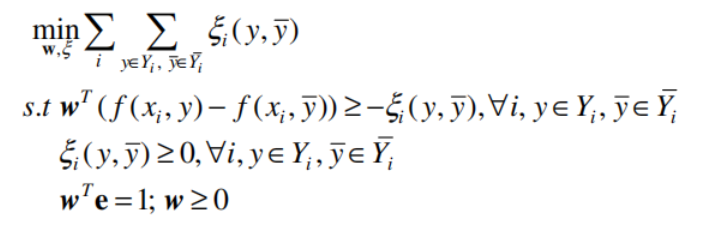
论文名：《基于多数据源排序和整合的计算药物重新定位》

看的这篇论文，感觉是可以做数据源整合技术的，但是有一点不同，这个方法是一个药物和一个疾病，计算预测评分，具体是一个药物dx与疾病s，药物dx的在某种数据源的相似性计算完成后，使用knn方法，将k个邻居选出，如果邻居中有可以治疗s疾病的，则进行计算整合，最后多种数据源计算后，再将多数据源整合。而我想找的是，可以将多个数据源相似性整合成一个相似性评判标准的方法，但是想自己尝试改进一下这个方法。

文中有个大边际法（最小化铰链损失）可以学习一下。主要是权重是如何计算的：

假设我们有m个数据源，{Dj, j = 1，，m}， n个药物{xi, i = 1，…，n}。每种药物从k适应症集合中分配到几种适应症。设Yi表示赋药xi的适应症集合，Yi表示赋药xi的适应症集合，yi表示不赋药xi的适应症集合。则设f(x,y)是一个长度为m的向量，其第j个元素为药物x在数据源Dj上的得分，表示y。通过求解下面的优化问题，得到用于m预测积分的权重向量w。



其中e是一个1向量。由此产生的凸优化问题可以使用标准的优化工具来解决，如CVX2。利用训练后的权重向量w，将不同数据源的药物指征评分加权平均为wT·f i (x,y)。

想着可以学习一下铰链损失，用权重分配，来整合多数据源。

在机器学习中，hinge loss是一种损失函数，它通常用于"maximum-margin"的分类任务中，如支持向量机。数学表达式为：

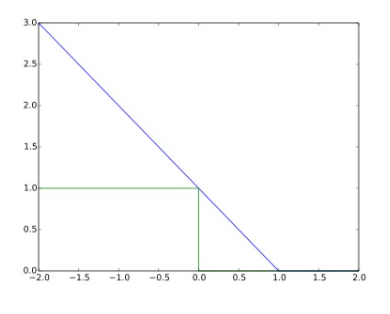


其中 表示预测输出，通常都是软结果（就是说输出不是0，1这种，可能是0.87。）， 表示正确的类别。

如果 ，则损失为： 

如果  ，则损失为：0

其函数图像如下，与0-1损失对比：



优化

以支持向量机为例，如果用hinge损失，其求导结果如下：



变种

实际应用中，一方面很多时候我们的y的值域并不是[-1,1]，比如我们可能更希望y更接近于一个概率，即其值域最好是[0,1]。另一方面，很多时候我们希望训练的是两个样本之间的相似关系，而非样本的整体分类，所以很多时候我们会用下面的公式：



其中，y是正样本的得分，y’是负样本的得分，m是margin（自己选一个数）

即我们希望正样本分数越高越好，负样本分数越低越好，但二者得分之差最多到m就足够了，差距增大并不会有任何奖励。

比如，我们想训练词向量，我们希望经常同时出现的词，他们的向量内积越大越好；不经常同时出现的词，他们的向量内积越小越好。则我们的hinge loss function可以是：



其中，w是当前正在处理的词， w+是w在文中前3个词和后3个词中的某一个词， w- 是随机选的一个词。