

# 模式识别

支撑向量机SVM

吴建鑫

南京大学计算机系，2017

# 统计学习方法的粗略分类

## ✓ Statistical learning methods

- $p(y = i)$ ,  $p(y = i|\mathbf{x})$ ,  $p(\mathbf{x}|y = i)$ ,  $p(\mathbf{x})$
- 还记得其含义吗?
- Generative (probabilistic) models: 估计 $p(\mathbf{x}|y = i)$ 和 $p(y)$ 
  - 然后用贝叶斯定理求 $p(y = i|\mathbf{x})$
  - 生成模型 (下一章)
- Discriminative (probabilistic) models: 直接估计 $p(y = i|\mathbf{x})$ 
  - 判别模型 (下一章)
- Discriminant function: 直接求一个把各类分开的边界
  - 不假设概率模型, 如FLD (上一章), SVM (本章)
  - 更多阅读PRML1.5.4

# 目标

- ✓ 理解并掌握SVM中主要思想的含义、描述、数学表述
- ✓ 如何将一个好的idea形式化
- ✓ 能实际应用SVM
- ✓ 提高目标
  - 理解相关推导，能在有文献帮助下自主完成推导
  - 进一步能通过独立阅读、了解统计学习

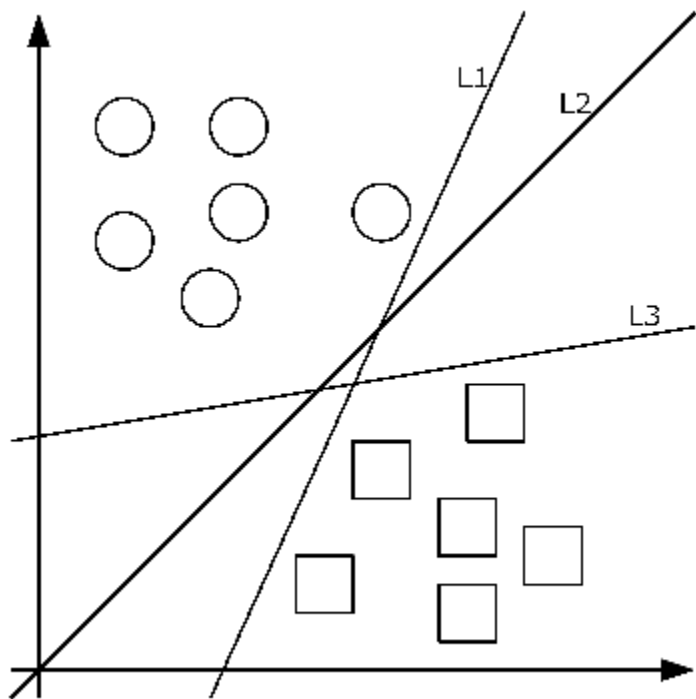
# SVM

---

Support vector machine 支撑向量机

注意SVM的**形式化**过程，和**简化**的思路

# large margin (最大边际?)



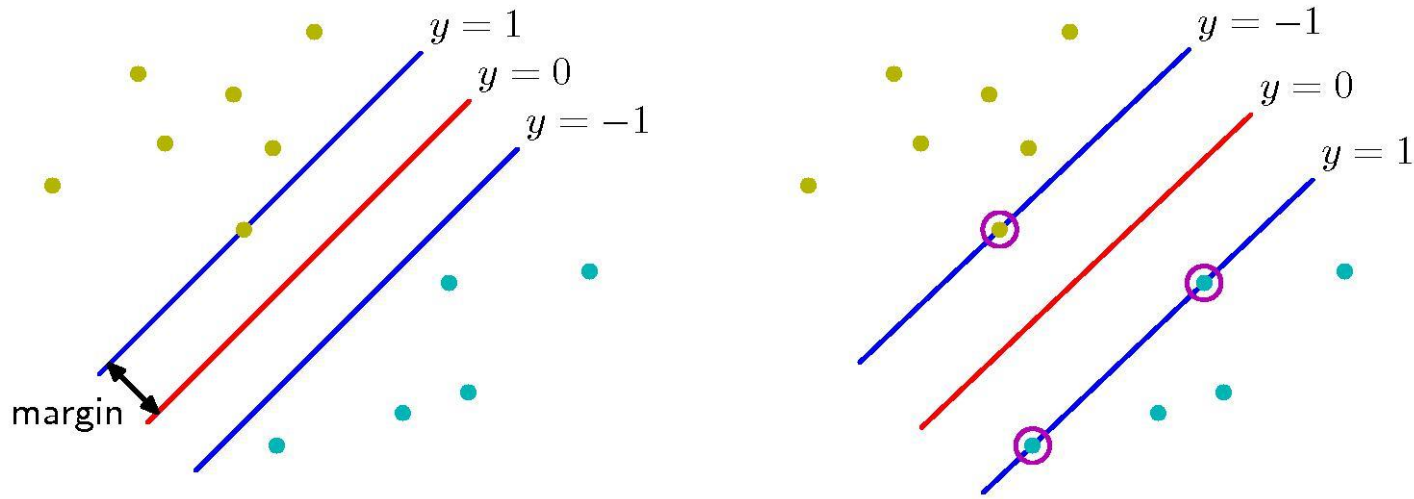
用线性边界分开2类

- 正类positive class,  $y_i = 1$
- 负类negative class,  $y_i = -1$
- 可以有很多边界L1, L2, L3, ..., 在训练集上都100%正确 (假设能完全分开)
- 哪个最好?

<http://zh.wikipedia.org/wiki/File:Classifier.svg>

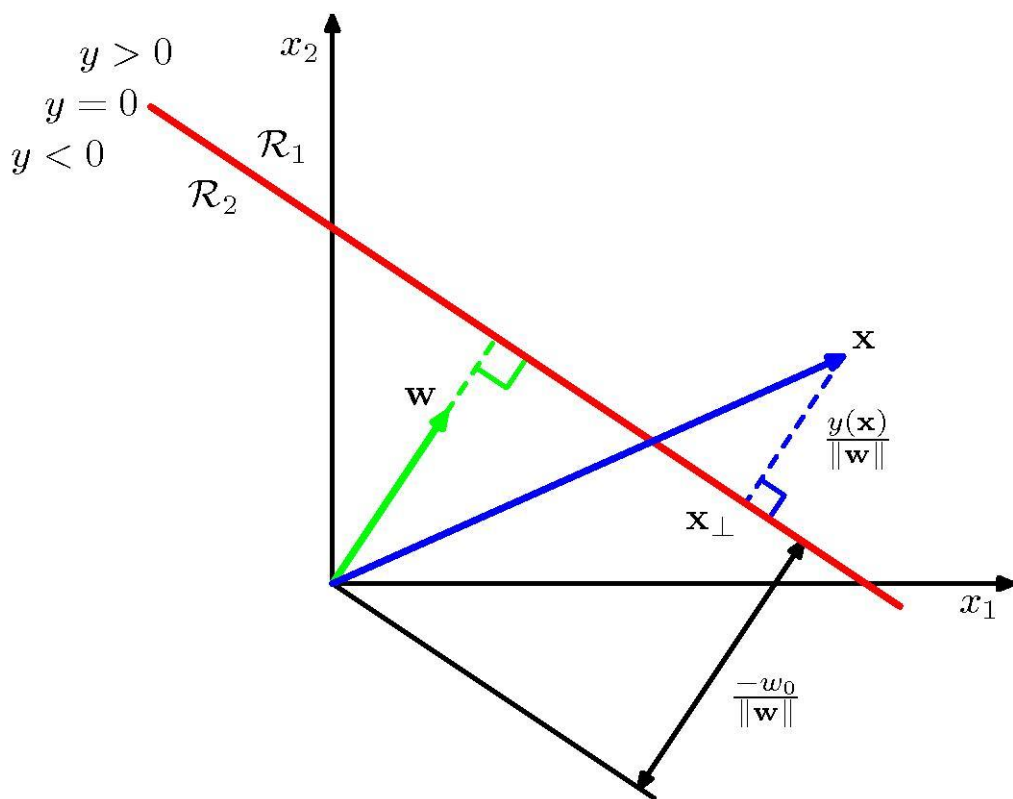
# margin

图片来自PRML第7章



- 一个点（样例）的边距margin是其到分界超平面 separating hyperplane 的垂直距离
- SVM最大化（所有训练样本的）最小边距
- 有最小边距的点称为支撑向量 (support vectors)
  - 所以叫支撑向量机 support vector machine

# 几何geometry示意图



- 分类超平面  
 $f(\mathbf{x}) = \mathbf{w}^T \mathbf{x} + b = 0$ 
  - 红色
  - 绿色为其法向量  
normal vector
- $\mathbf{x}$ 为任一点/样例
  - 其到超平面的距离为?

图片来自PRML第4章

# 计算margin

- ✓ 投影点为 $\mathbf{x}_\perp$ ,  $\mathbf{x} - \mathbf{x}_\perp$ 为距离向量
  - 其方向与 $\mathbf{w}$ 相同, 为 $\mathbf{w}/\|\mathbf{w}\|$
  - 其大小 $r$ 可为0, 或正, 或负; margin为其大小的绝对值
- ✓  $\mathbf{x} = \mathbf{x}_\perp + r \frac{\mathbf{w}}{\|\mathbf{w}\|}$ , 两边同乘以 $\mathbf{w}^T$ , 然后加上 $b$ 
  - $\mathbf{w}^T \mathbf{x} + b = \mathbf{w}^T \mathbf{x}_\perp + b + r \frac{\mathbf{w}^T \mathbf{w}}{\|\mathbf{w}\|}$
  - $f(\mathbf{x}) = f(\mathbf{x}_\perp) + r \|\mathbf{w}\|$  为什么?
  - $r = \frac{f(\mathbf{x})}{\|\mathbf{w}\|}$  为什么?
  - $\mathbf{x}$ 的margin是 $\frac{|f(\mathbf{x})|}{\|\mathbf{w}\|} = \frac{|\mathbf{w}^T \mathbf{x} + b|}{\|\mathbf{w}\|}$



# 分类、评价

## ✓ 怎么样分类？

- $f(\mathbf{x}) > 0$  - - 分为正类,  $f(\mathbf{x}) < 0$  - - 分为负类
- 那么  $f(\mathbf{x}) = 0$  怎么办？

## ✓ 对于任何一个样例，怎么知道预测的对错？

- $y_i f(\mathbf{x}_i) > 0$  正确       $y_i f(\mathbf{x}_i) < 0$  错误
- 即，**因为我们假设能完全分开**，所以

$$y_i f(\mathbf{x}_i) = |f(\mathbf{x}_i)|$$

# SVM的形式化描述

✓ 那么，SVM问题是什么？

$$\operatorname{argmax}_{\mathbf{w}, b} \left( \min_i \left( \frac{|\mathbf{w}^T \mathbf{x}_i + b|}{\|\mathbf{w}\|} \right) \right)$$

$$\operatorname{argmax}_{\mathbf{w}, b} \left( \min_i \left( \frac{y_i (\mathbf{w}^T \mathbf{x}_i + b)}{\|\mathbf{w}\|} \right) \right)$$

$$\operatorname{argmax}_{\mathbf{w}, b} \left( \frac{1}{\|\mathbf{w}\|} \min_i (y_i (\mathbf{w}^T \mathbf{x}_i + b)) \right)$$

✓ 非常难以优化，怎么办？

- 继续简化

# 换个角度看问题

✓ 到目前为止

- 对 $\mathbf{w}$ 没有限制，要求最大化最小的边际，难优化

✓ 判断对错：如果 $yf(\mathbf{x}) > 0$  即正确

- 即 $y(\mathbf{w}^T \mathbf{x} + b) > 0$ ，只需要方向，完全不需要大小！
- 如果 $(\mathbf{w}, b)$ 变为 $(\lambda \mathbf{w}, \lambda b)$ ，预测和边际会变吗？

✓ 那么我们可以限定 $\min_i (y_i(\mathbf{w}^T \mathbf{x}_i + b))$ 为1

- 问题变为：在限制 $\min_i (y_i(\mathbf{w}^T \mathbf{x}_i + b))$ 为1时，最大化 $\frac{1}{\|\mathbf{w}\|}$

$$\begin{aligned} & \underset{\mathbf{w}, b}{\operatorname{argmin}} && \frac{1}{2} \mathbf{w}^T \mathbf{w} \\ \text{s. t.} &&& y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1, \forall i \end{aligned}$$

# 拉格朗日乘子法, again

✓  $L(\mathbf{w}, b, \mathbf{a}) = \frac{1}{2} \mathbf{w}^T \mathbf{w} - \sum_{i=1}^n a_i (y_i (\mathbf{w}^T \mathbf{x}_i + b) - 1)$

• Subject to  $a_i \geq 0$

✓ 作业：证明最优化的必要条件

✓  $\frac{\partial L}{\partial \mathbf{w}} = \mathbf{0} \quad \rightarrow \quad \mathbf{w} = \sum_{i=1}^n a_i y_i \mathbf{x}_i$

✓  $\frac{\partial L}{\partial b} = 0 \quad \rightarrow \quad 0 = \sum_{i=1}^n a_i y_i$

✓ 在此两条件下，将两个等式代入回  $L$

$$\tilde{L}(\mathbf{a}) = \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n a_i a_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j$$

# SVM的对偶形式

- ✓ 在原来的空间（输入空间，input space）中
  - 变量是 $\mathbf{x}_i$ ，称为SVM的primal form
- ✓ 现在的问题里面
  - 变量是 $a_i$ ，即拉格朗日乘子，称为对偶空间dual space
  - 对偶空间完成优化后，得到最优的 $\mathbf{a}$ ，可以得到原始空间中的最优解 $\mathbf{w}$
- ✓ SVM的对偶形式dual form

$$\begin{aligned} \arg\max_{\mathbf{a}} \quad & \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n a_i a_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ \text{s.t.} \quad & a_i \geq 0 \\ & \sum_{i=1}^n a_i y_i = 0 \end{aligned}$$

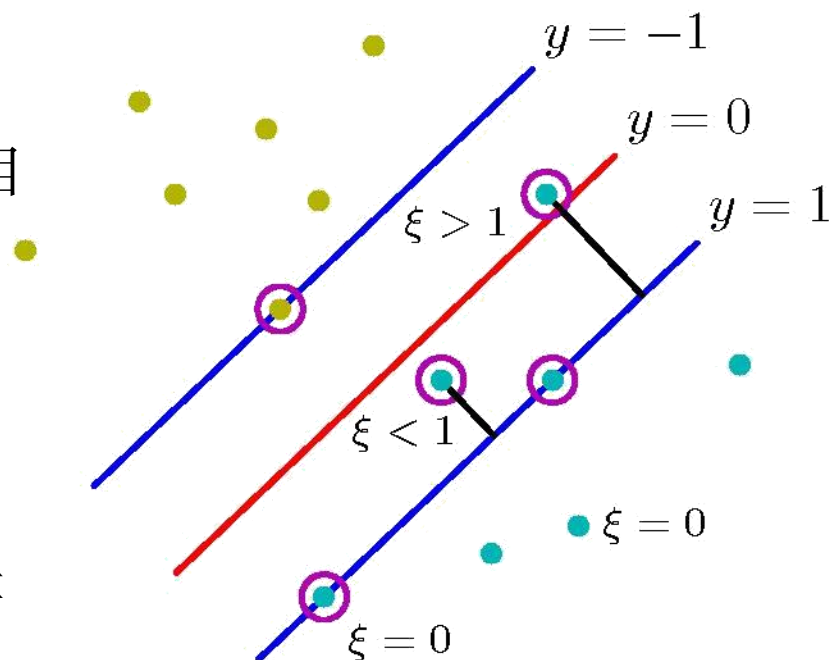
# 剩下的问题

- ✓ 如何最优化？
  - 对偶空间中
  - 原始空间中
- ✓ 如果能允许少数点  $y_i f(\mathbf{x}_i) < 1$ 
  - 如果允许一个点  $y_i f(\mathbf{x}_i) < 1$ ，但是大幅度增加margin呢？
- ✓ 如果不是线性可分的linearly separable，但是可以用非线性的边界分开non-linearly separable？
- ✓ 如果不是两个类，而是多个呢？

# Soft margin

- ✓ 可以允许少数点margin比1小
  - 但是犯错误是有惩罚的，否则？
  - $y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 \rightarrow y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i$
  - $\xi_i$ : 松弛变量slack variable, 即允许犯的错误
  - $\xi_i \geq 0$

•  $\xi_i = 0$ ,  $(0, 1)$ ,  $\xi_i = 1$ ,  $\xi_i > 1$ 各自代表什么？



图片来自PRML第7章

# 如何惩罚？

✓ Primal space

$$\begin{aligned} \underset{\mathbf{w}, b, \xi}{\operatorname{argmin}} \quad & \frac{1}{2} \mathbf{w}^T \mathbf{w} + C \sum_{i=1}^n \xi_i \\ \text{s. t.} \quad & y_i(\mathbf{w}^T \mathbf{x}_i + b) \geq 1 - \xi_i \\ & \xi_i \geq 0 \end{aligned}$$

✓  $C > 0$ : 正则化参数 regularization parameter

- $\xi_i$ —代价，我们要最小化代价函数（总代价）
- $\frac{1}{2} \mathbf{w}^T \mathbf{w}$ —正则项 regularization term，对分类器进行限制，使复杂度不至于太高（另一个角度，还是最大化边际）
- 那么，怎么确定  $C$  的值？



# Soft margin的对偶形式

✓ 自主阅读PRML

$$\begin{aligned} \operatorname{argmax}_a \quad & \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n a_i a_j y_i y_j \mathbf{x}_i^T \mathbf{x}_j \\ \text{s.t.} \quad & C \geq a_i \geq 0 \\ & \sum_{i=1}^n a_i y_i = 0 \end{aligned}$$

✓ 对偶形式仅依赖于内积！

# 内积：线性和非线性的联系

✓ 线性和非线性有时候紧密联系在一起——通过内积

✓  $\mathbf{x} = (x_1, x_2), \mathbf{z} = (z_1, z_2)$

✓  $K(\mathbf{x}, \mathbf{z}) = (1 + \mathbf{x}^T \mathbf{z})^2 = (1 + x_1 z_1 + x_2 z_2)^2$   
 $= 1 + 2x_1 z_1 + 2x_2 z_2 + x_1^2 z_1^2 + x_2^2 z_2^2 + 2x_1 z_1 x_2 z_2$

$$= \begin{pmatrix} 1 \\ \sqrt{2}x_1 \\ \sqrt{2}x_2 \\ x_1^2 \\ x_2^2 \\ \sqrt{2}x_1x_2 \end{pmatrix}^T \begin{pmatrix} 1 \\ \sqrt{2}z_1 \\ \sqrt{2}z_2 \\ z_1^2 \\ z_2^2 \\ \sqrt{2}z_1z_2 \end{pmatrix}$$

# Kernel trick

- ✓ 两个向量  $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ , 一个非线性函数  $K(\mathbf{x}, \mathbf{y})$
- ✓ 对于满足某些条件的函数  $K$ , 一定存在一个映射 (mapping)  $\phi: \mathbb{R}^d \mapsto \Phi$ , 使得对任意的  $\mathbf{x}, \mathbf{y}$ 
$$K(\mathbf{x}, \mathbf{y}) = \phi(\mathbf{x})^T \phi(\mathbf{y})$$
  - 非线性函数  $K$  表示两个向量的相似程度
  - 其等价于  $\Phi$  里面的内积
- ✓  $\Phi$ : 特征空间 feature space
  - 可以是有限维的空间, 但也可以是无穷维的空间  
infinite dimensional Hilbert space

# 什么样的限制条件？

- ✓ 必须存在特征映射feature mapping, 才可以将非线性函数表示为特征空间中的内积
- ✓ Mercer's condition (Mercer条件, 是充分必要的): 对任何满足  $\int g^2(\mathbf{u})d\mathbf{u} < \infty$  的非零函数, 对称函数  $K$  满足条件:  $\iint g(\mathbf{u})K(\mathbf{u}, \mathbf{v})g(\mathbf{v})d\mathbf{u}d\mathbf{v} \geq 0$
- ✓ 看上去眼熟? 另一种等价形式: 对任何一个样本集合  $\{\mathbf{x}_1, \dots, \mathbf{x}_n\}, \mathbf{x}_i \in \mathbb{R}^d$ , 如果矩阵  $K = [K_{ij}]_{i,j}$  (矩阵的第  $i$  行、第  $j$  列元素  $K_{ij} = K(\mathbf{x}_i, \mathbf{x}_j)$ ) 总是半正定的, 那么函数  $K$  满足 Mercer 条件
- ✓ 如何判定是否满足? 有几种方法?

# 核支撑向量机Kernel SVM

✓ 核函数kernel function:  $K$

✓ 对偶形式:

$$\operatorname{argmax}_a \sum_{i=1}^n a_i - \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n a_i a_j y_i y_j K(\mathbf{x}_i, \mathbf{x}_j)$$

✓ 分类边界:  $\mathbf{w} = \sum_{i=1}^n a_i y_i \phi(\mathbf{x}_i)$

✓ 怎样预测:  $\mathbf{w}^T \phi(\mathbf{x}) = \phi(\mathbf{x})^T \left( \sum_{i=1}^n a_i y_i \phi(\mathbf{x}_i) \right) = \sum_{i=1}^n a_i y_i K(\mathbf{x}, \mathbf{x}_i)$

- 线性:  $\mathbf{w} = \sum_{i=1}^n a_i y_i \mathbf{x}_i$ ,  $\mathbf{w}^T \mathbf{x}$  计算量为  $O(d)$
- 非线性 (核) 方法测试所需时间为?
- 假设计算  $K$  的时间为  $O(d)$ , 是  $O(nd)$  吗?

# Complementary Slackness

- ✓ 对所有 $i$ , KKT条件包括 $(C - a_i)\xi_i = 0$ 
  - 情况1:  $C > a_i > 0$ ,  $\xi_i = 0$ , 在特征空间中边际为1的两个超平面上
  - 情况2:  $a_i = C$ , 对 $\xi_i$ 没有限制
    - 可以在超平面上、介于两个超平面之间、或以外（即分类错误）
  - 情况3:  $a_i = 0$ , 在预测时不需要计算
- ✓ 这代表什么？
  - 复杂度由 $a_i > 0$ 的个数, 而非样本的总数目来决定
  - $\mathbf{a}$ 是稀疏的
  - 在soft margin SVM中, 称 $a_i > 0$ 对应的 $\mathbf{x}_i$ 为支撑向量

# 非线性核

- ✓ 线性核 linear kernel, dot-product kernel:

$$K(\mathbf{x}, \mathbf{y}) = \mathbf{x}^T \mathbf{y}$$

- ✓ 非线性核 non-linear kernel

- RBF (radial basis function)、高斯 (Gaussian) 核:

$$K(\mathbf{x}, \mathbf{y}) = \exp(-\gamma \|\mathbf{x} - \mathbf{y}\|^2)$$

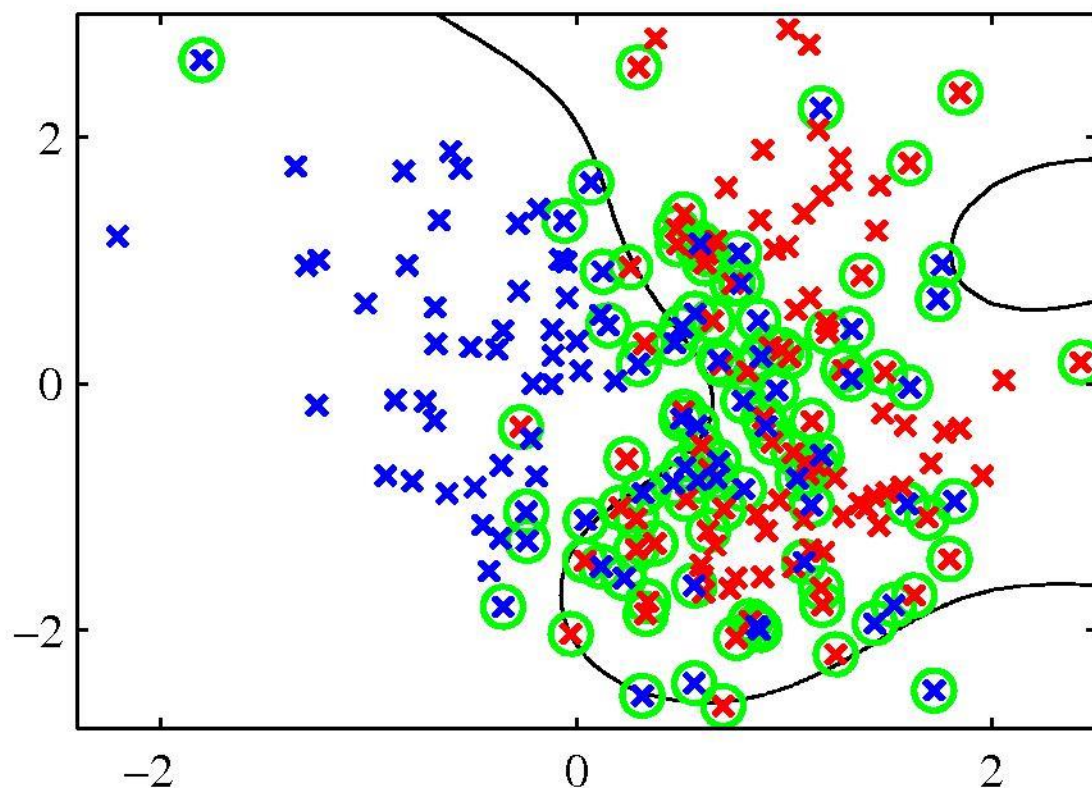
- 多项式核:  $K(\mathbf{x}, \mathbf{y}) = (\gamma \mathbf{x}^T \mathbf{y} + c)^d$

- ...

- 进一步阅读: 更多核函数

<http://www.zhizhihu.com/html/y2010/2292.html>

# 非线性核的例子 (RBF)



图片来自PRML第7章



# 超参数

## ✓ 如何决定 $C$ 、 $\gamma$ 、...

- 必须给定这些参数parameter的值，才能进行SVM学习，SVM本身不能学习这些参数！
- 称为超参数hyper-parameter
- 对SVM的结果有极大的影响！

## ✓ 用交叉验证在训练集上来学习

- 在训练集上得到不同参数的交叉验证准确率
- 选择准确率最高的超参数的数值

# 多类Multiclass (1)

- ✓ 思路：转化为2类问题
- ✓ 1-vs-1 (one versus one):  $C$  个类  $\{1, 2, \dots, C\}$ 
  - 设计  $\binom{C}{2}$  个分类器：用  $i$  和  $j$  ( $i > j$ ) 两类的训练数据学习
  - 一共  $C(C - 1)/2$  个，其中每个类出现  $C$  次
  - 对测试样本  $\mathbf{x}$ ，一共会得到  $C(C - 1)/2$  个结果，然后投票vote
  - 每个分类器  $f_i$  采用其二值输出，即  $\text{sign}(f_i(\mathbf{x}))$

	1	2	3
1			
2	1		
3	1	3	

# 多类Multiclass (2)

✓ 1-vs.-all (或1-vs.-rest)

- 设计 $C$ 个分类器，第 $i$ 个分类器用类 $i$ 做正类，把所有 $C - 1$ 个类别的数据合并在一起做负类
  - 和交叉验证的步骤有些类似
  - 每个新的分类器 $f_i$ 采用其实数值输出，即 $f_i(\mathbf{x})$
- $f_i(\mathbf{x})$ 的实数输出可以看成是其“信心” confidence
- 最终选择信心最高的那个类为输出

$$\operatorname{argmax}_i f_i(\mathbf{x})$$

# 多类Multiclass (3)

## ✓ 直接解决多类问题(进一步阅读)

- Crammer-Singer方法
- <http://jmlr.org/papers/v2/crammer01a.html>

## ✓ DAGSVM(进一步阅读)

- <http://research.microsoft.com/apps/pubs/?id=68541>

## ✓ ECOC(进一步阅读)

- <http://www.jair.org/papers/paper105.html>

# 从SVM的介绍学到的思想？

1. 确定问题，对问题有充分的认识（实践、理论）
2. 好的思路、想法idea（如margin）
  - 从理论（概率、统计？）中来
  - 或者实践（已有线性分类器的缺点，如感知机 perceptron）
3. 形式化
  - 用精确的数学形式表达出来
  - 如果不能精确描述，或说明你的idea有问题
  - 简化，开始时避免复杂、模糊的想法：限制条件（如，线性可分），从较小范围开始（如，2类）
4. 数学基础和研究
  - 用到的几何、凸优化、拉格朗日乘子法、Hilbert空间…
  - 经典的相关数学背景要熟悉：至少知道到哪里查

# 简化：一种可靠的思路

- ✓ 问题（特别是数学问题）难以解决时，尽量简化
  - 问题的表述，如果难以形式化，可以将问题简化
  - 简化后的问题可以去除很多复杂的考虑，但是
  - 原问题的核心要保持
  - 如SVM从二类、线性、可分的情况开始
- ✓ 有时可以通过换思路的方法等价简化
  - 如SVM限定 $\min_i (y_i(\mathbf{w}^T \mathbf{x}_i + b))$ 为1
  - 也可以对原问题做不重要的修改以使简化成为可能
  - 如（**进一步**阅读）LIBLINEAR假设不使用 $b$

# 进一步的阅读

- ✓ 如果对本章的内容感兴趣，可以参考如下文献
  - 凸函数、拉格朗日乘子法、KKT条件：
    - Convex Optimization第一、二、五章
  - SVM和统计学习
    - <http://research.microsoft.com/pubs/67119/svmtutorial.pdf>
    - 最新会议论文集：ICML、NIPS、AISTATS、COLT、...
  - SMO：
    - [http://en.wikipedia.org/wiki/Sequential\\_minimal\\_optimization](http://en.wikipedia.org/wiki/Sequential_minimal_optimization)
    - LIBSVM, SVMlight
  - Pegasos: <http://www.cs.huji.ac.il/~shais/code/>
  - DCD/LIBLINEAR:  
<http://www.csie.ntu.edu.tw/~cjlin/liblinear/>
  - 加性核：我的主页→publications页面→[W5]