机器学习

Machine Learning

主讲人：张敏 清华大学长聘副教授

ML 机器学习

MACHINE LEARNING-MIN ZHANG

Unit.07

支持向量机（I）

*图片均来自网络或已发表刊物

纲要

- 背景
- 线性支持向量机
- 核函数支持向量机
- 核心概要

分类

★★★★★ **An Awesome Movie!**
By Jokerz Wild on October 9, 2017
Format: Amazon Video | Verified Purchase

I love Iron Man!

★☆☆☆☆ **Comic book characters... making millions of horrible movies these days.**
By TylerVogt3329 on November 14, 2008
Format: DVD

You people these days consider this a good movie? Haha. Who in their right mind believes that a rich playboy can save the world from evil?
For good and REAL action check out WWE, ECW, or TNA.
For good classic Wrestling.. check out WCW and WWF.

★★★★☆ **Good solid film**
By M-M on July 30, 2013
Format: Amazon Video | Verified Purchase

It turned out to be entertaining and at the end I enjoyed the film. Good special effects, nice story line for "actions" and "comics". The protagonist (Tony Stark) looks natural: arrogant, brash, but at the same time clever, intelligent and ethic. The villain is a little bit overreacting, and annoying, as most of antagonists :) Overall that's a good movie.

Pos正面： This skirt is so beautiful!这件裙子真好看！

Neg负面: It looks ugly on my body. 这穿在我身上真难看。

Neural 中性: It is pure cotton. 它是纯棉的。

Rock: Michael Jackson
“Beat it”,
Hip Hop: Eminem
“Lose yourself”
Blues: Muddy Waters
“I can’t be satisfied”

分类方法

- 决策树
  - 样本的属性非数值
  - 目标函数是离散的
- 贝叶斯学习
  - 样本的属性可以是数值或非数值
  - 目标函数是连续的（概率）
- K-近邻
  - 样本是空间（例如欧氏空间）中的点
  - 目标函数可以是连续的也可以是离散的
- 支持向量机 (Support Vector Machine)
  - 样本是空间（例如欧氏空间）中的点
  - 目标函数可以是连续的也可以是离散的

背景信息

- 支持向量机 （Support Vector Machine，SVM）
- 当前版本的支持向量机大部分是由Vapnik和他的同事在AT&T贝尔实验室开发的
- 是一个最大间隔分类器（Max Margin Classifier）
- 最开始提出是做分类，后来很快被应用到了回归和时间序列的预测
- 最有效的监督学习方法之一
  - 曾被作为文本处理方法的一个强基准模型（strong baseline）

纲要

- 背景
- 线性支持向量机
  - 最大间隔线性分类器
  - 形式化对偶问题
  - 线性不可分情况
- 核函数支持向量机
- 核心概要

问题

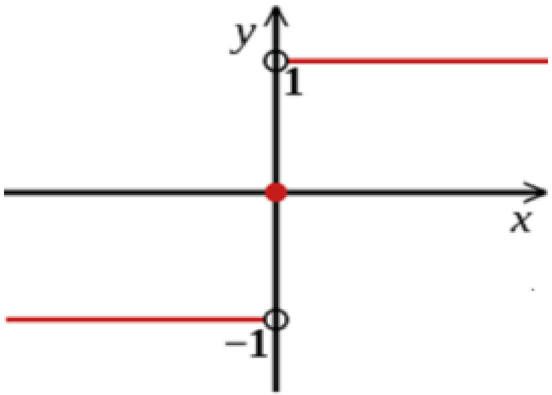
- 给定一个训练样本的集合

$$(x_1,y_1),(x_2,y_2),(x_N,y_N),x_i \in R^n,y_i \in \{-1,1\},$$

找到一个函数$f(x,\theta)$分类样本，使得

$$f(x_i,\theta)\begin{cases} >0, & \forall y_i=+1; \\ <0, & \forall y_i=-1, \end{cases}$$

其中 θ 表示参数　　　（符号函数）



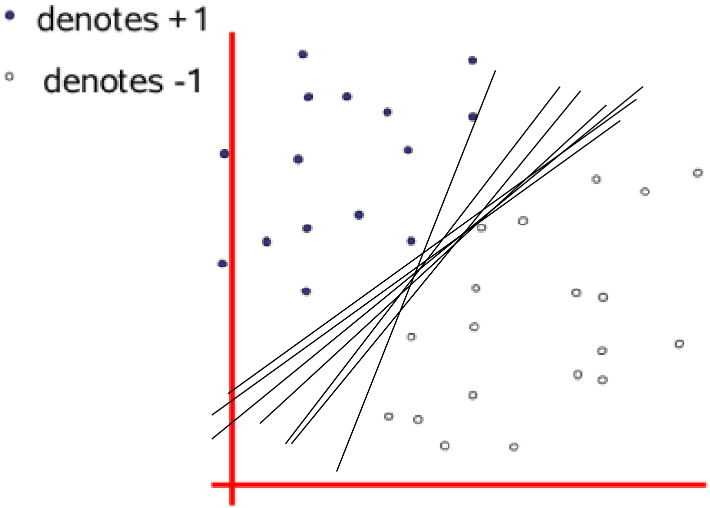
- 对一个测试样本x，我们可以预测它的标签为$[f(x,\theta)]$
- $f(x,\theta)=0$ 被称为 分类超平面（separation hyperplane）

线性分类器

- 线性超平面

$$f\left(x, w, b\right) = \langle x, w \rangle + b = 0$$

- 在线性可分的情况下，有无穷多个满足条件的超平面



如何分类数据?

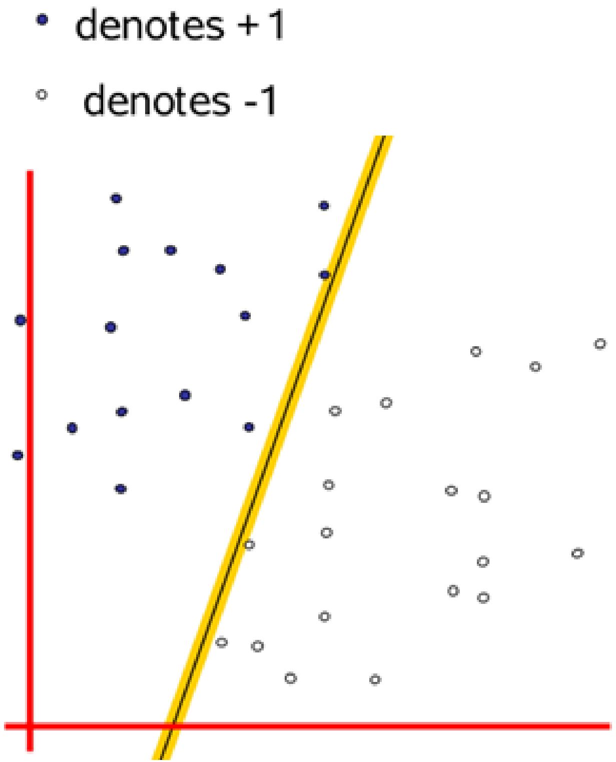
任何一条线都可以。

…… 但哪个最好?

线性分类器的间隔(Margin)

- 在分类分界面两侧分别放置平行于分类超平面的一个超平面， 移动超平面使其远离分类超平面
- 当他们各自第一次碰到数据点时，他们之间的距离被称为线性分类器的间隔

Margin (间隔): 分界在碰到数据点之前可以达到的宽度



最大间隔线性分类器 (Maximum margin linear classifie)

- 定义: 具有最大间隔的线形分类器

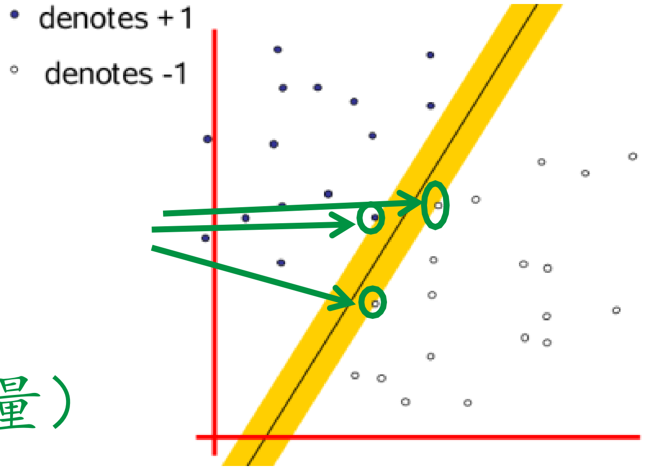
Support

vectors:

（支持向量）

那些阻挡间隔继续

扩大的数据点



问题形式化

- 形式化间隔，我们需要所有数据点满足

$$f(x_i,\alpha)=\langle x_i,w\rangle+b\begin{cases}\geq +1, & \forall y_i=+1;\\ \leq -1, & \forall y_i=-1.\end{cases}$$

（当然也可以不用1而用其他常数）

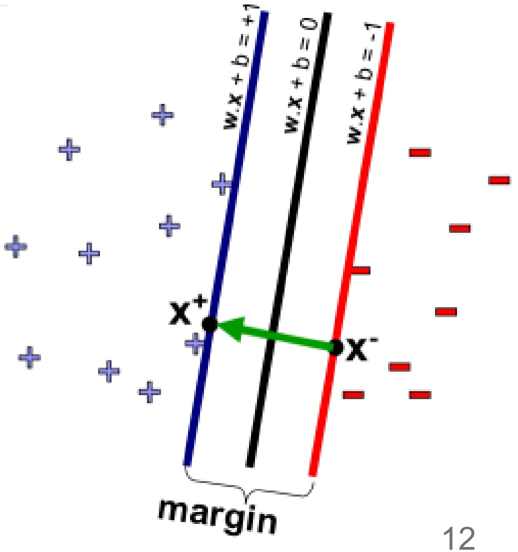
- 或者

$$y_i(\langle x_i,w\rangle+b)\geq 1,\quad \forall i=1,\ldots,N.$$

分类超平面：$\langle x,w\rangle+b=0$

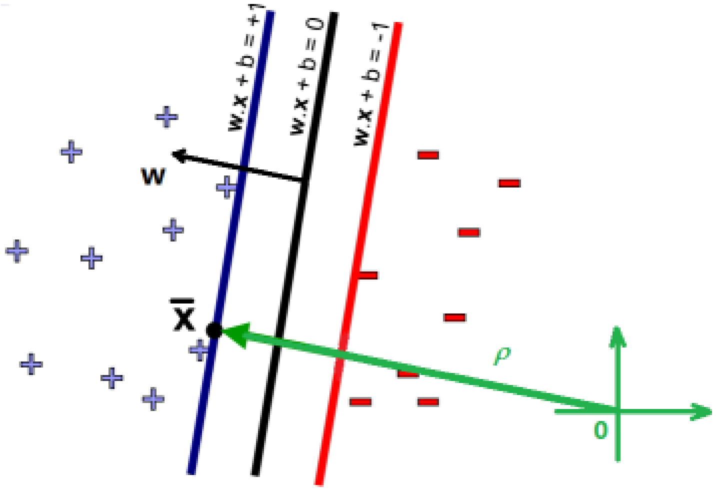
引入平行于分类超平面的两个额外超平面：

$$\langle x,w\rangle+b=\pm 1$$



问题形式化

- 什么是间隔?
  - 两个新的超平面（$\langle x, w \rangle + b = \pm 1$）之间的距离
- 间隔的表达式是什么?



- 从超平面$\langle x, w \rangle + b = 1$ 到原点的最小距离为ρ_1
- 从超平面$\langle x, w \rangle + b = -1$ 到原点的最小距离为 ρ_2

那么间隔是 $|\rho_1 - \rho_2|$

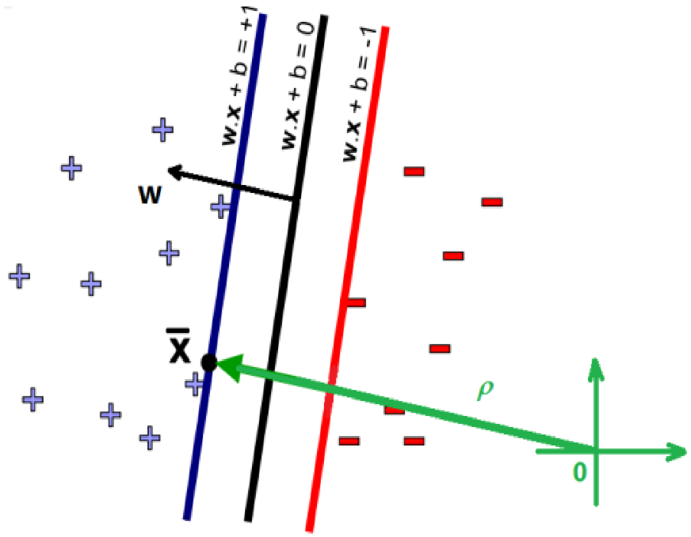
问题形式化

- 如何计算ρ_1 和 ρ_2?
- 记 $\bar{x} = \rho_1 w/||w||_2$， 其中 $w/||w||_2$ 是方向 w 的单位向量
- $\bar{x}$ 在蓝色的超平面， 那么

$$\langle \rho_1 w/||w||_2, w\rangle + b = 1$$

其中有 $\rho_1 = \frac{1-b}{||w||_2}$

- 类似可得 $\rho_2 = \frac{-1-b}{||w||_2}$
- 则间隔为 $|\rho_1 - \rho_2| = \frac{2}{||w||_2}$



问题形式化

- 最优化问题（使间隔$|\rho_1 - \rho_2|$最大化）

$$\begin{aligned} \max_{w,b} \quad & \frac{2}{\|w\|_2} \\ \text{s.t.} \quad & y_i\left(\langle w, x_i \rangle + b\right) \geq 1, \quad \forall i = 1, \ldots, N \end{aligned}$$

- 或等价于

$$\begin{aligned} \min_{w,b} \quad & \frac{1}{2}\|w\|_2^2 \\ \text{s.t.} \quad & y_i\left(\langle w, x_i \rangle + b\right) \geq 1, \quad \forall i = 1, \ldots, N \end{aligned}$$

虽然看起来似乎间隔只与w有关，
但b仍然通过约束w的取值，间接对间隔产生影响

纲要

- 背景
- 线性支持向量机
  - 最大间隔线性分类器
  - 形式化对偶问题
  - 线性不可分情况
- 核函数支持向量机
- 核心概要

Dual problem formulation 对偶问题

- 原始问题（Primal problem）

$$\min_{w,b} \quad \frac{1}{2}\|w\|_2^2$$
$$\text{s.t.} \quad y_i\left(\langle w, x_i\rangle + b\right) \geq 1, \quad \forall i = 1, \ldots, N$$

- 拉格朗日乘子法(Lagrange multipliers):对每条约束添加拉格朗日乘子

Lagrange function （拉格朗日函数）

$$L(w, b, \alpha) = \frac{1}{2}\|w\|_2^2 - \sum_i \alpha_i \left(y_i\left(\langle w, x_i\rangle + b\right) - 1\right) \qquad \alpha_i \geq 0$$

- KKT条件* (Karush[1939]-Khun-Tucker[1951] conditions, Kuhn-Tucker conditions)

$$\begin{cases} \dfrac{\partial L}{\partial w} = w - \sum_i \alpha_i y_i x_i = 0 \\ \dfrac{\partial L}{\partial b} = \sum_i y_i \alpha_i = 0 \end{cases} \qquad \begin{aligned} &\alpha_i\left(y_i\left(\langle w, x_i\rangle + b\right) - 1\right) = 0 \\ &y_i\left(\langle w, x_i\rangle + b\right) \geq 1 \\ &\alpha_i \geq 0 \end{aligned}$$

*非线性规划领域里最重要的理论成果之一，是确定某点为极值点的必要条件。如果所讨论的规划是凸规划，那么库恩-塔克条件也是充分条件。

形式化对偶问题

$$L(w,b,\alpha)=\frac{1}{2}\|w\|_2^2-\sum_i\alpha_i\left(y_i\left(\langle w,x_i\rangle+b\right)-1\right)$$

$$\begin{cases}\dfrac{\partial L}{\partial w}=w-\sum_i\alpha_i y_i x_i=0\\ \dfrac{\partial L}{\partial b}=\sum_i y_i\alpha_i=0\end{cases}$$

把这个结果带入 $L(w,b,\alpha)$

可以(自己尝试)得到对偶问题

$$\max_{\{\alpha_i\}}-\frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j\langle x_i,x_j\rangle+\sum_i\alpha_i$$

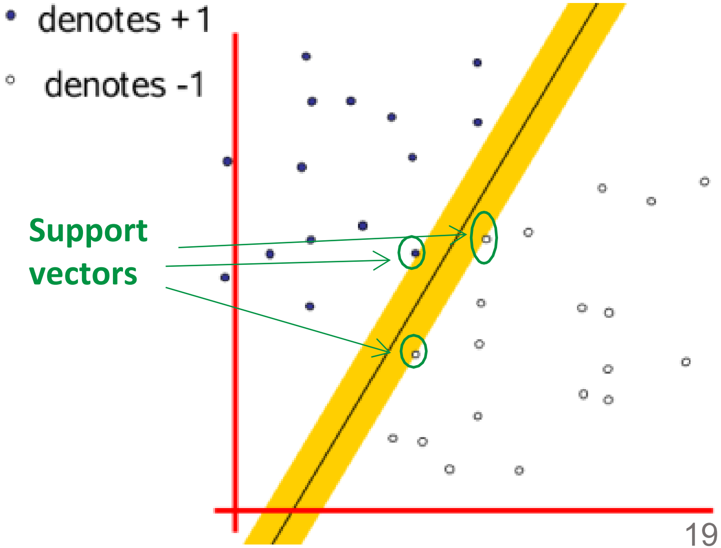
$$\begin{aligned}\min_{\{\alpha_i\}}\quad & \tfrac{1}{2}\textstyle\sum_{i,j}\alpha_i\alpha_j y_i y_j\langle x_i,x_j\rangle-\sum_i\alpha_i\\ \text{s.t.}\quad & \textstyle\sum_i y_i\alpha_i=0,\\ & \alpha_i\geq 0,\quad\forall i=1,\ldots,N\end{aligned}$$

支持向量

- 根据KKT条件 $\alpha_i\left(y_i\left(\langle w, x_i\rangle + b\right) - 1\right) = 0$
- α_i 非零 当且仅当 $y_i(\langle w, x_i\rangle + b) = 1$，即 x_i 在间隔的边界上

→ 这些 x_i 是 **支持向量 (SV)**

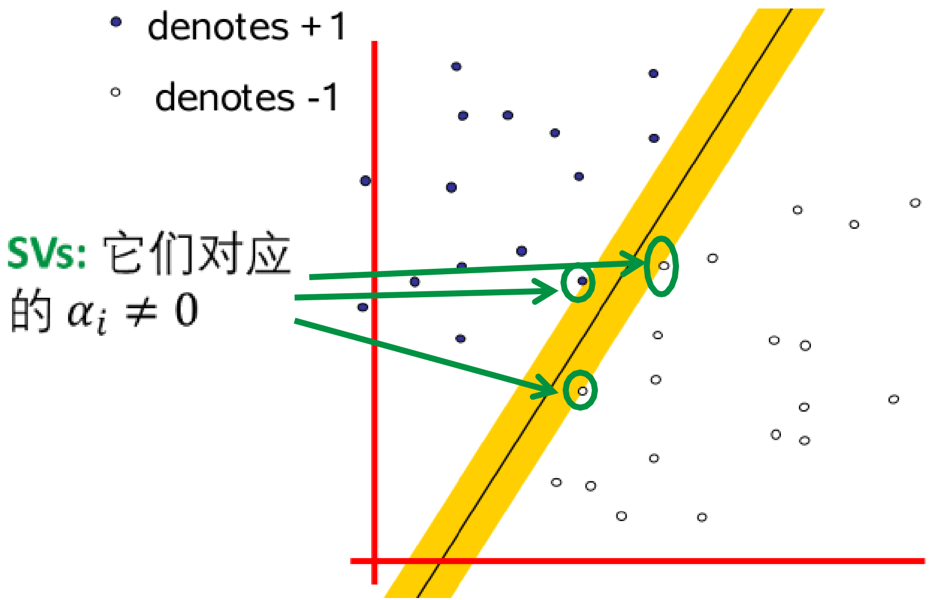
- 大多数 α_i 为零
  - 这些 (x_i, y_i) 对 $f(x)$ 没有影响
  - 稀疏解



如[Vapnik 1999]所述，支持向量机这个名字强调了此类学习器的关键是如何从支持向量构建出解；同时也暗示着起复杂度主要与支持向量机的数目有关。

目前为止

- SVM 在线性可分时
  - 最大化间隔



- 原始问题:

$$\min_{w,b} \quad \frac{1}{2}\|w\|_2^2$$
$$\text{s.t.} \quad y_i(\langle w, x_i\rangle + b) \geq 1, \quad \forall i = 1, \ldots, N$$

- 对偶问题:

$$\min_{\{\alpha_i\}} \quad \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle - \sum_i \alpha_i$$
$$\text{s.t.} \quad \sum_i y_i\alpha_i = 0,$$
$$\alpha_i \geq 0, \quad \forall i = 1, \ldots, N$$

纲要

- 背景
- 线性支持向量机
  - 最大化间隔分类器
  - 形式化对偶问题
  - 线性不可分情况 （待续）
- 核函数支持向量机
- 核心概要