《模式识别》

第八章 支持向量机

马锦华

https://cse.sysu.edu.cn/teacher/MaJinhua

SUN YAT-SEN University

中山大學 计算机学院
SCHOOL OF COMPUTER SCIENCE AND ENGINEERING, SYSU

课程目录（暂定）

统计学习方法的粗略分类

- Statistical learning methods
  - $p(y=i), p(y=i|\boldsymbol{x}), p(\boldsymbol{x}|y=i), p(\boldsymbol{x})$
  - 还记得其含义吗？
  - Generative models: 估计$p(\boldsymbol{x}|y=i)$和$p(y)$
    - 如GMM、HMM、GAN、VAE、Diffusion等
  - Discriminative models: 估计$p(y=i|\boldsymbol{x})$
    - 如k-NN、SVM、决策树、逻辑回归等
  - Discriminant function：一个把各类分开的函数
    - 求法：1. 利用生成模型和贝叶斯定理；2. 直接估计后验概率
    - 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

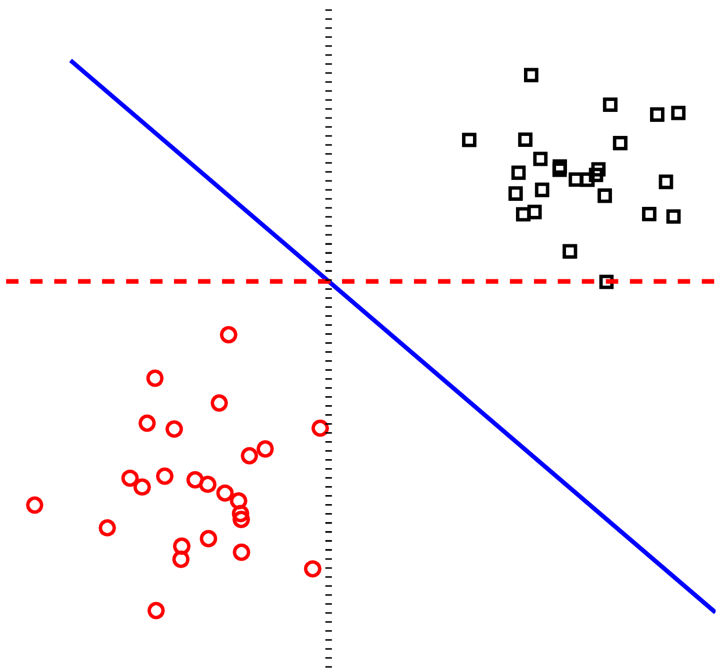
SVM

Support vector machine支持向量机

原始SVM算法是由弗拉基米尔·瓦普尼克（Vladimir Vapnik）等提出

注意SVM的形式化过程，和简化的思路

Large margin（最大间隔？）

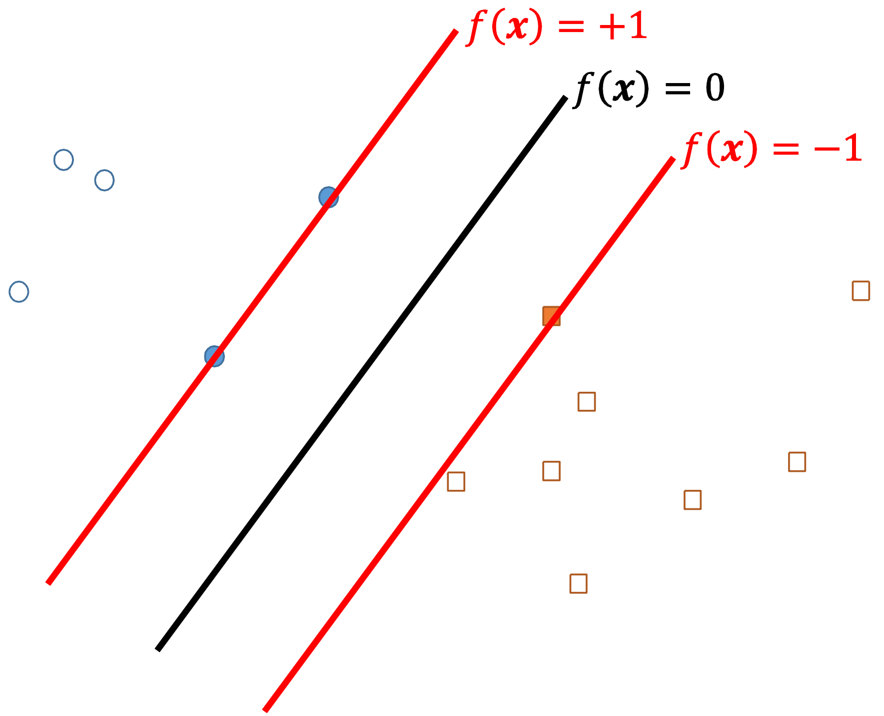


用线性边界分开2类

- 正类positive class, $y_i = 1$
- 负类negative class, $y_i = -1$

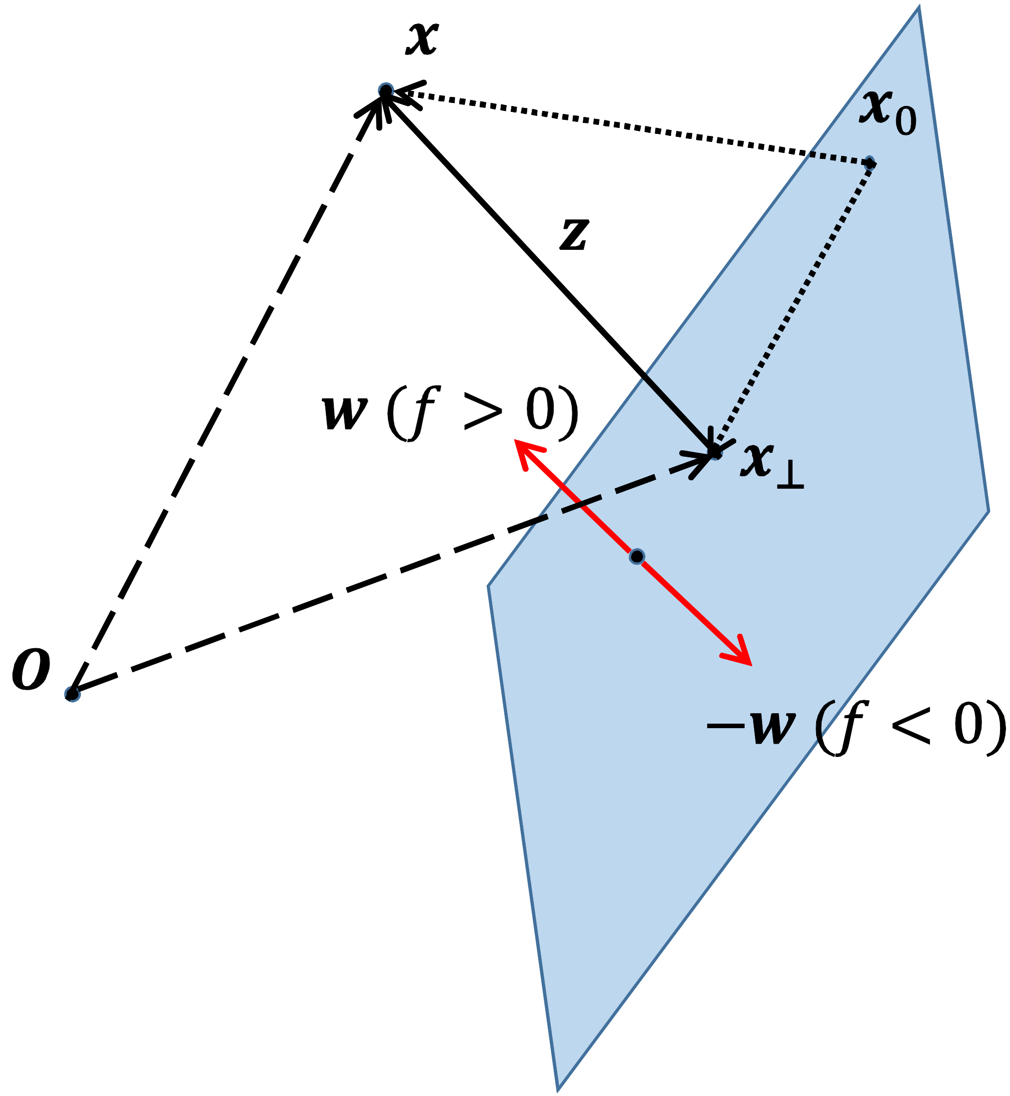
- 可以有很多决策边界，在训练集上都100%正确（假设能完美线性分开该数据集）
- 哪个边界最好？

margin



- 一个点（样本）的间隔margin是其到决策超平面separating hyperplane （$f(\boldsymbol{x}) = 0$）的垂直距离
- SVM最大化（所有训练样本的）最小间隔
- 有最小间隔的点称为支持向量(support vectors)
  - 所以叫支持向量机support vector machine

几何geometry示意图



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

计算margin

- 投影点为$\boldsymbol{x}_{\perp}$， $\boldsymbol{x}-\boldsymbol{x}_{\perp}$为垂直于决策超平面向量
  - 其方向与$\boldsymbol{w}$相同，为$\boldsymbol{w}/\|\boldsymbol{w}\|$
  - 其在$\boldsymbol{w}$上的投影系数r可为0，或正，或负; margin为其投影系数r的绝对值
- $\boldsymbol{x}=\boldsymbol{x}_{\perp}+r\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|}$，两边同乘以$\boldsymbol{w}^T$，然后加上$b$
  - $\boldsymbol{w}^T\boldsymbol{x}+b=\boldsymbol{w}^T\boldsymbol{x}_{\perp}+b+r\frac{\boldsymbol{w}^T\boldsymbol{w}}{\|\boldsymbol{w}\|}$
  - $f(\boldsymbol{x})=f(\boldsymbol{x}_{\perp})+r\|\boldsymbol{w}\|$ 为什么？
  - $r=\frac{f(\boldsymbol{x})}{\|\boldsymbol{w}\|}$ 为什么?
  - $\boldsymbol{x}$的margin是$\frac{|f(\boldsymbol{x})|}{\|\boldsymbol{w}\|}=\frac{|\boldsymbol{w}^T\boldsymbol{x}+b|}{\|\boldsymbol{w}\|}$

分类、评价

- 怎么样分类？
  - $f(\boldsymbol{x}) > 0$— 分为正类，$f(\boldsymbol{x}) < 0$ — 分为负类

- 对于任何一个样例，怎么知道预测的对错？
  - $y_i \in \{-1, +1\}$
  - $y_i f(\boldsymbol{x}_i)$>0 正确 $y_i f(\boldsymbol{x}_i) < 0$ 错误
  - 即，因为我们假设能完全分开，所以

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

SVM的形式化描述

- 那么，SVM问题是什么？

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

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

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

- 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$\boldsymbol{w}$没有限制，要求最大化最小的间距，难优化
- 判断对错: 如果$yf(\boldsymbol{x})$>0 即正确
  - 即$y(\boldsymbol{w}^T\boldsymbol{x}+b)$>0，只需要方向，完全不需要大小！
  - 如果$(\boldsymbol{w},b)$变为$(\lambda\boldsymbol{w},\lambda b)$，预测和间距会变吗？
- 那么我们可以限定$\min_i\left(y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)\right)$为1
  - 问题变为：在限制$\min_i\left(y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)\right)$为1时，最大化$\frac{1}{\|\boldsymbol{w}\|}$

$$\underset{\boldsymbol{w},b}{\operatorname{argmin}} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$$

$$s.t.\ \ y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)\geq 1, \forall i$$

拉格朗日乘子法again

- $L(\boldsymbol{w}, b, \boldsymbol{\alpha}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} - \sum_{i=1}^{n} \alpha_i(y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) - 1)$
  - Subject to $\alpha_i \geq 0$
- 证明最优化的必要条件
- $\frac{\partial L}{\partial \boldsymbol{w}} = \boldsymbol{0}$ ➔ $\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i y_i \boldsymbol{x}_i$
- $\frac{\partial L}{\partial b} = 0$ ➔ $0 = \sum_{i=1}^{n} \alpha_i y_i$
- 在此两条件下，将两个等式代入回L

$$\tilde{L}(\boldsymbol{\alpha}) = \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$

SVM的对偶形式

- 在原来的空间（输入空间，input space）中
  - 变量是$\boldsymbol{x}_i$，称为SVM的primal form
- 现在的问题里面
  - 变量是α_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$\boldsymbol{\alpha}$，可以得到原始空间中的最优解$\boldsymbol{w}$和b
- SVM的对偶形式dual form

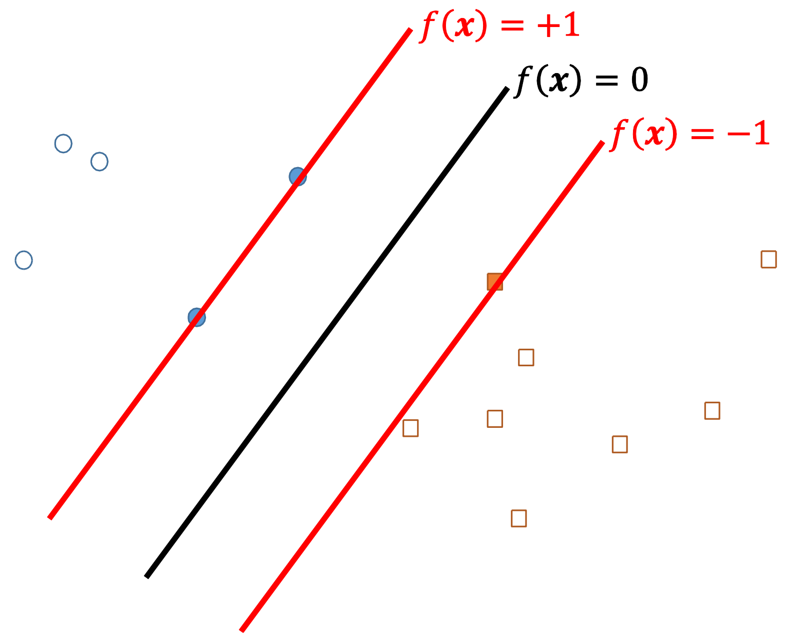
$$\underset{\boldsymbol{\alpha}}{\arg\max} \sum_{i=1}^{n} \alpha_i - \frac{1}{2} \sum_{i=1}^{n} \sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j \boldsymbol{x}_i^T \boldsymbol{x}_j$$

$$s.t. \qquad \alpha_i \geq 0$$

$$\sum_{i=1}^{n} \alpha_i y_i = 0$$

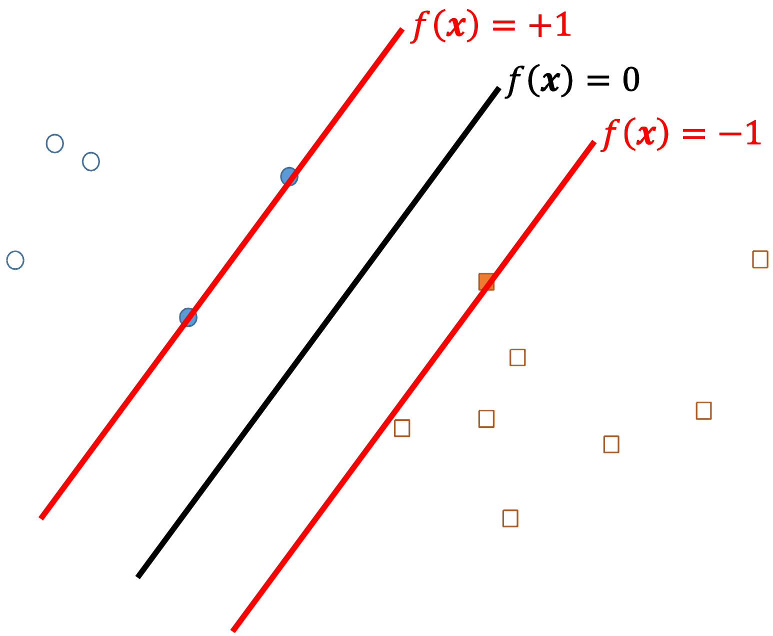
Complementary Slackness

- 对所有i，KKT条件包括$\alpha_i(y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)-1)=0$
  - 情况1：$\alpha_i>0, y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)=1$，即$\boldsymbol{x}_i$在到决策超平面（$y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)=0$）间隔为1的两个超平面上
  - 情况2：$\alpha_i=0$，$y_i(\boldsymbol{w}^T\boldsymbol{x}_i+b)>1$，即$\boldsymbol{x}_i$不在两个超平面



Complementary Slackness

- 这代表什么？
  - ❑ $\boldsymbol{w}$和b的计算只需要利用$\alpha_i > 0$的样本
  - ❑ $\boldsymbol{\alpha}$是稀疏的
  - ❑ 在SVM中，称$\alpha_i > 0$对应的$\boldsymbol{x}_i$为支持向量

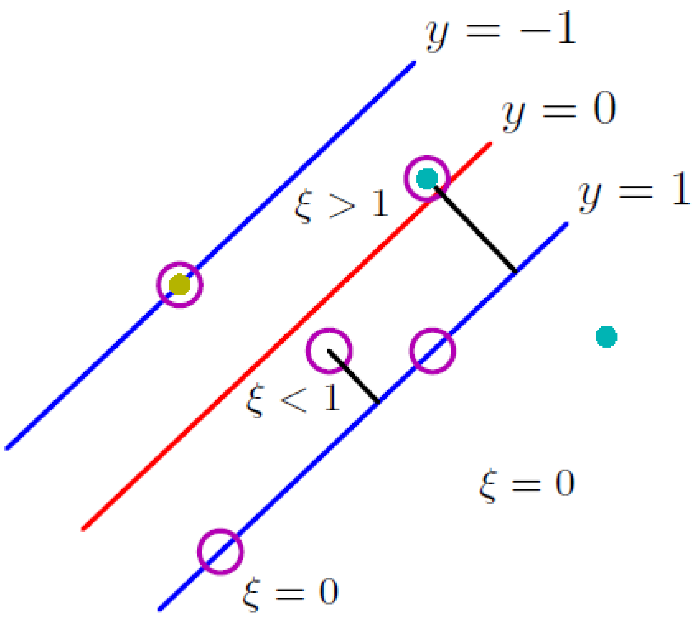


剩下的问题

- 如何最优化？
  - ❑ 对偶空间中
  - ❑ 原始空间中
- ✓ 如果能允许少数点$y_i f(\boldsymbol{x}_i) < 1$
  - ❑ 但是大幅度增加间隔margin（即减小$||\boldsymbol{w}||$）呢？
- 如果不是线性可分的（linearly separable）
  - ❑ 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？
  - $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1$ ➔ $y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i$
  - ξ_i：松弛变量slack variable，即允许犯的错误
  - $\xi_i \geq 0$
  - ξ_i=0, (0 1), =1, >1各自代表什么?



如何惩罚？

- Primal space

$$\underset{\boldsymbol{w},b,\xi}{\operatorname{argmin}} \quad \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{n}\xi_i$$

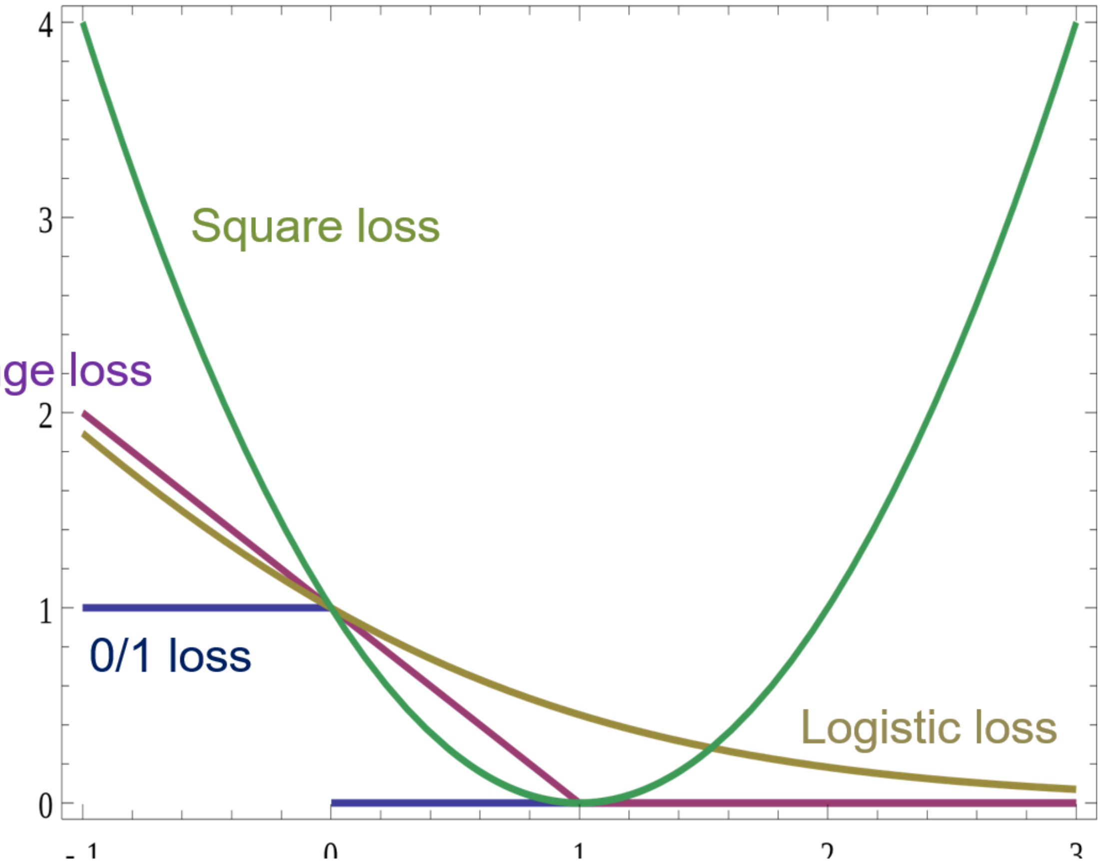
$$s.t. \quad y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b) \geq 1 - \xi_i$$

$$\xi_i \geq 0$$

- $C > 0$: 超参数hyperparameter
  - ξ_i—代价，我们要最小化代价函数（总代价）
  - $\frac{1}{2}\boldsymbol{w}^T\boldsymbol{w}$—正则化项（regularization term），对分类器进行限制，使复杂度不至于太高（另一个角度，还是最大化间隔）
  - 那么，怎么确定C的值？

经验损失函数

Loss Functions



PRML

Soft margin SVM的对偶问题

- 自学PRML
  - 拉格朗日函数：$L(\boldsymbol{w}, b, \boldsymbol{\xi}, \boldsymbol{\alpha}, \boldsymbol{\mu}) = \frac{1}{2}\boldsymbol{w}^T\boldsymbol{w} + C\sum_{i=1}^{n}\xi_i + \sum_{i=1}^{n}\alpha_i(1 - \xi_i - y_i(\boldsymbol{w}^T\boldsymbol{x}_i + b)) - \sum_{i=1}^{n}\mu_i\xi_i$
- 对偶问题：

$$\underset{\boldsymbol{\alpha}}{\operatorname{argmax}} \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n}\alpha_i\alpha_j y_i y_j \boldsymbol{x}_i^T\boldsymbol{x}_j$$

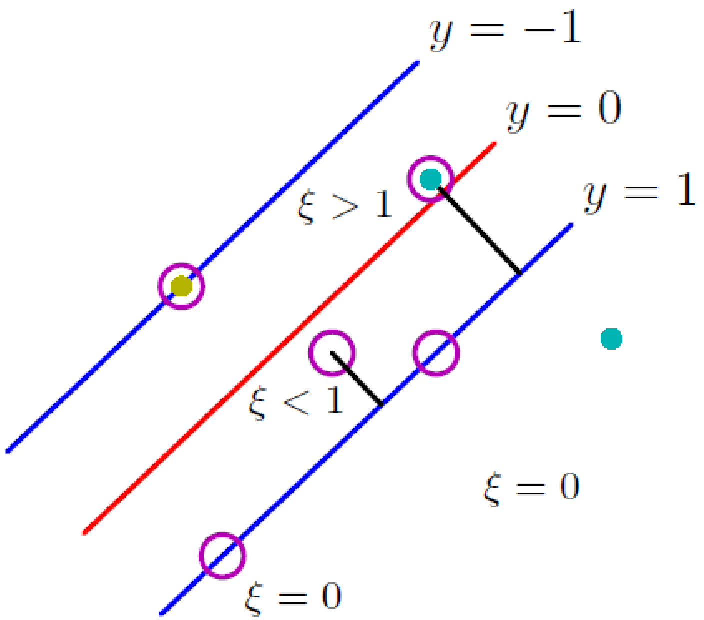
$$s.t. \quad C \geq \alpha_i \geq 0$$

$$\sum_{i=1}^{n}\alpha_i y_i = 0$$

- 对偶形式仅依赖于内积！

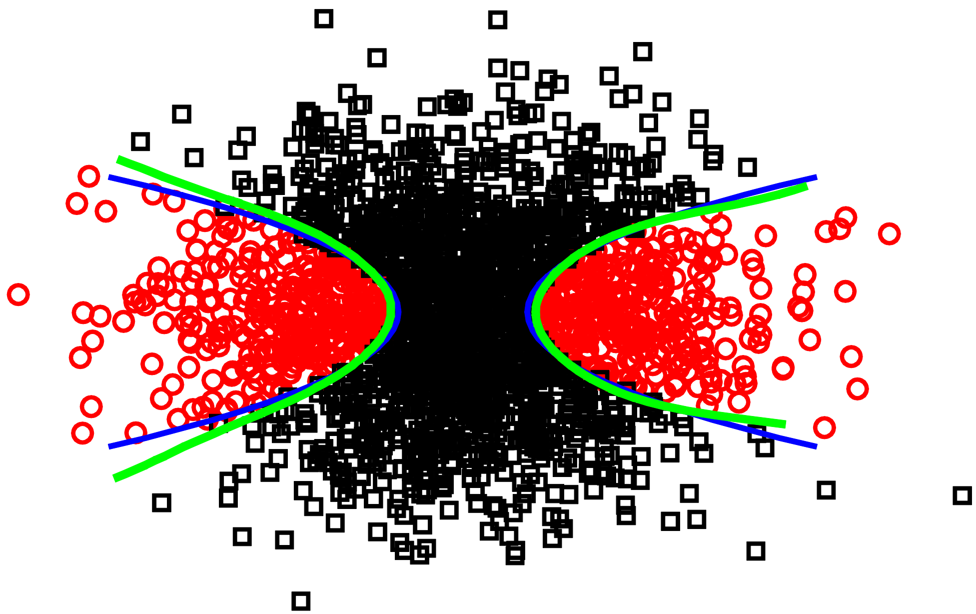
Complementary Slackness

- 对所有i，KKT条件包括$\mu_i\xi_i = (C - \alpha_i)\xi_i = 0$
  - 情况1：$C > \alpha_i > 0$，$\xi_i = 0$, 在特征空间中间隔为1的两个超平面上
  - 情况2：$\alpha_i = C$，对ξ_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：$\alpha_i = 0$，$\xi_i = 0$, 不在特征空间中间隔为1的两个超平面上，在预测时不需要计算



非线性可分例子

- 若使用软间隔SVM
  - ❑线性分类函数有时仍不能很好地解决分类问题



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

$$= 1 + 2x_1 z_1 + 2x_2 z_2 + x_1^2 z_1^2 + x_2^2 z_2^2 + 2x_1 z_1 x_2 z_2$$

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

Kernel trick

- 两个向量$\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^d$, 一个非线性函数$K(\boldsymbol{x}, \boldsymbol{y})$
- 对于满足某些条件的函数K，一定存在一个映射 (mapping) $\phi: \mathbb{R}^d \mapsto \Omega$，使得对任意的$\boldsymbol{x}, \boldsymbol{y}$

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

  - 非线性函数K表示两个向量的相似程度
  - 其等价于Ω里面的内积
- Ω：特征空间（feature space）
  - 可以是有限维的内积空间，也可以是无穷维的内积空间 (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$\int g^2(\boldsymbol{u})d\boldsymbol{u} < \infty$的非零函数，对称函数$K$满足条件： $\iint g(\boldsymbol{u})K(\boldsymbol{u},\boldsymbol{v})g(\boldsymbol{v})d\boldsymbol{u}d\boldsymbol{v} \geq 0$
- 看上去眼熟？另一种等价形式：对**任何**一个样本集合$\{\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n\}, \boldsymbol{x}_i \in \mathbb{R}^d$，如果矩阵$K = [K_{ij}]_{i,j}$(矩阵的第$i$行、第$j$列元素$K_{ij} = K(\boldsymbol{x}_i, \boldsymbol{x}_j)$)总是半正定的，那么函数$K$满足Mercer条件（可利用半正定矩阵性质证明）
- 如何判定是否满足？

核支持向量机Kernel SVM

- 对偶形式：
  $$\underset{\boldsymbol{\alpha}}{\operatorname{argmax}} \sum_{i=1}^{n} \alpha_i - \frac{1}{2}\sum_{i=1}^{n}\sum_{j=1}^{n} \alpha_i \alpha_j y_i y_j K(\boldsymbol{x}_i, \boldsymbol{x}_j)$$
- 分类边界： $\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i\, y_i \phi(\boldsymbol{x}_i)$
- 怎样预测： $\boldsymbol{w}^T \phi(\boldsymbol{x}) = \phi(\boldsymbol{x})^T \left(\sum_{i=1}^{n} \alpha_i\, y_i \phi(\boldsymbol{x}_i)\right) = \sum_{i=1}^{n} \alpha_i\, y_i K(\boldsymbol{x}, \boldsymbol{x}_i)$
  - 线性： $\boldsymbol{w} = \sum_{i=1}^{n} \alpha_i\, y_i \boldsymbol{x}_i$， $\boldsymbol{w}^T\boldsymbol{x}$ 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?
    - 复杂度由$\alpha_i > 0$的个数，而非样本的总数目来决定
    - $\boldsymbol{\alpha}$是稀疏的，复杂度远低于$O(nd)$

非线性核

- 线性核linear kernel, dot-product kernel：$K(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x}^T\boldsymbol{y}$
- 非线性核non-linear kernel
  - RBF(radial basis function)、高斯(Gaussian)核：$K(\boldsymbol{x}, \boldsymbol{y}) = \exp(-\gamma\|\boldsymbol{x}-\boldsymbol{y}\|^2)$
  - 多项式核： $K(\boldsymbol{x}, \boldsymbol{y}) = (\gamma\boldsymbol{x}^T\boldsymbol{y} + c)^d$
  - …

超参数

- 如何决定C、γ、…
  - 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$\{1,2,\ldots,C\}$
  - 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$C(C-1)/2$个输出分布在C个类别
  - 对测试样本$\boldsymbol{x}$，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$\mathrm{sign}(f_i(\boldsymbol{x}))$

	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，把其他所有$C-1$个类别的数据合并在一起做负类
    - 每个新的分类器f_i采用其实数值输出，即$f_i(\boldsymbol{x})$
  - ❑ $f_i(\boldsymbol{x})$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

$$\operatorname*{argmax}_i f_i(\boldsymbol{x})$$

多类Multiclass(3)

- 直接解决多类问题(进一步阅读)
  - Crammer-Singer方法
  - https://www.jmlr.org/papers/v2/crammer01a.html
- DAGSVM(进一步阅读)
  - https://papers.nips.cc/paper/1999/hash/4abe17a1c80cbdd2aa241b70840879de-Abstract.html
- ECOC(进一步阅读)
  - https://www.jair.org/index.php/jair/article/view/10127

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)，从简单情况开始（如，2类）
4. 数学基础和研究
   - 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
  - 但是要保持原问题的核心思想
  - 如SVM从二类、线性、可分的情况开始

进一步的阅读

- 如果对本章的内容感兴趣，可以进一步阅读
  - ❑ 拉格朗日乘子法、KKT条件：
    - Convex Optimization第五章
  - ❑ SVM和统计学习
    - Schölkopf, B. and Smola, A.J., 2002. Learning with kernels: support vector machines, regularization, optimization, and beyond. MIT press.
    - 最新会议论文集：ICML、NIPS、AAAI、IJCAI、…
  - ❑ SMO: http://en.wikipedia.org/wiki/Sequential_minimal_optimization
    - LIBSVM, SVMLight
  - ❑ Pegasos: http://www.cs.huji.ac.il/~shais/code/
  - ❑ LIBLINEAR: http://www.csie.ntu.edu.tw/~cjlin/liblinear/
    - Dual Coordinate Descent (DCD)
  - ❑ Additive Kernel: https://ieeexplore.ieee.org/abstract/document/6136519