

线性代数应该这样讲(三)-向量2范数与模型泛化

原创 夕小瑶 夕小瑶的卖萌屋 2017-07-19



在[线性代数（一）](#)中，小夕主要讲解了映射与矩阵的关系；在[线性代数（二）](#)中，小夕讲解了映射视角下的特征值与特征向量的物理意义。本文与下一篇会较为透彻的解析一下向量的**二范数**与**一范数**，他们在机器学习任务中是最常用，有时甚至是核心的东西哦。



首先，来一个俗俗的开篇。向量x的p范数表示如下：

$$||x||_p = \left(\sum_i |x_i|^p \right)^{\frac{1}{p}}$$

由此，p=1、p=2就分别代表1范数和2范数。本文只看p=2的情况。

二范数相信大家在大一学线性代数的时候就已经被灌输了“**用来度量向量长度**”、“**用来度量向量空间中两个点的距离**”这两个典型意义，但是却鲜有学校讲过最小化二范数会带来什么有趣的现象，而这正是二范数在机器学习中非常重要的应用。

我们经常在机器学习的loss函数中加上参数的2范数项，以减少模型对训练集的过拟合，即提高模型的泛化能力。那么问题来了，2范数凭什么可以提高模型的泛化能力呢？使用参数2范数约束项一定好吗？

首先我们把model的参数设为向量 $w=[w_1, w_2, \dots, w_n]$ 。这个w是什么呢？是model参数，更是**特征的权重**。更加具体点说，每个参数，决定了每个特征对决定样本所属类别的重要程度。

那么用参数向量的二范数做正则项时即(忽略归一化的问题)：

$$Reg = \sqrt{w_1^2 + w_2^2 + w_3^2 + \dots + w_n^2}$$

而我们训练的过程就是最小化loss函数的过程，因此一部分任务就是尽可能的减小Reg的值。那么怎样的w值才算是符合理想要求的值呢？比如维度n=5的情况，

1. 令 $w_1=w_2=w_3=w_4=w_5=2$
2. 令 $w_1=10, w_2=w_3=w_4=w_5=0$

1和2相比，哪个的Reg更小呢？显然前者的值只有20，而后者的值高达100！虽然1和2的情况下所有w的值加起来都等于10。

由这个例子可以看出，如果我们有10张用于决定类别的票分给各个特征，那么给每个特征分两张票带来的回报要远大于把这10张票分给一个特征！所以**二范数会削弱强特征，增强弱特征**，以共产主义为目标！反对资本主义！（什么鬼

然后在上面这个前提下，尽量的降低票数（然而这不重要，一共有10张票跟一共有100张票相比没有影响，毕竟真正起作用的是票的分配方式（当然，这里没有考虑梯度饱和等优化问题哈

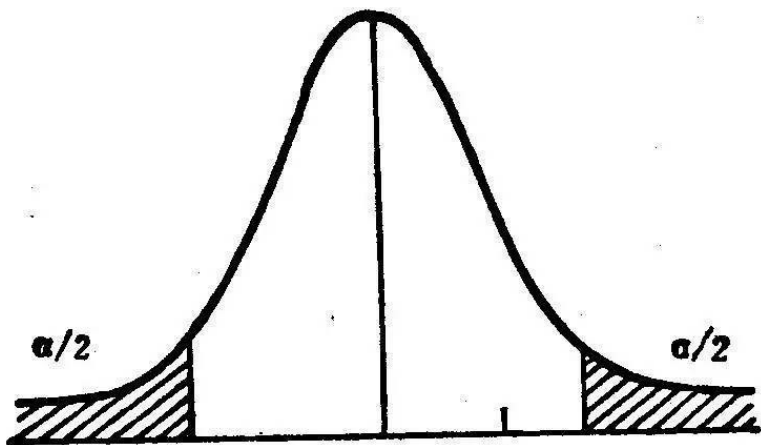
这种劫富济贫的方式有什么好处呢？



举个例子。假设我们要做文本的情感分类（判断一段文本是正面评价还是负面评价），将每个词作为一个特征（出现该词代表值为1，否则值为0）。

可想而知，有一些词本身就带了很多的情感极性，比如“不好”、“不满意”、“惊喜”等。而大部分词是弱极性的，但是多个弱极性的词同时出现的时候就会产生很强的情感极性。比如“总体”“来说”“还是”“可以”在文本中同时出现后基本就奠定了这篇文本的总体极性是正面的，哪怕文本中出现了（“待机”）“很烂”这种强负面词。

因此在二范数的约束下， w 这个随机变量的分布会趋向于方差 $\sigma=0$ 的**高斯分布**（正态分布）。

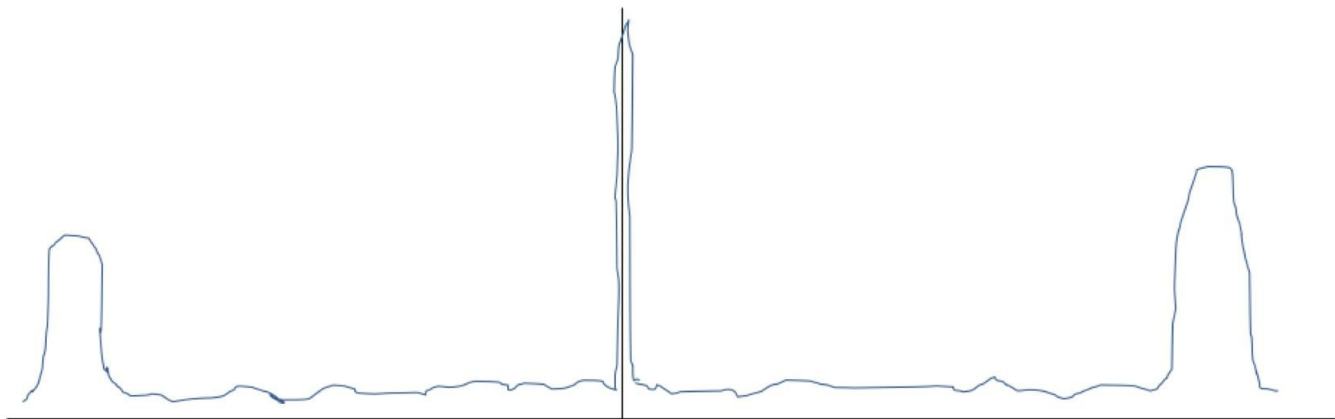


有人说，诶？那高斯分布的话，那也有极少的特征的值特别大呀~为什么没有被削弱呢？

这些特征当然就是超强特征啦，比如“力荐”这个特征一旦出现，基本整个文本的情感极性就确定了，其他的弱特征是很难与之对抗的。所以最小化二范数会让随机变量的采样点组成的向量趋向于期望 $=0$ 的高斯分布。



所以，若没有二范数的约束，弱特征会被强特征剥削，最终训练完后各个特征的权重很有可能是这样的：



（画的有点夸张啦，但是表达的意思是没错啦

这样会带来什么问题呢？这样就会导致模型过分依赖强特征。

首先，试想一下，这样的model拿到测试集上去后，一旦某个样本没有任何强特征，导致该样本的特征的权重几乎都是0，也就是这些特征都是被认为的中性词，那么就会导致这个样本的分类很随机了，哪怕这个样本的弱特征很多而且足以反映情感极性，然而学习的过程中这些弱特征被认为没有用而被当成了噪声，或者正值或者负值，那就悲剧啦。而在二范数约束带来的高斯分布下，弱特征们就会有条不紊的慢慢积累起很确定的情感极性完成置信度很高的分类过程。

再想一下，这样的model的抗噪声能力也会非常差，一旦测试集中的某个样本中出现了一个强特征词，就会直接导致整篇文章的情感极性随了这个强特征，哪怕这个样本的这个强特征词之外都是弱弱的相反极性词也无力挽救了。而在二范数约束带来的高斯分布下，手中多少也有点票的弱特征们就会聚沙成塔，合力打倒那个强特征的噪声。


那么是不是所有的机器学习任务加上二范数约束就一定好呢？




相信经过小夕上述的讲解，您心中已经有答案啦~在一些机器学习任务，尤其一些结构化数据挖掘任务和特征意义很模糊的机器学习任务（比如深度学习）中，特征分布本来就是就是若干强特征与噪声的组合，这时加上2范数约束反而会引入噪声，降低系统的抗噪性能，导致更差劲了。


因此，使用二范数去提高机器学习model的泛化能力大部分情况下是没错的，但是也不要无脑使用哦，懂得意义后学会根据任务去感性与理性的分析才是正解啦。

蟹蟹你o(≥v≤)o



小红包

 微信支付



Transfer to 夕小瑶