**前言**

这篇文章首先探讨了SVM为什么会对噪声敏感的问题，然后引出了解决办法——软间隔分类器，接着介绍了松弛变量和惩罚因子，并介绍了样本偏斜的解决办法，最后有一个总结。

**SVM为什么会对噪声敏感？**

之前看周志华老师的《机器学习》里说SVM对噪声敏感，后来查了一些资料，然后有了自己的理解，在这里做一个总结。

SVM的基本形态是一个硬间隔分类器，它要求所有样本都满足硬间隔约束(即函数间隔要大于1)，所以当数据集有噪声点时，SVM为了把噪声点也划分正确，超平面就会向另外一个类的样本靠拢，这就使得划分超平面的几何间距变小，降低模型的泛化性能。除此之外，当噪声点混入另外一个类时，对于硬间隔分类器而言，这就变成了一个线性不可分的问题，于是就使用核技巧，通过将样本映射到高维特征空间使得样本线性可分，这样得到一个复杂模型，并由此导致过拟合（原样本空间得到的划分超平面会是弯弯曲曲的，它确实可以把所有样本都划分正确，但得到的模型只对训练集有效）。

**如何克服？**

为了解决上述问题，SVM通过引入松弛变量构造了软间隔分类器，它允许分类器对一些样本犯错，允许一些样本不满足硬间隔约束条件，这样做可以避免SVM分类器过拟合，于是也就避免了模型过于复杂，降低了模型对噪声点的敏感性，提升了模型的泛化性能。

因为松弛变量是非负的，因此样本的函数间隔可以比1小。函数间隔比1小的样本被叫做离群点，我们放弃了对离群点的精确分类，这对我们的分类器来说是种损失。但是放弃这些点也带来了好处，那就是超平面不必向这些点的方向移动，因而可以得到更大的几何间隔（在低维空间看来，分类边界也更平滑）。显然我们必须权衡这种损失和好处。

**松弛变量**

松弛变量表示样本离群的程度，松弛变量越大，离群越远，松弛变量为零，则样本没有离群。

**惩罚因子**

惩罚因子表示我们有多重视离群点带来的损失，当C取无穷大时，会迫使超平面将所有的样本都划分正确，这就退化成了硬间隔分类器。

惩罚因子是事先指定的值，对应于libsvm库的参数C。

每个松弛变量对应的惩罚因子可以不一样。

我们可以用惩罚因子来解决样本偏斜的问题。

**什么是样本偏斜，这会带来什么后果**

样本偏斜是指数据集中正负类样本数量不均，比如正类样本有10000个，负类样本只有100个，这就可能使得超平面被“推向”负类（因为负类数量少，分布得不够广），影响结果的准确性。

**如何解决**

可以通过为正负类样本设置不同的惩罚因子来解决样本偏斜的问题。具体做法是为负类设置大一点的惩罚因子，因为负类本来就少，不能再分错了，然后正负类的惩罚因子遵循一定的**比例**（如下文三种情况所述），具体值要通过实验确定。

**三种情况：**

（1）首先，如果是**因为采样得到的样本不够多样性或者数量不够**而导致的样本偏斜(分布范围不均)，可以通过衡量正负类分布的程度来确定惩罚因子的比例。比如计算它们在空间中占据的体积，例如给负类找一个超球（就是高维空间里的球），它要包含所有负类的样本，再给正类找一个，比较两个球的半径，就可以大致确定分布的情况。显然半径大的分布比较广，就给小一点的惩罚因子。

（2）但是，如果是**因为正负类本身性质所导致的分布范围不均**，此时即使超球半径差异很大，也不应该赋予两个类别不同的惩罚因子。例如文本分类中，某些话题涉及的面很窄，例如计算机类的文章就明显不如文化类的文章那么“天马行空”，这就会导致它们的超球半径差异很大，但这是由它们自身性质决定的，与样本多样性和数量多少无关，换言之，这个时候**不存在样本偏斜**的问题。

参考链接：  
<https://blog.csdn.net/icefire_tyh/article/details/52135662>

https://blog.csdn.net/Dby\_freedom/article/details/82960422

（3）但是，实际中我们不会这么做，因为实在太麻烦了，**当正负类数量不均时，比如正负类数量比为100：1，则惩罚因子的比例直接就定为1:100**（libsvm中就是这么做的）。

**总结**

SVM是一个使用核技巧的软间隔分类器，它的完整形态如前式所示。

若令惩罚因子C为无穷大，则SVM退化为硬间隔分类器，此时使用线性核只能处理线性可分的样本（因为对于线性不可分的样本，无法找到一个超平面可以正确划分所有样本，于是模型参数不存在可行解）；

若令惩罚因子C为适当值，则使用线性核的SVM可以处理**线性或近线性**的样本（这是因为此时允许分类器在某些样本上出错，所以对于近线性的样本，可以找到一个超平面将大部分样本正确划分）。

我们当然要令C为适当值，毕竟软间隔分类器对噪声没那么敏感。