支持向量机

SVM核心思想

在机器学习领域，支持向量机Support Vector Machine (SVM)是一种经典、有效的分类模型。下面介绍SVM的核心思想：

在二分类任务中，记训练样本集合为 。其中表示样本的特征向量，为维空间的向量；表示样本的真实标签，-1表示该样本是负样本，+1表示该样本是正样本。

所有分类模型的目标都是，在训练集上获得一个能合理区分正负样本的划分规则，然后将此规则应用到真实标签未知的测试样本集合上，以期获得较高的分类精度。

出于简单假设，SVM认为划分正负样本的规则应该是一个超平面，对应的方程为：

在长成的空间中，该划分超平面由参数和唯一确定；但在训练样本集合线性可分的情况下这样的划分超平面有很多可能，通过等价的放缩，可以得到如下不等式组：

可以证明符合上述不等式约束的超平面，都可以成为训练样本集上的一个划分超平面，并且能合理区分正负样本。

为何降低过拟合风险，提高模型的泛化性能，SVM通过最大化（硬）间隔的方式，找到唯一的划分超平面。（硬）间隔是两个异类支持向量到划分超平面的距离之和，而所谓的支持向量便是使上述不等式组等号成立的样本特征向量。（硬）间隔的计算式如下：

所以SVM基本型的优化目标可表示为：

上述为SVM的基本型，也是SVM的核心思想。当最大间隔划分超平面找到后，可以通过如下判别函数，对新样本进行类别划分。

其中为符号函数。

SVM基本模型

事实上，基本型的SVM很难有效的解决实际的问题，因为绝大多数场景的分类任务中：一方面，能合理区分正负样本的划分规则往往与特征向量的关系是非线性的，所以想要沿用SVM的划分超平面假设需要找到非线性到线性的映射关系；另一方面，很多时候需要允许SVM在少量训练样本上划分错误，以进一步减小过拟合的风险。下面我们介绍，为了解决上述问题，而引入“核函数”和“软间隔”的SVM的基本模型。

**核函数**的引入是为了将特征向量非线性映射到某个高维空间的中，在这个高维空间中训练样本集可以被超平面划分。记这样的非线性映射函数为，一般符合下列等式：

其中表示映射后的特征向量，表示核函数，所以在计算高维（无限维）空间中特征向量内积时，可以避免直接计算的困难，转化为计算原空间中核函数的值，也称为核技巧。常用的核函数有：

* 多项式核 多项式次数
* 径向基核（高斯核） 带宽
* Sigmoid核

当参数确定后，每种核函数唯一确定一个核映射函数，用于将特征向量非线性映射到对应的高维空间。

**软间隔**的引入是为了增加SVM在一些样本的上的容错机制。在软间隔的设定下，训练样本集合上所有样本不再严格满足不等限定，即不再严格要求正负样本全部落在划分超平面和支持向量规定的对应区域内，而是给出了一定的松弛度。记对于样本的松弛度为，则新的不等式约束转换为 当然，这样的松弛度也不是无限大的，所以在软间隔SVM中，通过在最小化目标项中加上所有样本的松弛度之和，对松弛度进行正则化，控制整体的松弛度。

综上，SVM基本模型对应的优化目标为：

其中为整体松弛度的正则化系数，越大整体的松弛度越小，软间隔愈加趋向于（硬）间隔。当最大间隔划分超平面找到后，可以通过如下判别函数，对新样本进行类别划分。

其中为符号函数，为核函数确定的核映射函数。

SVM的求解

对于SVM基本模型的优化目标，首先采用拉格朗日乘子法，并令对应的偏导为零，可以转化为：

并且在偏导等于零时，有,

注意到不等式约束，可以转化为KKT(Karush-Kuhn-Tucker)条件约束：

其中为拉格朗日乘子，在KKT条件上采用Sequential Minimal Optimization(SMO)算法可以近似的求出，并在不等式成立时的支持向量集合。

实际中，的计算式如下：

SVM的另一个参数一般不直接求解，而是在判别函数中带入：

多分类扩展

如上述，SVM被设计用作二分类任务，通常也可以one-versus-rest(OVR)或者one-versus-one(OVO)策略扩展到多分类任务中。记样本的多分类类别数为n\_class，则两种策略描述如下：

* OVR策略：在训练样本集合上，为每类样本训练一个SVM模型，正样本为当前类所有样本，负样本为剩余所有的样本，所以一共需要训练n\_class个SVM；在新测试样本集合上，每个SVM为样本给出属于当前类的置信度，样本最终的分类划分结果是置信度最高的SVM对应的类别。
* OVO策略：在训练样本集合上，为两两类间训练一个SVM，正样本为其中一类所有样本，负样本为另一类所有样本，所以一共需要训练n\_class\*(n\_class-1)/2个SVM；在新测试样本集合上，每个SVM为样本给出一个类别判别作为一次投票，样本最终的类别是对应得票最多的类别。

两种策略的分类精度，取决于具体数据分布，一般情况下，性能差不多。

主成分分析

PCA 核心思想

Principle Component Analysis(PCA)是一种常用的无监督降维算法，可以从数据中提取主要构成成分并用于分析，在人脸识别、手写识别等领域应用广泛。下面介绍PCA的核心思想：

在没有（或不考虑）真实标签的情况下，记训练样本集合为。其中表示样本的特征向量，为维空间的向量。通常数据降维算法，希望找到一个更低的维的空间，并在此低维空间中获得样本集合新的表示作为“低维嵌入”，一般这样的嵌入在保留主要信息的同时允许部分信息的丢失。

出于简单假设，PCA算法认为从原空间到嵌入空间的转化经过是一种（线性）坐标转换。我们假设样本（已经进行中心化处理），以为参考的低维空间的标准正交基记为其中，张成的空间记为。则对应的坐标变换结果记为，有**。**当然，这样的低维嵌入空间以及对应的标准正交基，有很多可能的情况，也对应着不同的“重构误差”。从一种思路上看，PCA的核心思想是通多设定最小化“重构误差”为优化目标，则N个样本的计算式如下：

其中，依据嵌入结果可以（部分）重构，记为重构项，重构项与原样本的重构误差定义为向量间的欧氏距离。实际上，最小化重构误差，就等效于在低维嵌入的过程中尽可能多的保留信息，优化目标的解被称为“投影矩阵”，对应的个向量，被称为“主成分”（可以乘上任一非零系数），也就是低维嵌入空间的标准正交基。

注：另外一条理解PCA的思路是从最大化“可分性”出发。

PCA 模型

从PCA的核心思想的论述中，我们可以看出PCA算法的核心是求解出最小化重构误差对应的“投影矩阵”。经推导，上述优化问题可以转化为：

其中表示矩阵的迹，为训练样本集合的协方差矩阵（数据首先要进行中心化），为投影矩阵。

显然，上式优化问题的求解过程，对应的是协方差矩阵特征值分解：求解的矩阵特征值记为，一一对应的单位特征向量记为；依据目标嵌入维度，选取前大的特征值，以及对应的构成。

是PCA算法的重要参数，可以直接指定主成分个数，合理选取十分重要。一方面可以结合后续的学习任务，通过交叉验证，设置使学习器性能最佳的值；如果无法直接考虑后续的学习任务，可以通过设定使下式成立的最小的值:

上式大致衡量了PCA算法在数据降维时信息的保留比，其中一般取95%（可以选择其他阈值），保证了最低保留比。

在得到“投影矩阵”后，对于新样本，进行中心化后，可以直接进行变换，。这样便完成了从高维到低维的嵌入，实现了数据降维任务。