**机器学习阅读笔记**

蒋明鑫 17301064

**一、文献信息**

作者：John Canny

论文题目： Designing, Visualizing and Understanding

Deep Neural Networks

（设计、可视化和理解深层神经网络）

发表时间：CS 182/282A Spring 2019

**二、问题意义**

作者在前面的内容里面定义了模型损失的概念，它常常是用来衡量预测值$\hat{y}$和实际值y之间的误差的一个量，那么对于一个理想的预测模型，就一定要有最小化预测损失的工作。如果对于一个开放空间，由于自身的训练数据极为有限，不一定能对新数据进行合理的预测和判决，所以损失是在所难免的（此类情况，文中定义为风险）。而对于有限数据子集来说，又会产生一定的经验风险。所以在机器学习模型中，人们一般来说会寻求解决方案来最小化风险。

所以，为了减小预测模型损失，作者提出了两种预测方程$\hat{y}=f(x)$，一种是线性回归模型，该模型只能够拟合直线，因此具有较高的偏差；但是，由于运用于不同数据集的线性模型变化很小，所以它们往往具有比较低的方差。另外一种是逻辑回归模型，它是通过高次多项式拟合而成的复杂曲线，能够尽可能地去适应每一个数据点的变化，所以其有较小的偏差；但是由于多项式往往会产生对数据样本的过度拟合，往往会导致方差较大。所以，本文的工作之一也是对预测偏差和方差进行权衡，以实现最佳系统性能。

除此之外，作者还定义了线性分类器的裕度和合页损失函数，以实现一个支持向量机（SVM）分类器的最大边缘化。最后，作者用交互可视化验证了模型性能，并将交叉验证作为更准确的模型预测方法。

**三、思路方法及实验结论**

**1.偏差-方差的权衡以及正则化处理**

前面我们提到了损失的定义，其实偏差就是在不同训练数据集D上预测值与实际值的预期差异，定义为：

$$Bias\left(\hat{f}\left(x\right)\right)=E\left[\hat{f}\left(x\right)-f\left(x\right)\right]=\overbar{f}\left(x\right)-f(x)$$

所以，方差就是预测值的方差，定义为：

$$Variance\left(\hat{f}\left(x\right)\right)=E[\left(\hat{f}\left(x\right)-\overbar{f}\left(x\right)\right)^{2}]$$

因此，对于一个系统的平方损失问题，总平方误差可分解为偏差和方差，其流程如下：

$$E\left[\left(\hat{f}\left(x\right)-f\left(x\right)\right)^{2}\right]=E\left[\left(\hat{f}\left(x\right)-\overbar{f}\left(x\right)+\overbar{f}\left(x\right)-f\left(x\right)\right)^{2}\right]$$

$$=E\left[\left(\hat{f}\left(x\right)-\overbar{f}\left(x\right)\right)^{2}\right]+Bias^{2}-2E[(\hat{f}\left(x\right)-\overbar{f}\left(x\right))Bias]$$

上式最后一项，预测值减去期望预测值近似为0，所以预期的总误差可表示为：

$$=Variance\left(\hat{f}\left(x\right)\right)+Bias(\hat{f}\left(x\right))^{2}$$

我们可以看到，一个系统的总误差是来源于方差和偏差的共同贡献，所以我们对这两种贡献进行权衡便可以使总误差降至最低。

当$Variance\gg Bias^{2}$，其表现为过度拟合，如下右图；

当$Variance\ll Bias^{2}$，其表现为欠拟合，如下左图。



由于深度网络具有高维参数化，因此通常存在于右侧图。与其它模型相比，它们往往具有高方差和低偏差。为了解决这个问题，作者采用正则化处理来见皇帝模型的方差。

在这里，我们先定义一个$k×m$阶的矩阵A，其最小平方误差模型定义为：$\hat{A}=M\_{yx}M\_{xx}^{-1}$，而$M\_{xx}=\sum\_{i=1}^{n}x\_{i}x\_{i}^{T}$，$M\_{yx}=\sum\_{i=1}^{n}y\_{i}x\_{i}^{T}$，公式中的y是由线性回归y=Ax得到的。其次，我们定义$λ\_{i}$为矩阵$M\_{xx}$的特征值，在推到过程中我们发现，哪怕是很小的$λ\_{i}$也会造成模型的高方差，所以我们在损失之中添加了一个正则化项，如下：

$$L\left(A\right)=\sum\_{i=1}^{n}\left(x\_{i}^{T}A^{T}-y\_{i}^{T}\right)\left(Ax\_{i}-y\_{i}\right)+λ\sum\_{i,j}^{}A\_{ij}^{2}$$

如果我们对其求梯度，并令梯度为0，我们就可以得出一个新的模型$A=M\_{yx}(M\_{xx}+λI)^{-1}$，则此时$(M\_{xx}+λI)$对于的特征值即为$λ\_{1}+λ,…,λ\_{m}+λ$。相比与前面的特征值，此时的新模型的特征值变大了，从而在一定程度上减小了方差。

该正则化又称为L2正则化，在深度网络中有较为广泛的应用。同时，我们还可以通过调整$λ$的值来调整偏差与方差的权衡关系，如：

当$λ$较大时，称为强正则化，此时有较低的方差和较高的偏差；

当$λ$较小时，称为弱正则化，此时有较高的方差和较低的偏差。

**2.支持向量机**

假设我们要对一系列属于两类的二维观测的坐标点进行分类，我们可以在图中分别用黑点和白点表示，我们可以创建出线性分类器，产生三种不同的权重向量的决策边界对二维观测点进行分类，如下图：



可以发现，$H\_{1}$没有正确地进行分类，因而不予考虑。那么现在就有一个问题：究竟是$H\_{2}$的分类效果好呢，还是$H\_{3}$的分类效果好？接下来，我们便予以讨论。

直观上来看，$H\_{3}$具有较大的边距，所以功能可能分类出正确的观测点，所以我们便要对其模型损失进行分析。作者希望在保持最佳分类效果的同时，还可以尽可能地最大化分类器边界。作者定义了一个能够表示到判决边界距离的函数$f\left(x\right)=w^{T}x+b$，其中$w$为权向量且

其模值为1，接下来就可以创建$f\left(x\right)=1$和$f\left(x\right)=-1$的边距，并且定义：

$$f\left(x\right)\geq 1(x\in C)$$

$$f\left(x\right)\leq -1(x\notin C)$$

之后再对其进行标记：

$$y=\{\begin{matrix}-1&x\notin C\\1&x\in C\end{matrix}$$

然后我们便可以得到一个约束关系为：$yf(x)\geq 1$。

既然要衡量该系统分类能力的好坏，那么我们就需要一个标准来评判上述约束关系被违反的程度。所以，作者定义了合页损失函数。对于一系列观测点，合页损失函数就是每个观察到损失的总和，如下：

$$l=\sum\_{i=1}^{n}max⁡(0,1-y\_{i}f\left(x\_{i}\right))$$

显然，如果观测点分类正确，损失为0；如果观测点分类正确但却在距离判决边界为1的边界以内，损失为正值，并且会随着离判决边界的距离的增加而增加；当观察点分类错误，损失为正值。所以，合页损失函数可以衡量一个分类模型的性能，当其值越大，说明其判决性能越差，反之则越优。

**四、启发思考**

本文重点对偏差和方差进行了定义，并且讨论了两者之间的权衡关系，给出了欠拟合以及过度拟合所造成的误差影响。考虑到现有的逻辑回归模型都会伴随着较高方差的问题，作者是通过利用矩阵$A$的先验分布已知的条件下，在损失中添加了一个正则化项，增大了特征值，减小了方差。先验分布，顾名思义，就是“事前分布”，其与实验结果无关，仅仅是在统计实验之前通过数学推导和参数计算而得。所以，我们遇到问题时，首先要运用理论知识对问题进行简化，再结合实验论证。

此外，作者还定义了合并损失函数来衡量判决系统的性能。其间推导的重要一步就是运用了元素标记法，把对观测点分布的两个约束条件化为了一个，进而导出了损失函数的表达式。所以我们在分析中，也要尝试把多元条件化一，以便进一步处理。

文章中还简要地对介绍了多类支持向量机分类器，以及交叉验证测试模型性能的方法，文章的结构主要就是先提问，再分析，后对比，最后测试，思路较为清晰，也在一定程度上反映了我们解决问题所需要的一般思路。