第7课时 支持向量机SVM

主讲教师：欧新宇

January 25, 2020

Outlines

- Explain SVM like I am a 5 year old
- 支持向量机的基本原理
- 支持向量机的数学表达（略）
- 核函数介绍
- 不同核函数的对比
- 超参数调节和分析
- SVM实例——波士顿房价回归分析

第7课时 支持向量机SVM

原始SVM算法是由弗拉基米尔·万普尼克和亚历克塞·泽范兰杰斯于1963年发明的。1992年，Bernhard E. Boser、Isabelle M. Guyon和弗拉基米尔·万普尼克提出了一种通过**将核技巧应用于最大间隔超平面来创建非线性分类器**的方法。当前标准的前身（软间隔）由Corinna Cortes和Vapnik于1993年提出，并于1995年发表。

上个世纪90年代，由于人工神经网络的衰落，SVM在很长一段时间里都是当时的明星算法。被认为是一种理论优美且非常实用的机器学习算法。

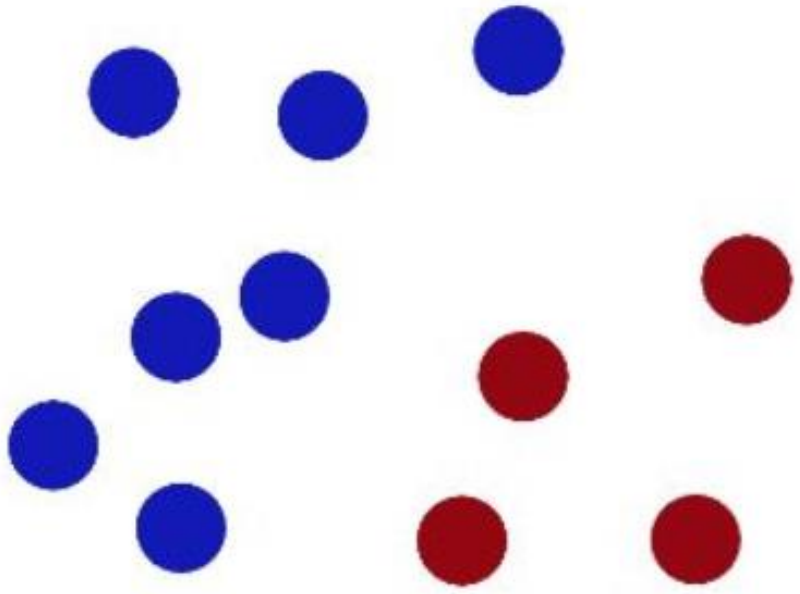
第7课时 支持向量机SVM

在理论方面，SVM算法涉及到了非常多的概念：**间隔(margin)**、**支持向量(support vector)**、**核函数(kernel)**、**对偶(duality)**、**凸优化**等。有些概念理解起来比较困难，例如kernel trick和对偶问题。在应用方法，SVM除了可以当做有监督的分类和回归模型来使用外，还可以用在无监督的聚类及异常检测。相对于现在比较流行的**深度学习**（适用于解决**大规模非线性问题**），**SVM**非常擅长解决复杂的具有**中小规模训练集的非线性问题**，甚至在特征多于训练样本时也能有非常好的表现*（深度学习此时容易过拟合）*。但是随着样本量m的增加，SVM模型的计算复杂度会呈 m^2 或 m^3 增加。

Explain SVM like I am a 5 year old

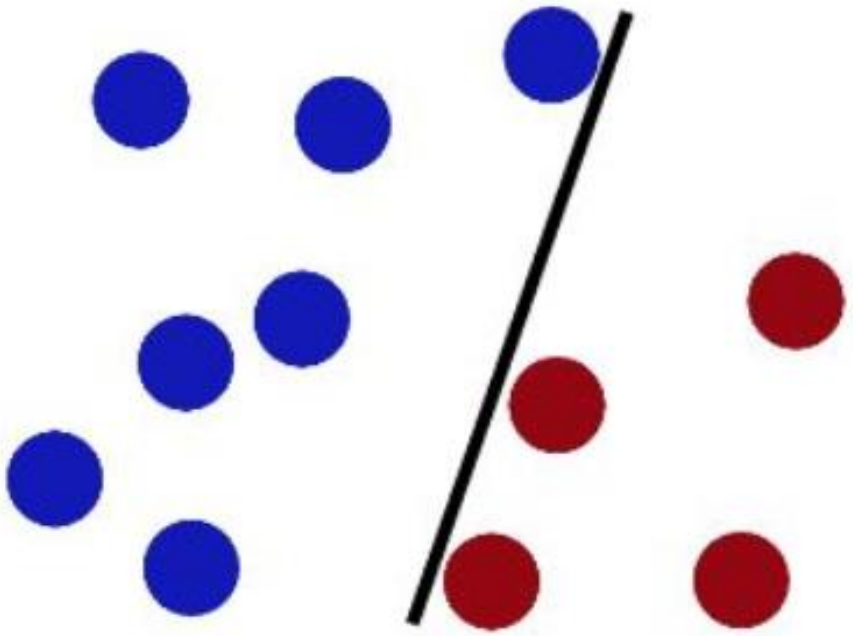
很久很久以前，有个大侠的爱人被魔鬼抓走了，魔鬼要这位大侠和它玩一个游戏才能放了大侠的爱人。

魔鬼在桌子上似乎很有规律地方了两种颜色的球，然后说到："你需要用一根棍子将它们分开，并且要求在后续放更多球之后，仍然适用。"



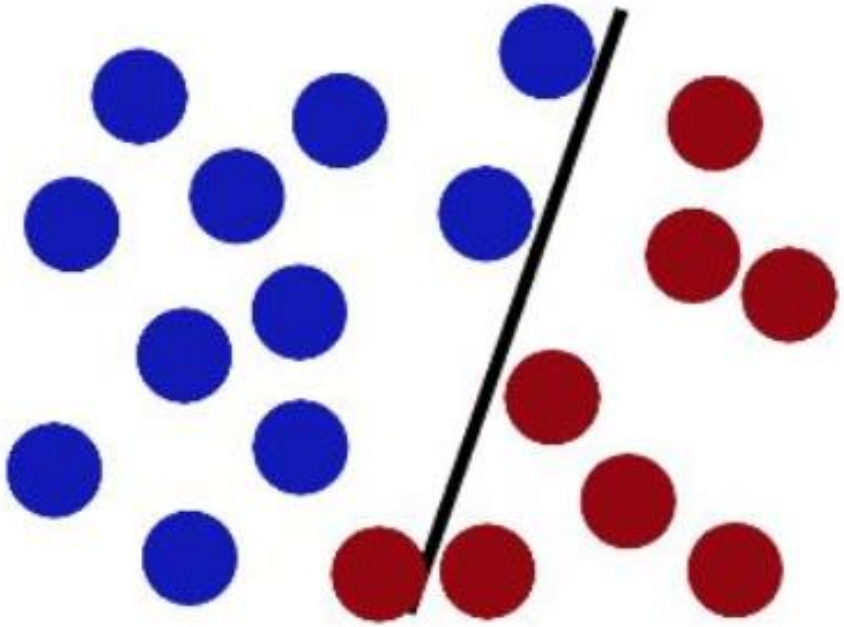
Explain SVM like I am a 5 year old

于是乎，大侠这样放下了棍子，看起来还不错。



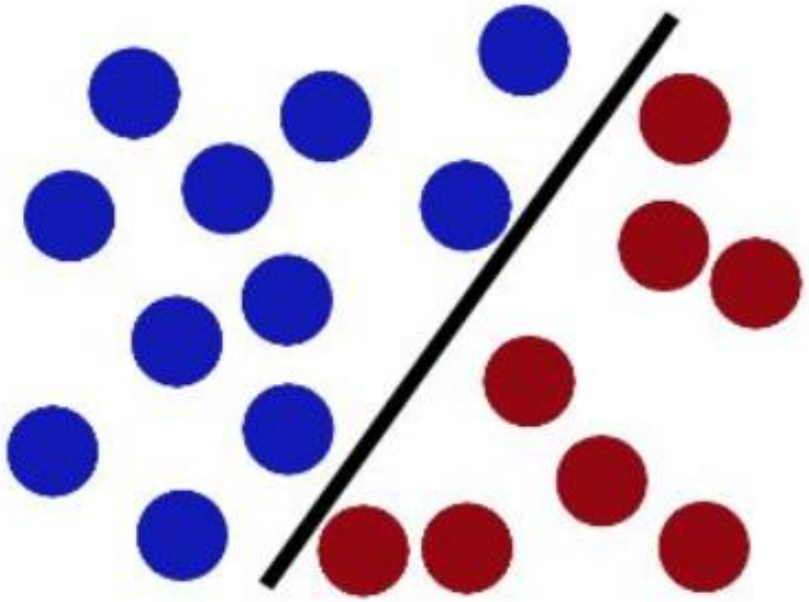
Explain SVM like I am a 5 year old

魔鬼又在桌上放下了更多的球，似乎也还不错，不过有一个球站错阵营了。



Explain SVM like I am a 5 year old

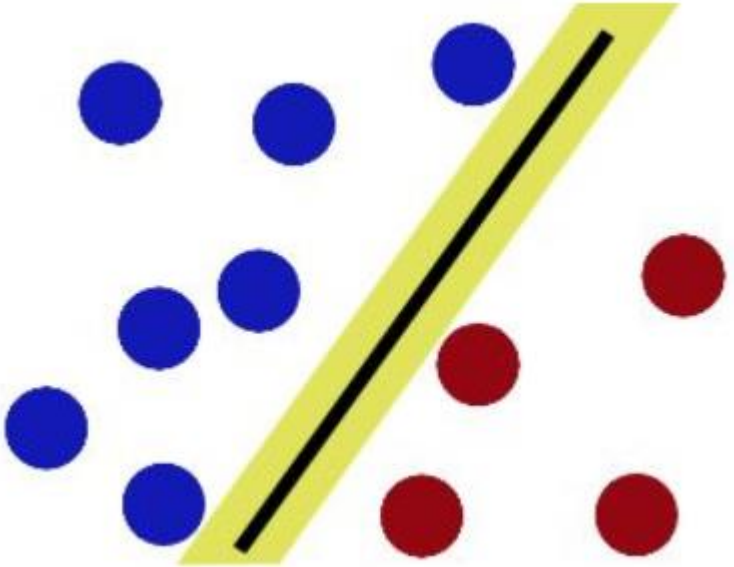
于是，大侠祭出了一件神级法宝 —— 支持向量机（Support Vector Machine，SVM）。利用SVM，大侠让棍子再一次完美地充当了一条分割线。



Explain SVM like I am a 5 year old

SVM的特性一：

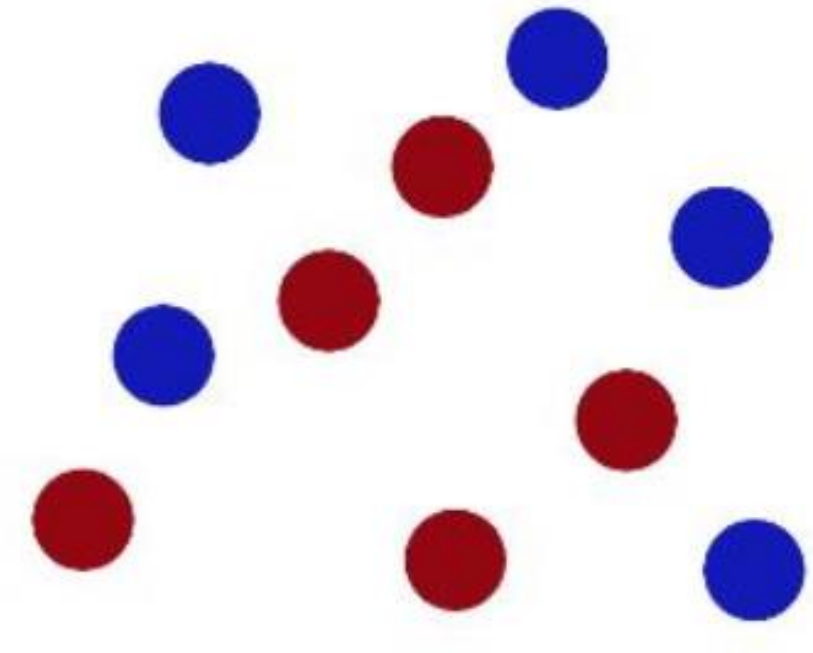
试图建立一条完美的分界线，让该分界线处于最佳的位置，让分界线两边与样本间有尽可能大的间隙。



Explain SVM like I am a 5 year old

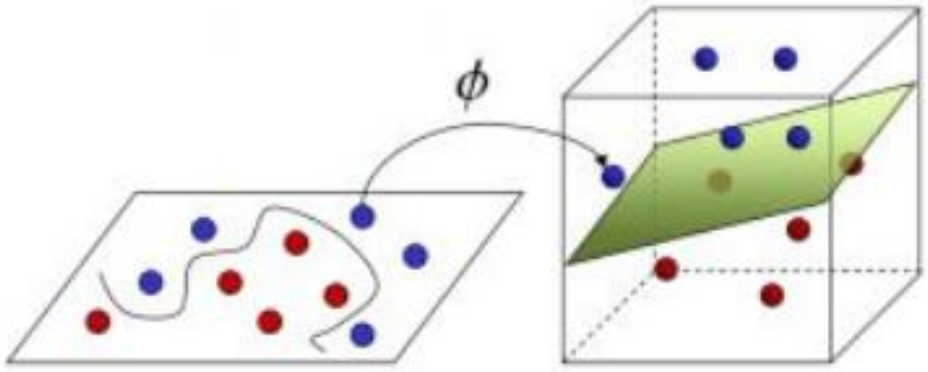
借助神级法宝SVM的第一个trick —— **最大类间间隙**，大侠度过了第一关。

于是，魔鬼给了大侠一个更困难的挑战。



Explain SVM like I am a 5 year old

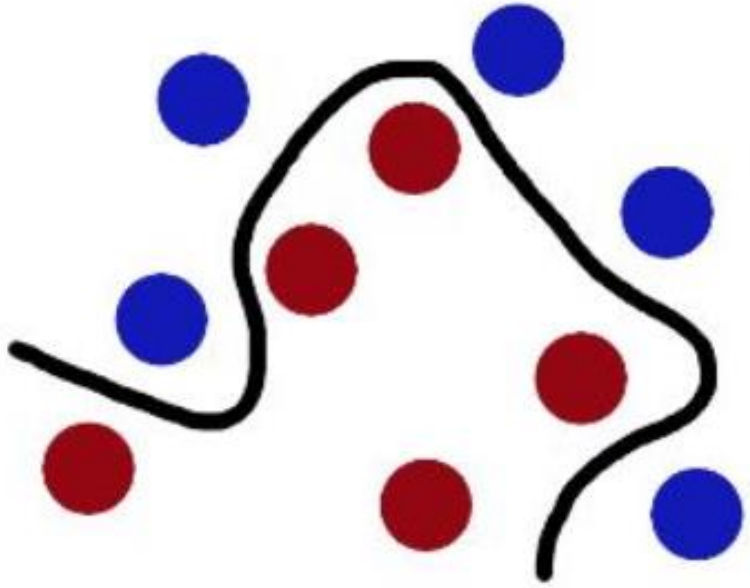
这个问题让大侠范畴了，似乎一条棍子根本没办法将两种颜色的球进行分隔。此时，大侠想到了法宝SVM的另一个神技 —— **超平面**。于是乎，大侠用力一拍桌子，所有的球都飞到了空中。凭借大侠"快准狠"的身手，他迅速将一张纸插入到球的中间。



输入空间　　特征空间

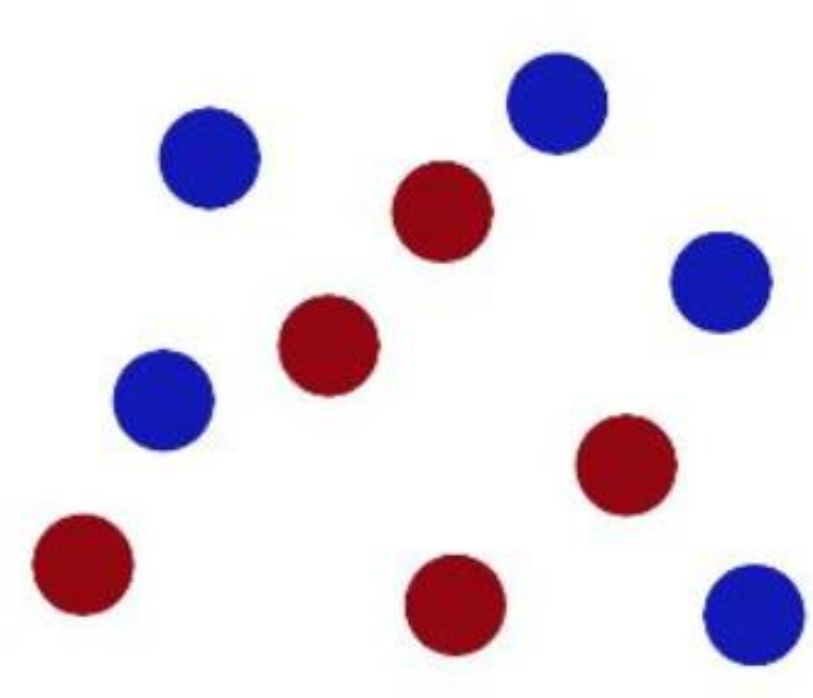
Explain SVM like I am a 5 year old

此时，从魔鬼的角度来看，这些球看起来就像是被一条曲线给分开了。



Explain SVM like I am a 5 year old

借助神级法宝SVM的第一个trick —— **最大类间间隙**，大侠度过了第一关。于是，魔鬼给了大侠一个更困难的挑战。



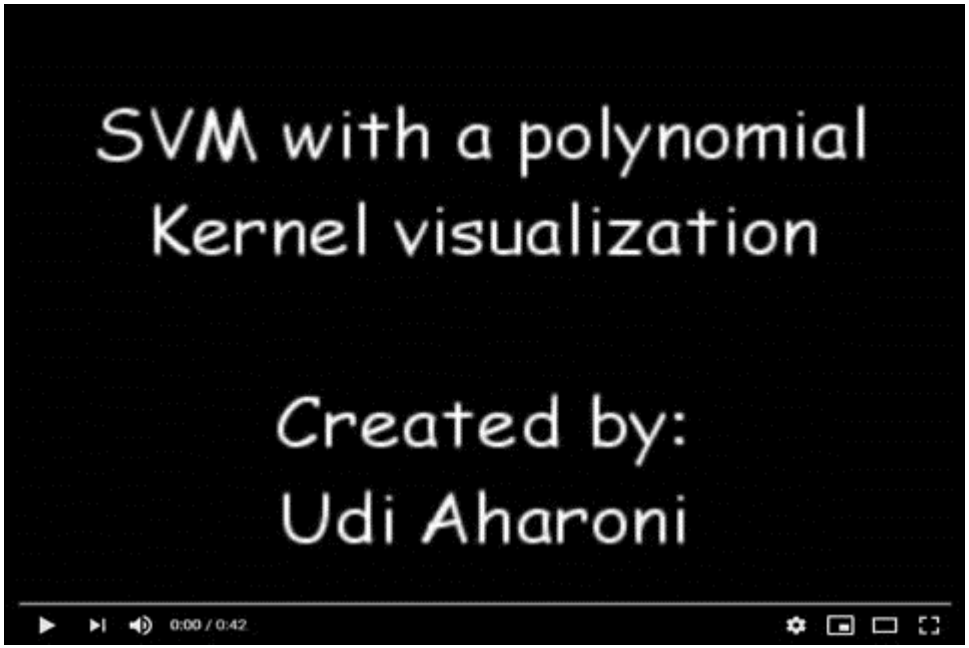
Explain SVM like I am a 5 year old

SVM的特性二：

由于样本特征的特性，当样本在原始特征空间中线性不可分时，我们可以将其转换到高维空间，并利用高维空间中的超平面（HpyerPlane）对样本进行分隔。

Explain SVM like I am a 5 year old

很多年以后，神界无聊众神们认真总结并研究了这个故事。它们把那些带颜色的球称为数据 (data)，把棍子称为分类器 (classifier)，把最大间隙trick称为优化 (optimization)，拍桌子的绝招称为核化 (kernelling)，而那张纸就是超平面 (hyperplane).



支持向量机SVM的基本原理

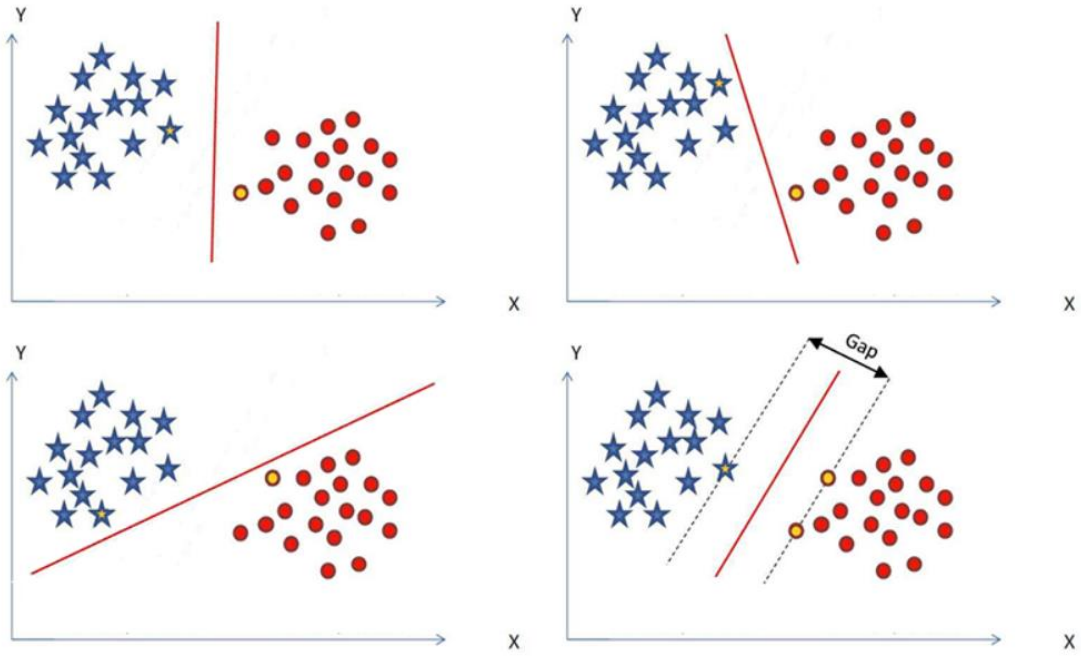
SVM（Support Vector Machines）支持向量机是在所有知名的数据挖掘及传统机器学习算法中最健壮，最准确的方法之一，它属于二分类算法，可以支持线性和非线性的分类。

当然，SVM也可以支持多分类。

支持向量机SVM的基本原理

● 基本原理

首先，我们来了解一下线性分类器。假设在一个二维线性可分的数据集中，如下图所示，我们需要找一个超平面把两组数据分开。图中的四条直线都可以实现分隔两种数据，然而哪一条直线能够达到更好的泛化能力呢？换句话说，我们需要找到一个能够使两个类的空间最大的超平面。



支持向量机SVM的基本原理

● 基本原理

在二维空间中，超平面就是一条直线（例如上图中的分割线，或大侠放的棍子）；而在三维空间中就是一个平面（例如大侠插入球中间的纸）。

我们将这个划分数据的决策边界统称为超平面。距离这个超平面最近的点就叫做**支持向量**，点到超平面的距离叫做**间隔**。

支持向量机就是要使超平面和支持向量之间的间隔尽可能的大，这样超平面才可以更好地将两类样本进行准确分隔。

【重点】支持向量机的核心任务是：

最大化类间距，最小化类内距。

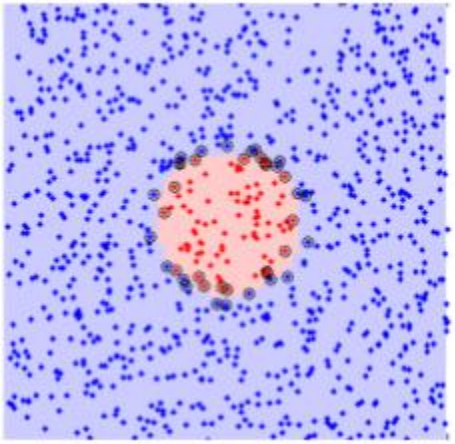
支持向量机SVM的数学表达

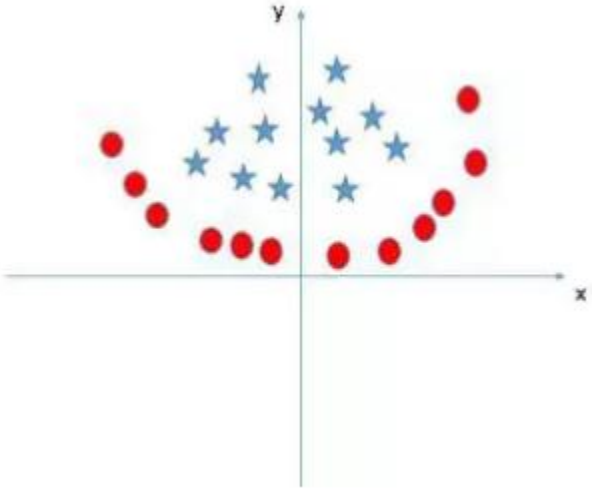
- SVM的数学表达

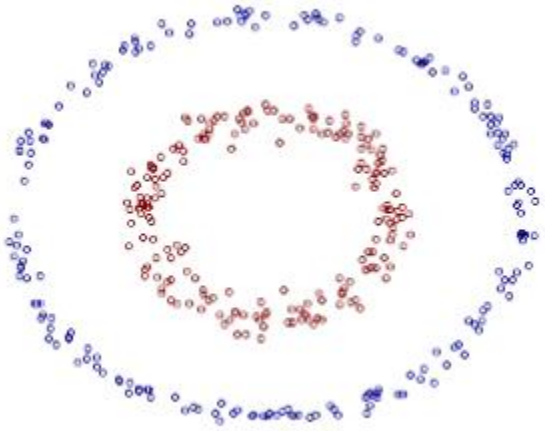
暂略...

支持向量机SVM的核函数

- 为什么要使用核函数?







对于非线性分布的样本（即线性不可分问题 ）我们该如何进行分类呢？

- Linear模型很难进行处理
- SVM有一个优秀的trick —— 核函数K(*,*)，它通过将数据映射到高维空间，来解决原始空间线性不可分的问题。

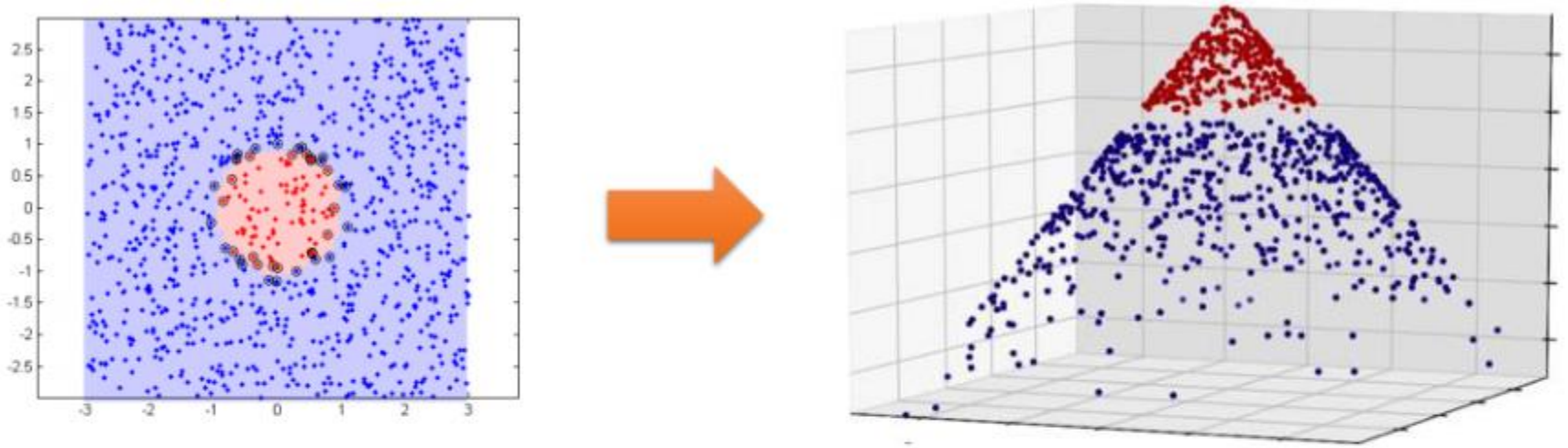
支持向量机SVM的核函数

- 基于核函数的SVM的基本工作流程是：
  - 在低维空间中完成特征计算
  - 通过核函数将输入空间中的特征映射到高维特征空间
  - 在高维特征空间中构造最优分离超平面

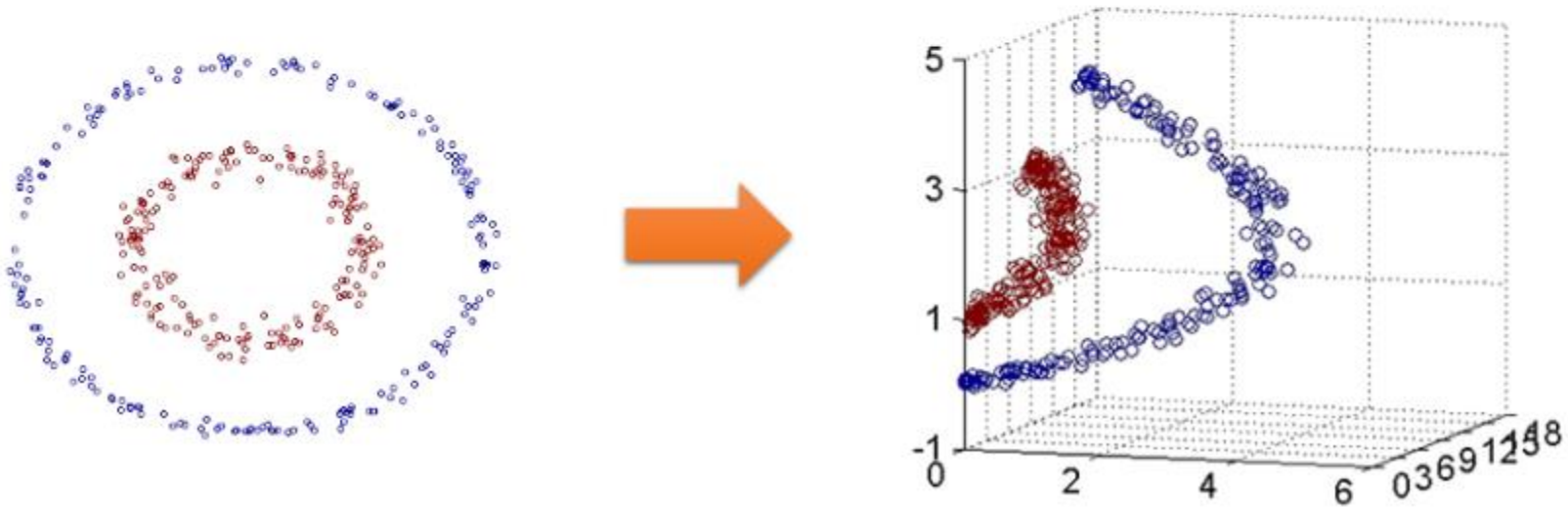
通过以上的操作，SVM可以实现将原始特征空间不好分隔的非线性数据进行最优分隔。

支持向量机SVM的核函数

图一：



图二：



支持向量机SVM的核函数

- 小结：
  - SVM的春天在于核函数；
  - SVM不仅仅能用于二分类，也同样可以用于多分类；同时也可以实现回归和聚类；
  - 核函数的主要功能是将特征映射投射到高维空间以实现"线性可分"；
  - 相比于简单地将特征映射到高维空间，核函数的价值在于：它所有的计算都是基于原始空间，只是将实质的分类效果表现到高维空间。这种机制避免的维度爬升的不可预见性（维度灾难），以及高维空间计算的复杂性。

SVM的使用

- 基于SVM的分类
  - 线性核Linear kernel (Ch0701introLinearSVM.py)
  - RBF核 (Ch0702introRBFSVM.py)
  - 基于不同核函数的SVM对比 (Ch0703KernelCompare.py)
  - 超参数调节与分析
    - RBF核的Gamma值 (Ch0704RBFGramma.py)
    - 多项式核的Degree超参数 (Ch0705PolyDegree.py)
- 基于SVM的回归——波士顿房价回归分析 (Ch0706CaseBoston.py)
  - 模型优化——利用正则化优化各特征之间的量级差 (..norm.py)
  - 模型优化——超参数调整 (..Hyperparameter.py)

欧老师的联系方式

读万卷书 行万里路 只为最好的修炼

QQ：14777591 (宇宙骑士)

Email：ouxinyu@alumni.hust.edu.cn

Tel：18687840023