrac{|w^T x + b|}{\|w\|}$

超平面间隔：$m = \dfrac{2}{\|\mathbf{w}\|}$



数学推导：

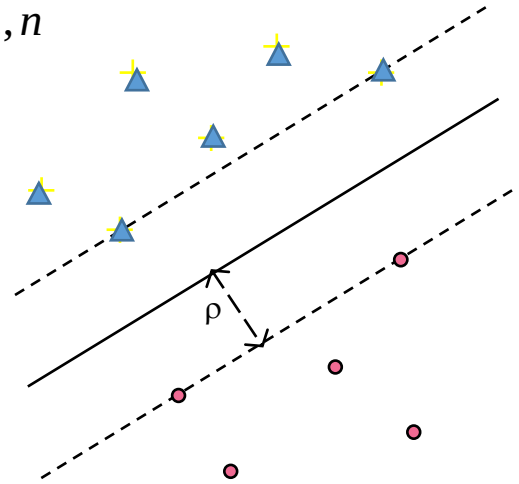
设有两类样本的训练集：

$$D=\{(x_1,y_1),(x_2,y_2),...,(x_n,y_n)\}$$

$$x_i \in X \subset R^m, y_i \in \{1,-1\}, \quad i=1,...,n$$

线性可分情况意味着存在超平面使训练点中的正类和负类样本分别位于该超平面的两侧。

$$(w\cdot x)+b=0$$



如果能确定这样的参数对（w, b）的话，就可以构造决策函数来进行识别新样本。

$$f(x)=\mathrm{sgn}((w\cdot x)+b)$$

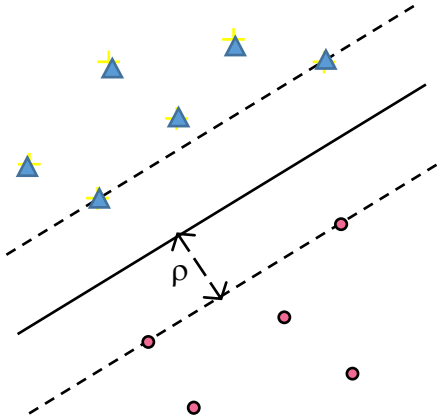
使用最大间隔原则，取使得间隔最大的参数对（w,b）：

在规范化下，超平面的几何间隔为：

$$\max_{w,b} \frac{2}{\|w\|}$$

$$x_i^T w + b \geq 1 \qquad y_i = 1$$

$$x_i^T w + b \leq -1 \qquad y_i = -1,$$



为方便计算将上述max函数改为下述min函数：

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2 \qquad (1)$$

$$s.t. \quad y_i((w \cdot x_i) + b) \geq 1, i = 1, \cdots, n$$

拉格朗日乘子法

（1）等式约束条件

$$\min f(\vec{x}) \qquad L(\vec{x},\lambda) = f(\vec{x}) + \lambda g(\vec{x})$$

$$s.t.\ g(\vec{x}) = 0$$

（2）不等式约束条件

$$\min f(\vec{x}) \qquad L(\vec{x},\lambda) = f(\vec{x}) + \lambda g(\vec{x})$$

$$s.t.\ g(\vec{x}) \le 0$$

（3）总结

$$\min f(\vec{x})$$

$$s.t.\ (1)\ g_i(x) \le 0, i = 1...m$$

$$(2)\ h_i(x) \le 0, i = 1...k$$

$$L(\vec{x},\alpha,\beta) = f(\vec{x}) + \sum_{i=1}^{m} \alpha_i g_i(\vec{x}) + \sum_{i=1}^{k} \beta_i h_i(\vec{x})$$

对偶问题

（1）弱对偶问题

原问题　　对偶问题

$$\min_{x}\max_{\alpha,\beta} L(\vec{x},\vec{\alpha},\vec{\beta}) \quad \geq \quad \max_{\alpha,\beta}\min_{x} L(\vec{x},\vec{\alpha},\vec{\beta})$$

（2）强对偶问题

原问题　　对偶问题

$$\min_{x}\max_{\alpha,\beta} L(\vec{x},\vec{\alpha},\vec{\beta}) \quad = \quad \max_{\alpha,\beta}\min_{x} L(\vec{x},\vec{\alpha},\vec{\beta})$$

支持向量机的KKT条件

$$a_i \geq 0$$

$$1 - y_i(w^T x_i + b) \leq 0$$

$$a_i(1 - y_i(w^T x_i + b)) = 0$$

$$\frac{\partial L}{\partial W} = 0 \ , \ \frac{\partial L}{\partial b} = 0$$

引入Lagrange函数，使用Lagrange乘子法将其转化为对偶问题。于是引入Lagrange函数：

$$L(w,b,\alpha)=\frac{1}{2}\|w\|^2-\sum_{i=1}^{n}\alpha_i(y_i((w\cdot x_i)+b)-1) \quad (2)$$

其中，$\alpha=(\alpha_1,\alpha_2,...,\alpha_n)^T\in R_+^n$ 称为Lagrange乘子

首先求Lagrange函数关于w, b的极小值。由极值条件有：

$$\nabla_b L(w,b,\alpha)=0,\quad \nabla_w L(w,b,\alpha)=0$$

得到：

$$\sum_{i=1}^{n}y_i\alpha_i=0 \quad (3)$$

$$w=\sum_{i=1}^{n}y_i\alpha_i x_i \quad (4)$$

将(3)式代入Lagrange函数，并利用(4)式，则原始的优化问题转化为如下的对偶问题(使用极小形式)

$$L(\alpha)=\max_{\alpha}\sum_{j=1}^{n}\alpha_j-\frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}y_i y_j\alpha_i\alpha_j(x_i\cdot x_j) \quad (5)$$

$$s.t. \quad \sum_{i=1}^{n}y_i\alpha_i=0,$$

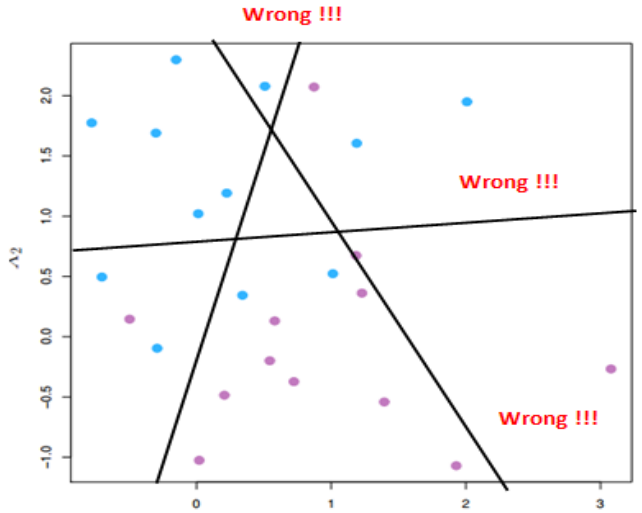
$$\alpha_i\geq 0, \qquad i=1,\cdots,n$$

软间隔

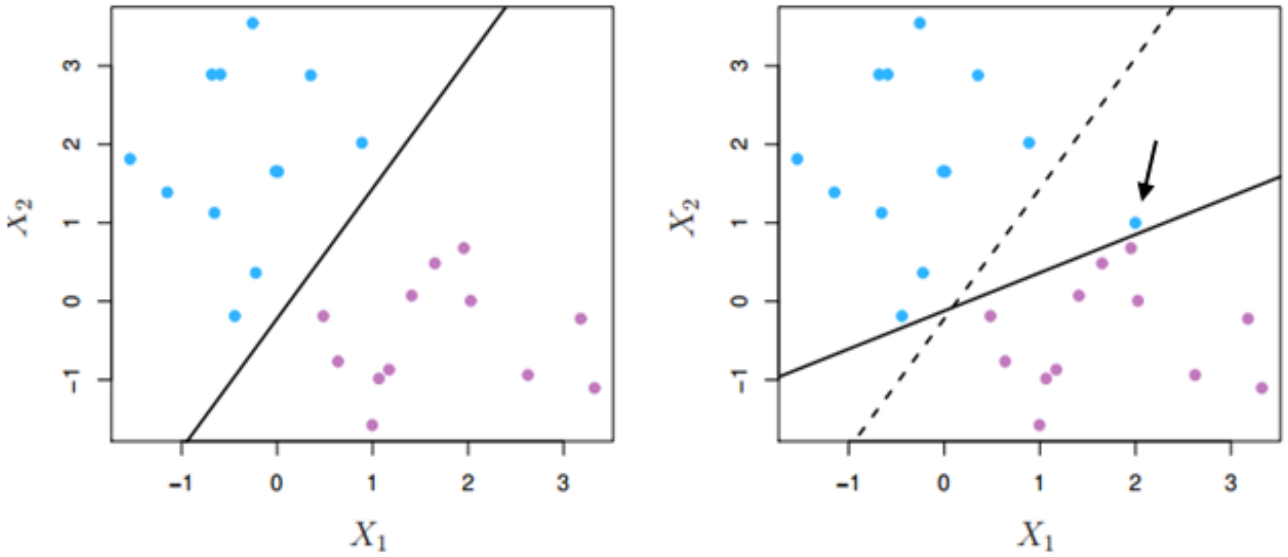
硬间隔有两个缺点：

1. 不适用于线性不可分数据集
2. 对离群点（outlier）敏感

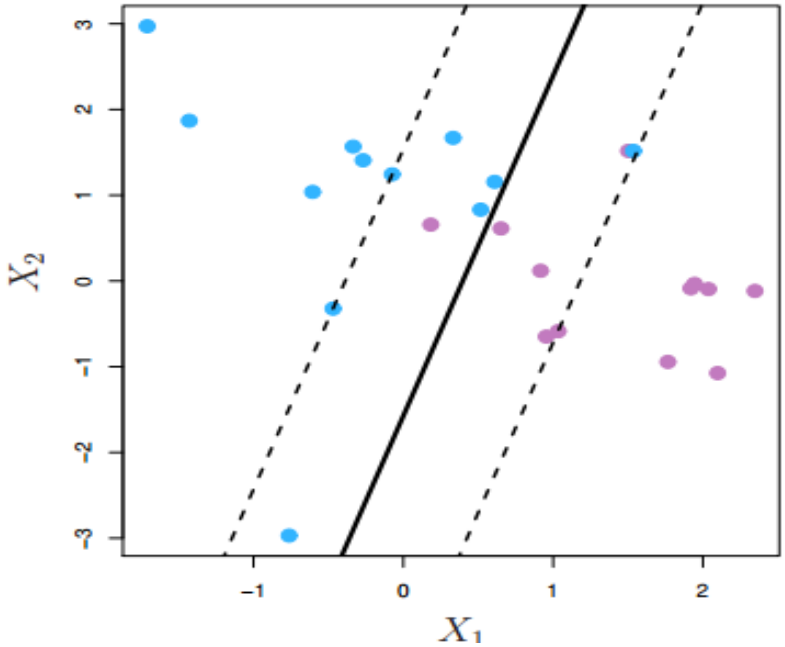
比如下图就无法找到一个超平面将蓝点和紫点完全分开：



下图显示加入了一个离群点后，超平面发生了很大的变动，最后形成的间隔变得很小，这样最终的泛化效果可能不会太好。



为了缓解这些问题，引入了“软间隔（soft margin）”，即允许一些样本点跨越间隔边界甚至是超平面。如下图中一些离群点就跨过了间隔边界。



于是为每个样本点引入松弛变量 ξi，优化问题变为：

$$\min_{w,b} \quad \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i$$

$$s.t. \quad y_i((w\cdot x_i)+b) \geq 1-\xi_i, \quad i=1,\cdots,n$$

$$\xi_i \geq 0, \quad i=1,2,3,...,m$$

引入拉格朗日乘子：

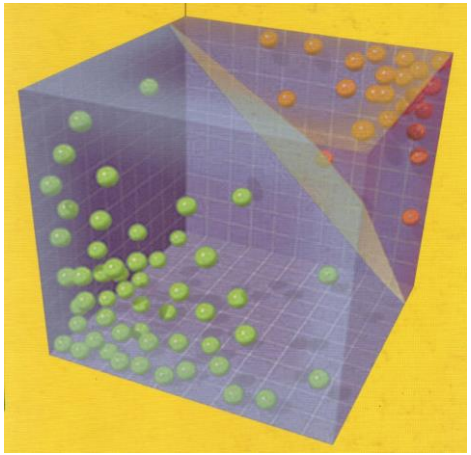
$$L(w,b,\xi,\alpha,\beta) = \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{m}\xi_i + \sum_{i=1}^{m}\alpha_i[1-\xi_i - y_i(w^T x_i + b)] + \sum_{i=1}^{m}\beta_i(-\xi_i)$$

$$\alpha_i \geq 0, \beta_i \geq 0$$

2. 非线性可分情况

例子

$$a\,x_1^2 + b\,x_2^2 = 1 \longrightarrow [w]_1 z_1 + [w]_2 z_2 + [w]_3 z_3 + b = 0$$



非线性分类

设训练集 $T=\{(x_i,y_i),i=1,\cdots l\}$，其中 $x_i=([x_i]_1,[x_i]_2)^T, y_i\in\{1,-1\}$

假定可以用 $([x]_1,[x]_2)$ 平面上的二次曲线来划分：

$$[w]_1+2[w]_2[x]_1+2[w]_3[x]_2+2[w]_4[x]_1[x]_2+[w]_5[x]_1^2+[w]_6[x]_2^2+b=0$$

现考虑把2维空间 $x=([x]_1,[x]_2)^T$ 映射到6维空间的变换

$$\phi(x)=(1,\sqrt{2}[x]_1,\sqrt{2}[x]_2,\sqrt{2}[x]_1[x]_2,[x]_1^2,[x]_2^2)^T$$

上式可将2维空间上二次曲线映射为6维空间上的一个超平面：

$$[w]_1[X]_1+\sqrt{2}[w]_2[X]_2+\sqrt{2}[w]_3[X]_3+\sqrt{2}[w]_4[X]_4+[w]_5[X]_5+[w]_6[X]_6+b=0$$

非线性分类

可见，只要利用变换，把x所在的2维空间的两类输入点映射x 所在的6维空间，然后在这个6维空间中，使用线性学习机求出分划超平面：

$$(w^* \cdot x) + b^* = 0, \text{ 其中 } w^* = ([w^*]_1, \cdots [w^*]_6)^T$$

最后得出原空间中的二次曲线：

$$[w^*]_1 + 2[w^*]_2[x]_1 + 2[w^*]_3[x]_2 + 2[w^*]_4[x]_1[x]_2 + [w^*]_5[x]_1^2 + [w^*]_6[x]_2^2 + b = 0$$

核函数

求解最优化问题

$$\min_{\alpha} \quad \frac{1}{2}\sum_{i=1}^{l}\sum_{j=1}^{l} y_i y_j \alpha_i \alpha_j \left(\phi(x_i)\cdot\phi(x_j)\right) - \sum_{j=1}^{l}\alpha_j$$

$$s.t. \quad \sum_{i=1}^{l} y_i \alpha_i = 0$$

$$0 \le \alpha_i \le C, i = 1, \cdots l$$

得到最后的决策函数：

$$f(x) = \mathrm{sgn}(\sum_{i=1}^{l} y_i \alpha_i (\phi(x_i) \bullet \phi(x)) + b)$$

为此引入核函数：

$$\mathrm{K}(x_i, x_j) = (\phi(x_i)\cdot\phi(x_j)) = ((x_i \cdot x_j) + 1)^2$$

核函数的选择

目前研究最多的核函数主要有三类：

- 多项式内核

$$K(x, x_i) = [(x \cdot x_i) + c]^q$$ 得到q阶多项式分类器

- 高斯径向基函数内核RBF

$$K(x, x_i) = \exp\{-\frac{|x - x_i|^2}{\sigma^2}\}$$

每个基函数中心对应一个支持向量，它们及输出权值由算法自动确定

- Sigmoid内核

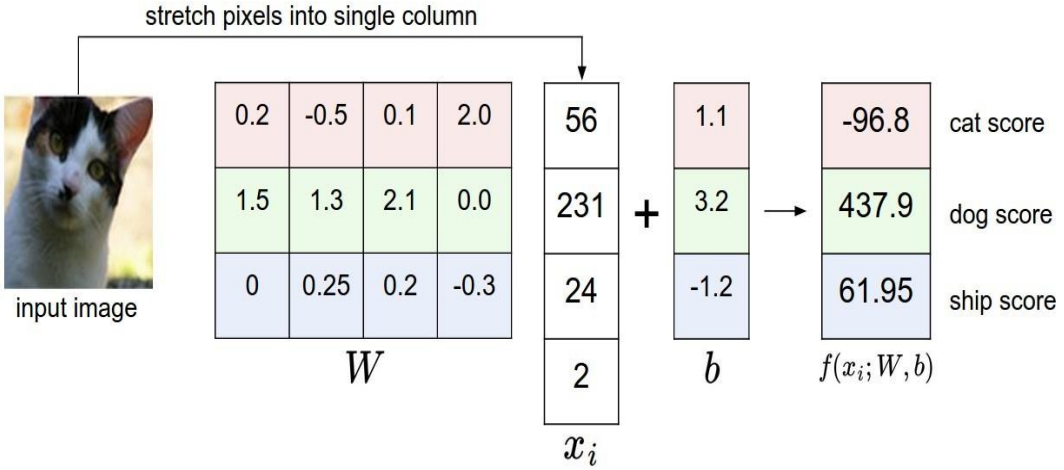
$$K(x, x_i) = \tanh(\upsilon(x \cdot x_i) + c)$$

包含一个隐层的多层感知器，隐层节点数是由算法自动确定

多分类svm

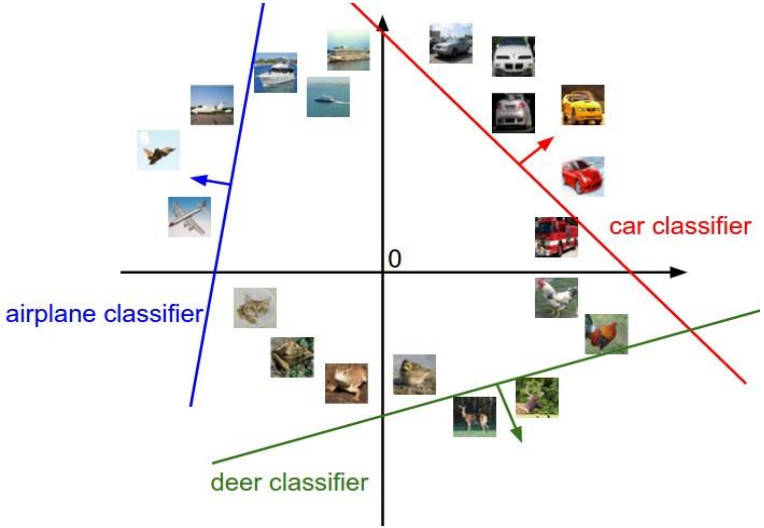
构造一个线性分类器：

$$f(x_i, W, b) = Wx_i + b$$



有3个分类（红色代表猫，绿色代表狗，蓝色代表船，注意，这里的红、绿和蓝3种颜色仅代表分类，和RGB通道没有关系）。首先将图像像素拉伸为一个列向量，与W进行矩阵乘，然后得到各个分类的分值。需要注意的是，这个W一点也不好：猫分类的分值非常低。从上图来看，算法倒是觉得这个图像是一只狗。

而线性分类器就是在高维度空间中的一个超平面，将各个空间点分开，如下图所示：



而我们要做的就是寻找一个W和一个b，使得这个超平面能很好的区分各个类。寻找方法就是不停的改变w和b的值，即不停的旋转平移，直到它使分类的偏差较小。

损失函数：

$$L_i = \sum_{j \neq y_i} max(0, w_j^T x_i - w_{yi}^T x_i + \Delta)$$

举例：用一个例子演示公式是如何计算的。假设有3个分类，并且得到了分值s=[13, −7, 11]。其中第一个类别是正确类别，即yi=0。同时假设Δ是10。上面的公式是将所有不正确分类加起来，所以得到两个部分：

Li=max(0,−7−13+10)+max(0,11−13+10)

正则化：为了降低过拟合，我们加上正则化。

$$L = \frac{1}{N} \sum_i \sum_{j \neq yi} [\max(0, f(x_i; W)_j - f(x_i; W)_{yi} + \Delta)] + \lambda \sum_k \sum_l W_{k,l}^2$$

支持向量机的优缺点

优点：

- 能够处理大型特征空间
- 能够处理非线性特征之间的相互作用
- 无需依赖整个数据，只需依赖支持向量
- 占用内存小，因为不需要保留所有样本，只需要保留支持向量即可，能够获得一个较好的效果。
- 泛化性能良好，学习效果具有较好的推广性

缺点：

- 当观测样本很多时，效率并不是很高
- 有时候很难找到一个合适的核函数

问题：

（1）SVM的原理是什么：

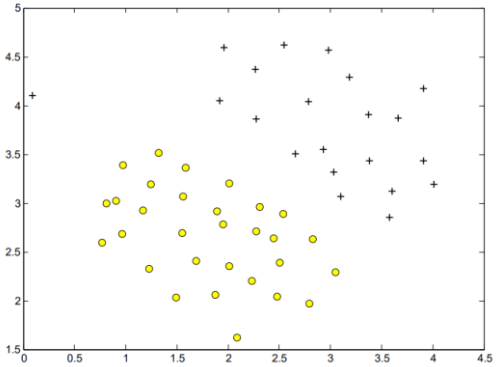
支持向量机是一种二分类模型，它的基本模型定义在特征空间上的间隔最大的线性分类器，间隔最大使它有别于感知机；支持向量机还包括核技巧，这使他成为实质上的非线性分类器。支持向量机的学习策略就是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数的最大化问题。支持向量机的学习算法是求解凸二次规划的最优化算法。

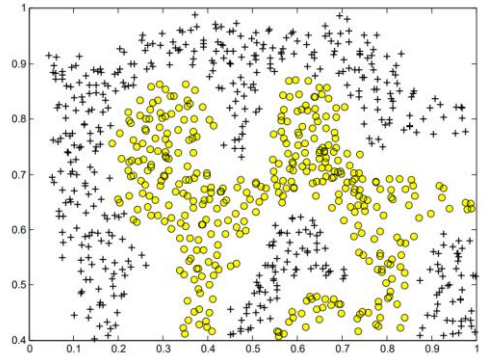
（2）SVM为什么采用间隔最大化?

支持向量机学习的基本思想是求解能够正确划分训练集并且几何间隔最大的最大超平面。对线性可分数据集而言，线性可分分离超平面有无穷多个，但是几何间隔最大的分离超平面是唯一的。

间隔最大化的直观解释是： 对训练集找到几何间隔最大的超平面意味着以充分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(离超平面最近的点)也有最大的确信度将它们分开**。这样的超平面应该对未知的新实例有很好的分类预测能力(具有良好的鲁棒性)。

作业4:基于SVM的垃圾邮件分类





- 推荐编程环境:Anaconda+Jupyter notebook

  安装教程： 点这

- 用SVM解决线性可分和非线性可分数据集

参考文献：

1. Aharo M. M. Elad, and A. Bruckstein. K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation.[C] IEEE Transctions on Image Processing, 54(11),2006:4311-4322.
2. Akaike, H. A new look at the statistical model identification[J]. IEEE Transαctions on Automatic Control, 19(6), 1974:716-723.
3. Baraniuk, R. G. Compressive sensing[J]. IEEE Signal Processing Magzine,24 (4),2007:118-121
4. Bengio, S., F. Pereira, Y. Singer, and D. Strelow. Gro sparse coding[C]. In Advances in Neural Information Processing Systems 22 (NIPS),2009:82-89.
5. Blum, A. and P. Langley. Selection of relevant features and examples in machine learning[J]. Artificial lntelligence, 97(1-2):245-271.

相关论文会放到课程网页中，如有需要请自行下载。