SPSS神经网络

常见用途

1. 产生随机数来选择样本数据集；
   转换→随机数生成器→设置起点→固定值→值（此处需输入数值）→确定”
   <用于设定随机数种子，保证在以后生成的随机数一致>

   > 菜单“转换→计算变量→目标变量（此处需输入自定义变量名称）→数字表达式（输入“2*RV.BERNOULLI(0.7)-1”）→确定”

<用于产生bernoulli（伯努利）分布数列，数列名即为输入的自定义名，括号中的0.7是对样本数据进行划分的标准，如随机抽取数据的70%值为1,30%值为-1>

2. 生成多层感知器。
   https://blog.csdn.net/qq_51746700/article/details/121281475
   分析→神经网络→多层感知器→变量（分别选入因变量、因子和协变量，按照数据集情况选择“协变量重新标度”）
   →分区（有“根据个案的相对数目随机分配个案”和“使用分区变量来分配个案”两种选择）
   →体系结构（可选择“体系结构自动选择”或“定制体系结构（可以在“隐藏层数”中选择生成两层隐含层）”）多增一层可能会导致数据过拟合
   →训练（选择“训练类型”和 “优化算法”）
   →输出（选择需要输出的结果显示）→保存/导出/选项→确定”

多层感知机的SPSS操作

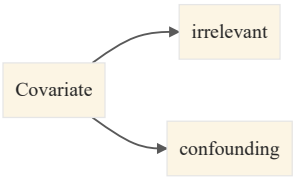
https://www.bilibili.com/read/cv11986460

Multilayer Perceptron，缩写MLP
前向结构的人工神经网络 ，映射一组输入向量到一组输出向量。 .
MLP可以被看作是一个有向图，由多个的节点层所组成，每一层都全连接到下一层。
除了输入节点，每个节点都是一个带有非线性激活函数的神经元（或称处理单元）。
一种被称为 反向传播算法 的 监督学习 方法常被用来训练MLP。
多层感知器遵循人类神经系统原理，学习并进行数据预测。
它首先学习，然后使用权重存储数据，并使用算法来调整权重并减少训练过程中的偏差，即实际值和预测值之间的误差。

选择

因子：分类自变量；
协变量：连续型自变量。
因子：一般有分类变量。协变量一般有尺度变量或者连续的自变量。

1．因变量，即用于检验影响是否显著的变量。多方差因素分析只选择一个因变量。自变量是指在实验中由实验者操作的变量，它被认为不会受其他变量的影响（即独立性）。
2．固定因子，即用于检验是否有显著影响的因素变量。
3．协变量，与因变量存在相关关系的变量。实验中除自变量以外的影响实验变化和结果的潜在因素或条件，但并非实验所感兴趣的变量。协变量对主要变量分析最重要的影响，指的是它往往会造成实验性研究干预措施疗效评价的偏倚。协变量，经常成为混杂变量。



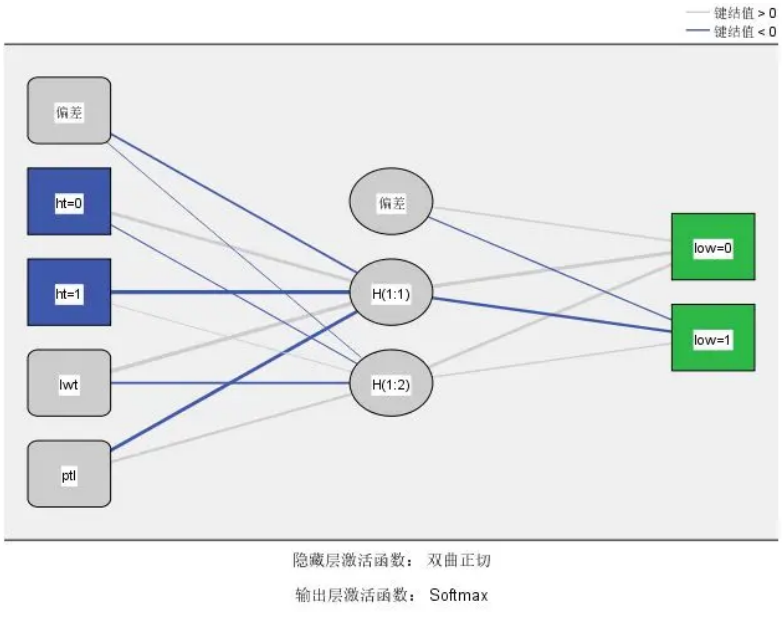
> 如果您有离散变量而且想要将其包括在回归或方差分析模型中，可以决定是将其视为连续预测变量（协变量），还是类别变量（因子）。如果离散变量具有许多水平，那么最好将其视为连续变量。将预测变量视为连续变量意味着简单线性或多项式函数足以描述响应和预测变量之间的关系。当您将预测变量视为类别变量时，离散响应值将与变量的每个水平拟合，而不必考虑预测变量水平的顺序。

结果分析

个案处理摘要

		N	百分比
样本	训练	132	69.8%
	检验	57	30.2%
有效		189	100.0%
排除		0	
总计		189	

为了防止过度拟合，在神经网络中需要对样本进行拆分，一般按照7：3或者4:3:3的比例，随机形成训练集、验证集和支持集。这里采用SPSS默认的7：3。



神经网络的结果示意图
因子“ht”分为两个哑变量（0或者1的人工变量）节点
ht=0（没有高血压）,ht=1（患有高血压）

相应的两个分类的因变量也以low=0和low=1两个哑变量输出。
协变量“lwt”和“ptl”各自以一个节点方式纳入模型。

在相邻的两层中，以两种颜色区分连接权重的正负，连接线的粗细代表权重绝对值的大小
从图中可以看出，自变量ht对模型的贡献较大，且输入层的ht=1的节点通过隐藏层H(1:1)节点与输出层low=1节点有较强的连接权重，这表示有高血压的孕妇更容易生出低体重婴儿。

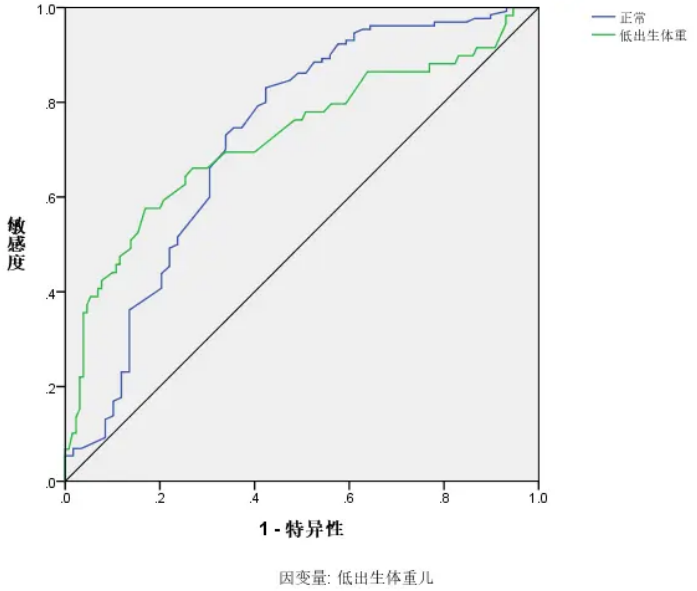
相对错误
都在25%-30%之间，等于说，预测准确率在70%-75%。
训练集和测试集对出生婴儿低体重的预测准确率为29.5%和26.7%。

ROC曲线

Receiver Operating Characteristic Curve，中文名字叫“受试者工作特征曲线”
该曲线的横坐标为假阳性率（False Positive Rate，FPR），N是真实负样本的个数， FP是N个负样本中被分类器预测为正样本的个数。纵坐标为真阳性率（True Positive Rate，TPR）
TPR = TP/P
AUC（Area under roc Curve）面积指ROC曲线下的面积大小.AUC的取值一般在0.5~1之间。AUC的值越大，说明该模型的性能越好。
反映模型在选取不同阈值的时候其敏感性（sensitivity，FPR）和其精确性（specificity，TPR）的趋势走向

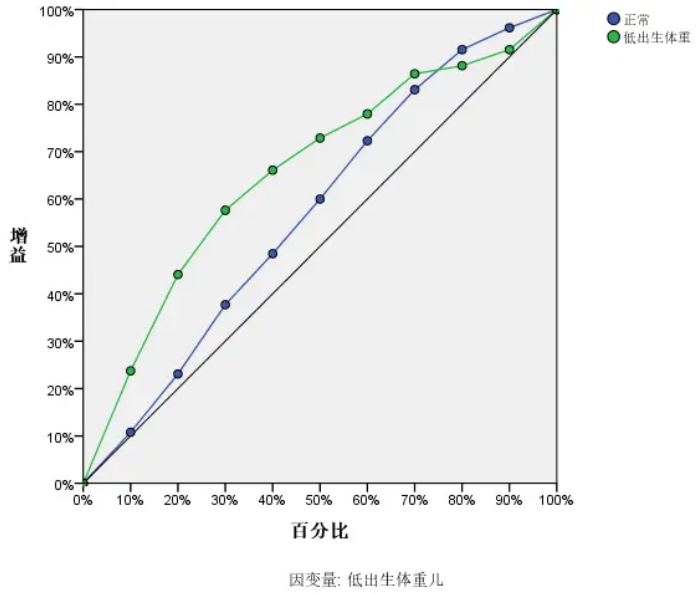
很大的优势

当正负样本的分布发生变化时，其形状能够基本保持不变，而P-R曲线的形状一般会发生剧烈的变化
能降低不同测试集带来的干扰，更加客观的衡量模型本身的性能



分别针对正常和低出生提供的两个类别的ROC曲线

增益累积曲线



因变量: 低出生体重儿

5横轴代表进入预测的个案比例
纵轴代表某类别中已被正确预测的样本占该类别所有被正确预测样本的比例。
基线（斜45度直线）代表随机选择得到的结果，模型累积增益线代表使用模型之后的预测结果。图中累积增益线从一开始就明显高于基线，在某一点之后逐渐开始靠近基线并且最终重叠。
若累积增益图从左到右开始阶段越陡峭，而且下面所包围的面积越大，则模型的效果越好。

镜像基函数

在两者的操作界面中，除了“体系结构”选项不同之外，其余选项卡的内容完全相同。

体系结构

多层感知器的“体系结构”选项用来指定神经网络结构，一般软件程序会自动选择最好的结构，不需要手动修改。
径向基函数中“体系结构”是用来建立一个隐藏的径向基函数层。
两个神经网络模型中“体系结构”的各参数，绝大多数时候采用自动搜索确定最佳的单位数和最佳允许重叠数量即可，因此很多时候也不需要修改。
如果在做神经网络分析时，特别强调需要修改隐含层或者输出层激活函数类型时，才会手动修改.

相同的分析方式，比较最终数值（训练集和测试集预测正确率，整体预测错误率）得出更适合什么样的预测结果