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（II）

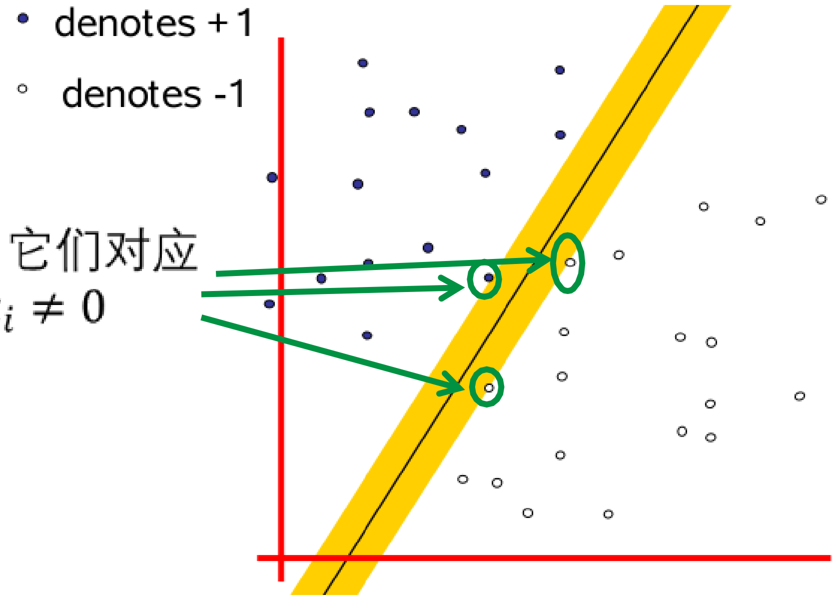
*图片均来自网络或已发表刊物

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
- 核函数支持向量机
- 核心概要

回顾

- SVM 在线性可分时
  - 最大化间隔



SVs: 它们对应的 $\alpha_i \neq 0$

- 原始问题:

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

- 对偶问题:

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

线性不可分情况

- 在线性可分情况下：

保证训练分类错误为零

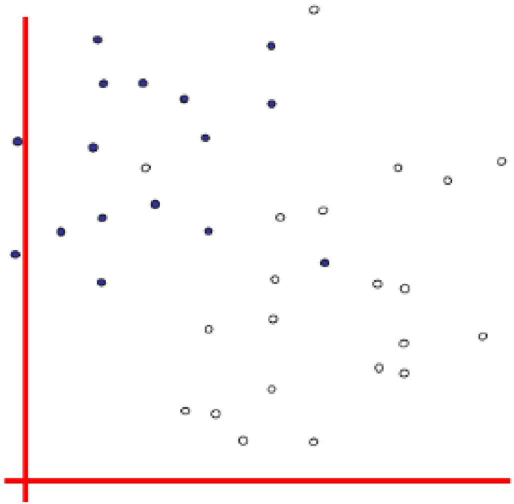
$$\min_{w,b} \quad \frac{1}{2}\|w\|_2^2$$
$$\text{s.t.} \quad y_i\left(\boxed{\langle w, x_i\rangle + b) \geq 1, \quad \forall i = 1, \ldots, N}\right.$$

- 但是在线性不可分情况下，一定会有错误。

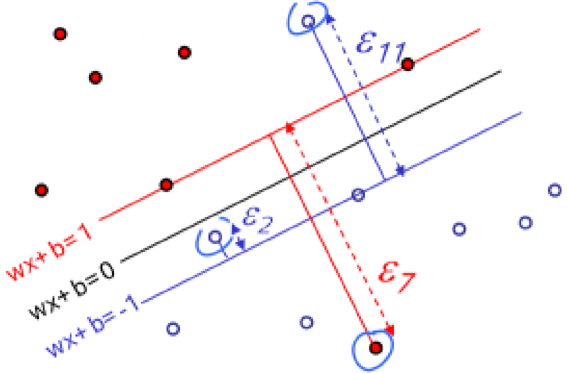
我们需要最小化 $\|w\|_2^2$ 和 训练分类错误!

$$\min_{w,b} \frac{1}{2}\|w\|_2^2 \boxed{+ C \times (\text{loss for errors})}$$

其中 $C > 0$ 是一个用于平衡这二者的常数



损失函数：0/1 损失 和 Hinge 损失



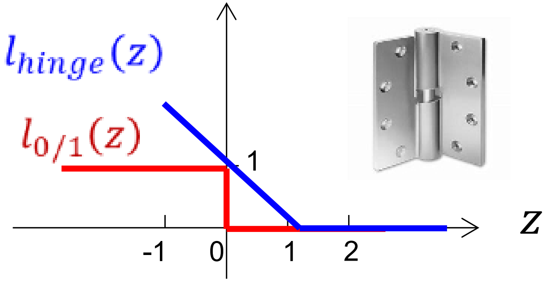
- 回顾正确的预测：$y_i(\langle w, x_i \rangle + b) \geq 1$
- 定义：$z_i \triangleq y_i(\langle w, x_i \rangle + b)$
- 对每个样本 x_i
    - 0/1 损失：

      PS: 分类超平面是 $\langle w, x_i \rangle + b = 0$, i.e., $z_i = 0$

      $l_{0/1}(z_i) \triangleq 1(z_i < 0)$ 非凸非连续，不容易求解
    - Hinge 损失：

      $l_{hinge}(z_i) \triangleq \max(0, 1 - z_i)$

      线性惩罚 z_i <1



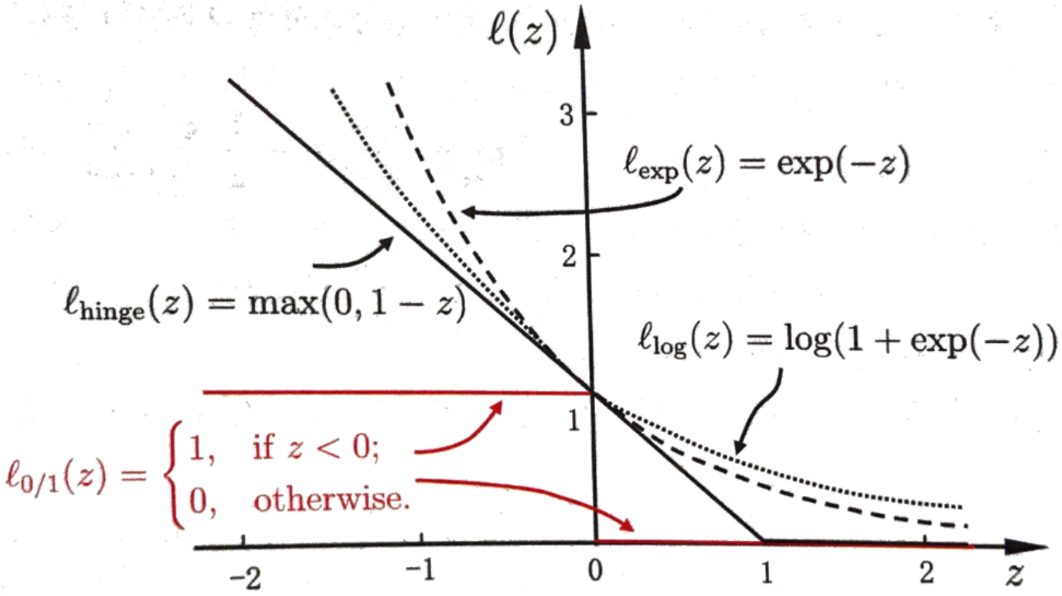
更多损失函数

- 指数损失

$$l_{exp}(z) \triangleq e^{-z_i}$$

- 对率损失

$$l_{log}(z) \triangleq \log(1 + e^{-z_i})$$



三种常用的损失函数，可替代0/1损失

形式化损失函数

- 引入松弛变量（slack variables） $\varepsilon_i \geq 0$:

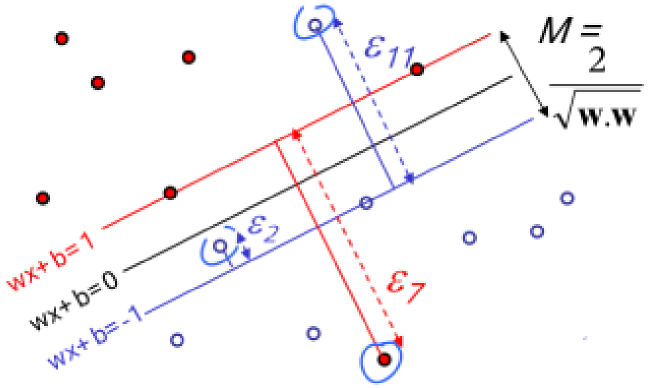
$$\begin{aligned} \min_{w,b,\varepsilon_i} \quad & \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{N}\varepsilon_i \\ \text{s.t.} \quad & y_i\left(\langle w, x_i\rangle + b\right) \geq 1 - \varepsilon_i, \quad \varepsilon_i \geq 0, \quad \forall i = 1, \ldots, N \end{aligned}$$

新变量

与线性可分情况对比:

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

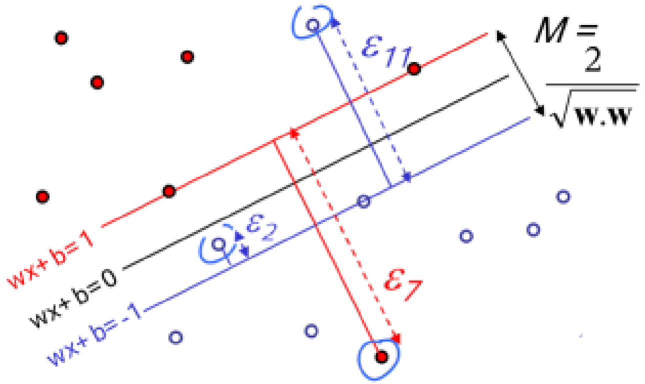
软间隔（soft margin）



$$
\begin{aligned}
\min_{w,b,\varepsilon_i} \quad & \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{N}\varepsilon_i \\
\text{s.t.} \quad & y_i\left(\langle w, x_i\rangle + b\right) \geq 1 - \varepsilon_i, \\
& \varepsilon_i \geq 0, \quad \forall i = 1, \ldots, N
\end{aligned}
$$

- 仍然希望找到最大间隔超平面，但此时:
  - 我们允许一些训练样本被错分类
  - 我们允许一些训练样本在间隔区域内

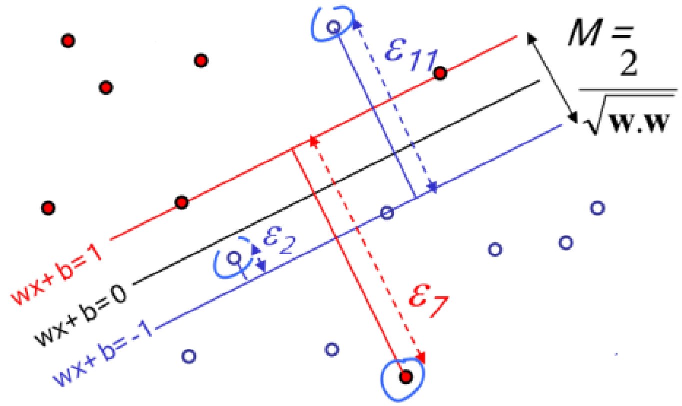
软间隔



$$
\begin{aligned}
\min_{w,b,\varepsilon_i} \quad & \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{N}\varepsilon_i \\
\text{s.t.} \quad & y_i\left(\langle w, x_i\rangle + b\right) \geq 1 - \varepsilon_i, \\
& \varepsilon_i \geq 0, \quad \forall i = 1, \ldots, N
\end{aligned}
$$

- $\varepsilon_i = 0$时，数据点在间隔区域的边界上 或 在间隔区域外部正确分类的那一侧
- $0 < \varepsilon_i \leq 1$时，数据点在间隔区域内但在正确分类的一侧
- $\varepsilon_i > 1$时，数据点在分类超平面 错误分类 的一侧.

软间隔



$$\min_{w,b,\varepsilon_i} \quad \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{N}\varepsilon_i$$
$$\text{s.t.} \quad y_i(\langle w, x_i\rangle + b) \geq 1 - \varepsilon_i,$$
$$\varepsilon_i \geq 0, \quad \forall i = 1, \ldots, N$$

- 正值常数 C **平衡着** 大间隔和小分类错误
  - Structure risk (结构风险) vs. empirical risk (经验风险)
  - 大 C: 更偏向于错误小
  - 小 C: 更偏向于间隔大

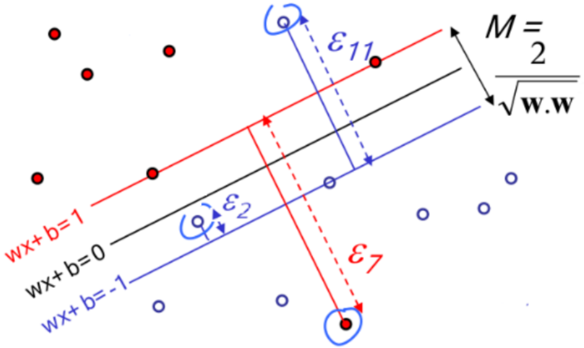
对偶问题

- 线性可分情况下的对偶问题

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

- 线性不可分情况下的对偶问题

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



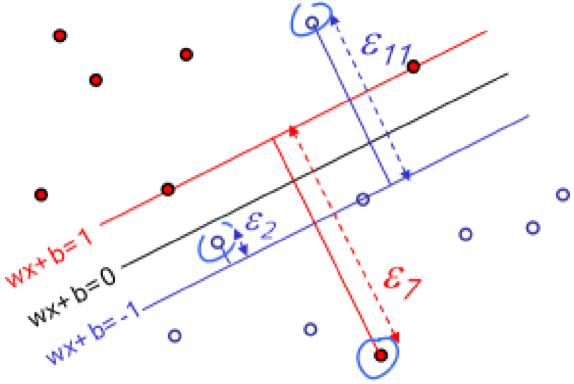
$\alpha_i = 0$: non-SVs

$\alpha_i > 0$: SVs

(除了间隔边界上的数据点以外，那些在间隔区域内的、以及在错误一侧的数据点，也都是SV)

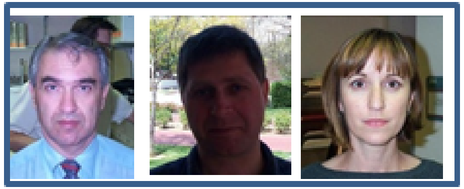
目前为止

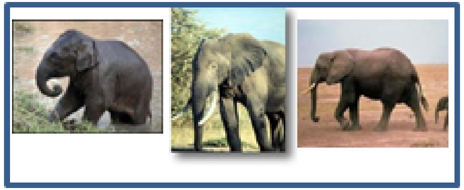
- 线性可分情况下的SVM： 原始问题和对偶问题
- 线性不可分情况下的SVM
  - 原始问题: $$\begin{aligned} \min_{w,b,\varepsilon_i} \quad & \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^{N}\varepsilon_i \\ \text{s.t.} \quad & y_i\left(\langle w, x_i\rangle + b\right) \geq 1 - \varepsilon_i, \\ & \varepsilon_i \geq 0, \quad \forall i = 1,\dots,N \end{aligned}$$
  - 对偶问题: $$\begin{aligned} \min_{\{\alpha_i\}} \quad & \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle - \sum_i \alpha_i \\ \text{s.t.} \quad & \sum_i y_i\alpha_i = 0, \\ & C \geq \alpha_i \geq 0, \quad \forall i = 1,\dots,N \end{aligned}$$

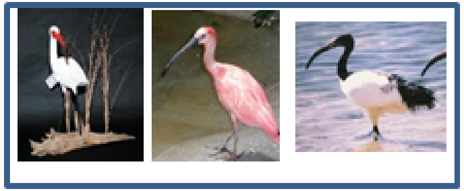


应用举例: 利用LLC 和 SVM的图像分类

- 数据集: Caltech101
  - 9144 张图片
  - 102 类别
- 预处理
  - 转化为灰度图
  - 放缩图片 e.g. 较长边有120 像素
- 用LLC (底层特征，Low Level Content) 来提取图像特征
- 用SVM训练和测试
- 测试结果
  - 训练时每个类别用 15 张图片: 70.16%
  - 训练时每个类别用 30 张图片: 73.44%







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 （Kernel SVM）
  - 从输入空间到特征空间
  - 核函数及其构造
  - 基于核的学习
- 核心概要

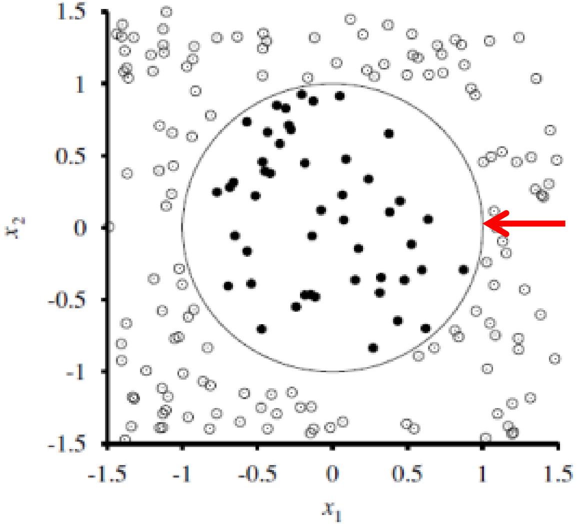
特征空间

- 输入空间(input space) 到 特征空间(feature space)

$$\Phi(x) : R^n \rightarrow F$$

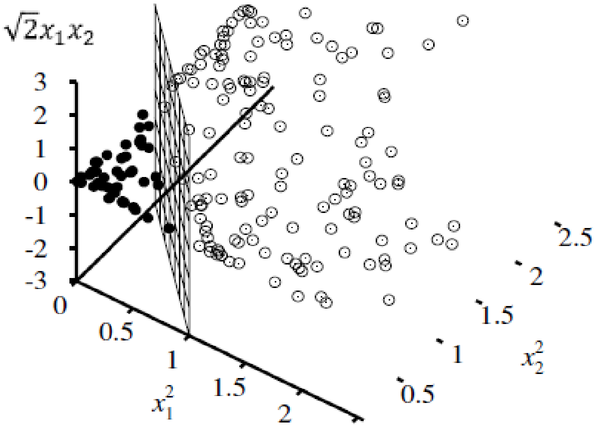
- 样本 x_i 在输入空间中线性不可分，但可能在特征空间中线性可分

在2D空间(x_1, x_2) 上



$x_1^2 + x_2^2 \leq 1$

同样的点在空间 $(x_1^2, x_2^2, \sqrt{2}x_1x_2)$上



特征空间中的问题

线性不可分情况

	原始问题	对偶问题
输入空间	$\min_{w,b,\varepsilon_i} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^N \varepsilon_i$ s.t. $y_i\left(\langle w, x_i\rangle + b\right) \geq 1-\varepsilon_i,$ $\varepsilon_i \geq 0, \quad \forall i = 1,\ldots,N$	$\min_{\{\alpha_i\}} \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \langle x_i, x_j\rangle - \sum_i \alpha_i$ s.t. $\sum_i y_i\alpha_i = 0,$ $C \geq \alpha_i \geq 0, \quad \forall i = 1,\ldots,N$
特征空间	转换 x_i 为 $\Phi(x_i)$ $\min_{w,b,\varepsilon_i} \frac{1}{2}\|w\|_2^2 + C\sum_{i=1}^N \varepsilon_i$ s.t. $y_i\left(\langle w, \Phi(x_i)\rangle + b\right) \geq 1-\varepsilon_i,$ $\varepsilon_i \geq 0, \quad \forall i = 1,\ldots,N$	转换 x_i 为 $\Phi(x_i)$ $\min_{\{\alpha_i\}} \frac{1}{2}\sum_{i,j}\alpha_i\alpha_j y_i y_j \underline{\langle \Phi(x_i), \Phi(x_j)\rangle}$ $-\sum_i \alpha_i$ s.t. $\sum_i y_i\alpha_i = 0,$ $C \geq \alpha_i \geq 0, \quad \forall i = 1,\ldots,N$

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 （Kernel SVM）
  - 从输入空间到特征空间
  - 核函数及其构造 （Kernel Function）
  - 基于核的学习
- 核心概要

核技巧 (Kernel Tricks)

- 为了*在特征空间中*求解对偶问题且找到分类超平面，我们只需要知道

$$\langle \Phi(x), \Phi(y) \rangle$$

而不是分别的 $\Phi(x)$ 和 $\Phi(y)$

- 如果我们已知一个函数$k(x,y)$，它等于

$$\langle \Phi(x), \Phi(y) \rangle$$

那么我们就没有必要显式地表示这些特征

- $k(x,y)$ 称作 核函数（kernel function）

常用核函数

- 齐次多项式 Homogeneous polynomials $k(x,y)=(\langle x,y\rangle)^d$
- 非齐次多项式 Inhomogeneous polynomials $k(x,y)=(\langle x,y\rangle+1)^d$
- 高斯核函数 Gaussian Kernel $k(x,y)=\exp\left(-\frac{\|x-y\|^2}{2\sigma^2}\right)$
- Sigmoid核函数 Sigmoid Kernel $k(x,y)=\tanh(\eta\langle x,y\rangle+v)$

*tanh: 双曲正切函数

(一个使用了sigmoid核函数的SVM 模型等价于一个两层的感知机神经网络)

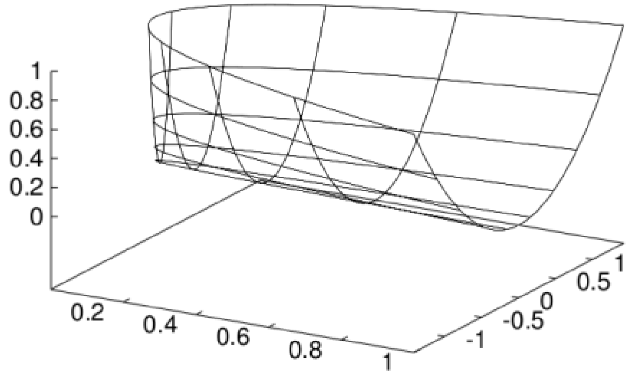
更多常用核函数的列表: https://blog.csdn.net/chlele0105/article/details/17068949

核函数应用举例

- 多项式核函数 $k(x,y) = (\langle x,y\rangle)^d$

  当 n = 2 (x 和 y的维度为2) 且 d = 2，则

$$k(x,y) = \left\langle \begin{pmatrix} x_1 \\ x_2 \end{pmatrix}, \begin{pmatrix} y_1 \\ y_2 \end{pmatrix} \right\rangle^2 = (x_1y_1 + x_2y_2)^2$$



- 存在一个映射：$\Phi(x) = \left(x_1^2, \sqrt{2}x_1x_2, x_2^2\right)$

  证明:

$$\langle\Phi(x),\Phi(y)\rangle = \left\langle \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix}, \begin{pmatrix} y_1^2 \\ \sqrt{2}y_1y_2 \\ y_2^2 \end{pmatrix} \right\rangle = (x_1y_1 + x_2y_2)^2$$

- 映射和特征空间都不是唯一的

$$\Phi(x) = (x_1^2, x_1x_2, x_1x_2, x_2^2) \qquad \Phi(x) = \frac{1}{\sqrt{2}}\left(x_1^2 - x_2^2, 2x_1x_2, x_1^2 + x_2^2\right)$$

三种构造核函数的方法

1. 选择一个特征函数$\Phi(\cdot)$，然后构造核函数 $k(x,x') = \langle \Phi(x), \Phi(x') \rangle$
2. 直接选择一个合理的核函数而不用显式构造 $\Phi(\cdot)$
3. 利用简单核函数构造新的核函数

已知合理的核函数 $k_1(x,x')$ 和 $k_2(x,x')$，则下面的新的核函数也是合理的:

$$
\begin{aligned}
k(x,x') &= ck_1(x,x') \\
k(x,x') &= f(x)k_1(x,x')f(x') \\
k(x,x') &= q(k_1(x,x')) \\
k(x,x') &= \exp(k_1(x,x')) \\
k(x,x') &= k_1(x,x') + k_2(x,x') \\
k(x,x') &= k_1(x,x')k_2(x,x') \\
k(x,x') &= x^T A x'
\end{aligned}
$$

其中 $c>0$： 常数， $f(\cdot)$： 任何函数， $q(\cdot)$： 有非负系数的多项式，

A： 半正定对称矩阵（symmetric positive semidenite matrix）

举例: 构造高斯核函数

- 高斯核函数 $k(x,x') = \exp(-\|x-x'\|^2/2\sigma^2)$
- 注意 $\|x-x'\|^2 = \langle x,x\rangle + \langle x',x'\rangle - 2\langle x,x'\rangle$

$$k(x,x') = \exp\left(-\frac{\langle x,x\rangle}{2\sigma^2}\right)\exp\left(\frac{\langle x,x'\rangle}{\sigma^2}\right)\exp\left(-\frac{\langle x',x'\rangle}{2\sigma^2}\right)$$

- 即然 $\langle x,x'\rangle$ 是一个核函数，根据之前讲义中的规则， 高斯函数也是一个合理的核函数
- 注意关于该高斯核函数的特征向量的维度可以是无穷

软件工具

- Libsvm
  - http://www.csie.ntu.edu.tw/~cjlin/libsvm/
- Liblinear
  - http://www.csie.ntu.edu.tw/~cjlin/liblinear/
- SVMlight
  - http://svmlight.joachims.org/

SVM的核心概要

- 线性支持向量机
  - 线性可分问题: 最大化间隔
    - 原始问题和对偶问题
  - 线性不可分问题: 最大化间隔且最小化分类错误
    - 原始问题和对偶问题
- 核函数支持向量机
  - 映射到特征空间: $\Phi\colon R^n \to F$
  - 核技巧: $k(x,y) = \langle \Phi(x), \Phi(y) \rangle$
  - 3 种构造核函数的方法

SVM的优缺点

- 优点
  - 很好的数学基础
  - 最大化间隔使得方法的鲁棒性非常高，泛化能力强
  - 用线性的方法解决线性不可分问题（两种思路）
    - 利用soft margin: 最大化间隔且最小化分类错误
    - 通过从输入空间到特征空间的变换
      - 本质上是将在低维空间上线性不可分的问题，通过变换（大多是升维）变成线性可分问题
      - 利用Kernel Trick，并不需要知道该变换是什么
  - 在实际应用中效果往往不错
- 缺点
  - 有多种核函数可用，针对具体问题，哪个核函数最好？—— 尚未找到理论上可证