什么是支持向量机

支持向量机（Support Vector Machine, SVM）是一个非常特立独行的存在，它和符号学派不沾边，既不属于贝叶斯学派，也不属于连接学派，基本的假设就是样本离哪个类近，就属于哪个类。尽管假设如此简单，SVM 却是机器学习中应用最广泛的分类器，在绝大多数任务中表现出优异的性能。

事实上，很少有一个模型能包含两种天才思想，SVM 却是个例外，而这两个天才思想都是由 Vladimir Vapnik 本人或联合其他合作者提出的，不得不说 Vapnik 是天才中的天才。

1963 年，Vladimir N. Vapnik 和 Alexey Ya. Chervonenkis 首先提出了第一个天才思想，后人称之为“最大边界准则”。什么叫最大边界呢？如图 1 所示，如果我们想找一条分界线将蓝圈和红框分开，一个天然的想法是这两类之间距离最大。两类之间的距离称为边界（Margin），最大化这一边界即是最大边界准则。看起来似乎很自然，但里边的内涵却非常深刻。列举几点如下：(1)我们在选择分类面时，只需考虑每边界上的点，类内的点则不需考虑。换句话说，SVM 是“捡硬柿子捏”的，硬柿子都捏了，软柿子就更不在话下。
因此，SVM 具有天然的鲁棒性。（2）只考虑边界点意味着 SVM 不受数据分布假设的限制，因而可应用于分布复杂的真实数据；（3）只考虑边界点使得训练不受训练样本量的影响，因而对类间的数据不均衡具有天然鲁棒性。

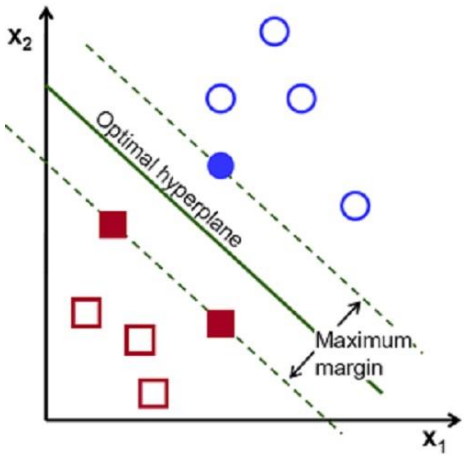
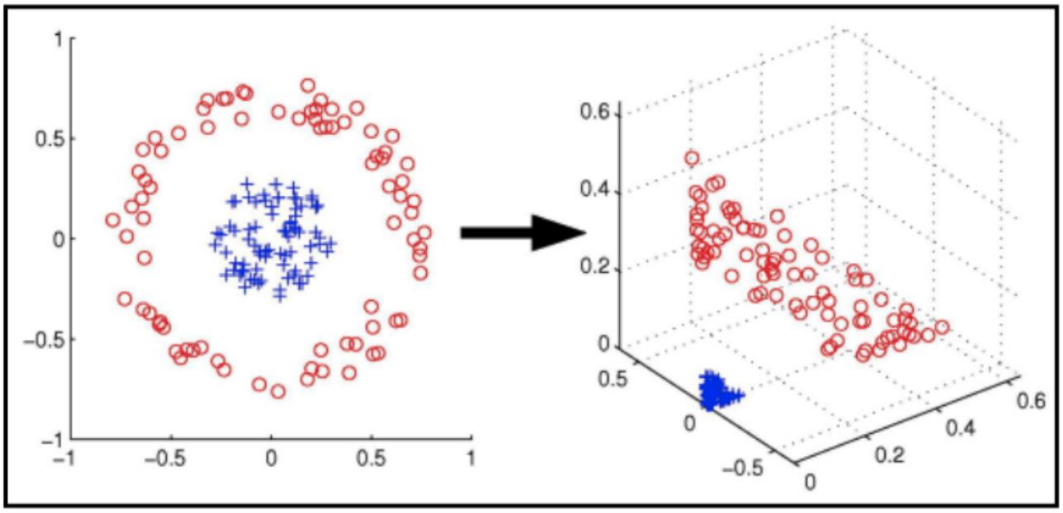


图 1：SVM 模型最大化两类的边界

1992 年，Bernhard Boser, Isabelle Guyon and Vladimir Vapnik 等人提出了第二个天才思想，后人称之为“核方法”。我们知道数据如果比较复杂的话，线性模型是无法实现合理分类的。如图 2 左图所示，一类数据围绕在另一类数据周围，这时无法找到一个好的线性模型将这两类分开。一种方法是用一个非线性映射将数据映射到一个高维特征空间，并在该空间中训练线性分类器模型，如图 2 中右图所示。



如何得到这个非线性映射呢？神经网络采用的是一种数据学习方法，而 Vapnik 的核方法采用了另一种思路：把映射函数的设计归结为数据间距离度量的设计，这意味着如果能设计出一个合理的距离度量，就等价于设计了一个非线性映射。基于这一距离度量，就可以在特征空间中设计线性模型，并将该模型上的所有操作转化为对距离度量的操作，从而实现“隐性”的特征映射和线性建模。利用距离来定义映射是个非常有趣的思路，一是对距离的定义相对直观、简单，且有很多先验知识可用； 二是通过定义距离可以定义非常复杂的映射，把数据映射到极高维度的特征空间甚至是无穷维空间，极大提高了模型的表达能力；第三，也是最重要的，即便距离定义的再复杂，在映射空间中建立的依然是线性模型，因此依然可得到全局最优解。保持训练过程的全局可优化是核方法对神经网络模型的最大优势。

核方法极大扩展了 SVM 模型建模能力，使 SVM 的性能大幅提高。1993 年，Corinna Cortes 和 Vapnik 引入松驰变量以处理错分类的训练样本，进一步扩展了 SVM 模型，奠定了目前广泛使用的 SVM 模型的基础。值得说明的是，线性 SVM 中的两大思想--最大边界和核方法--是互相独立的，各自都有广泛应用。