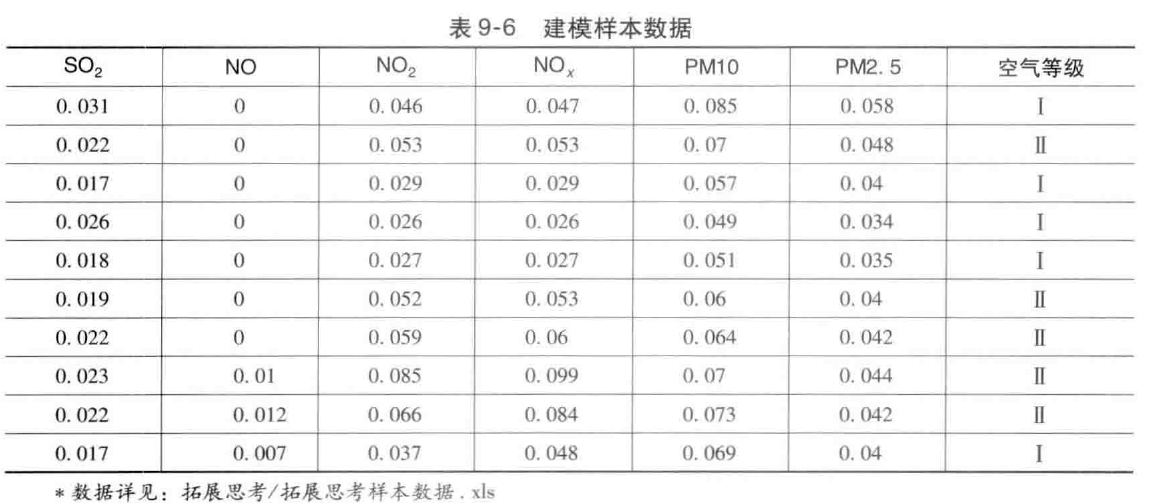
有人根据空气中SO2、NO、NO2、NOx、PM10和PM2.5的含量，建立分类预测模型，对空气质量进行评价。在某地实际监测的部分原始样本数据经预处理后如表9-6所示（具体数据见附件“拓展思考样本数据.csv”）。请采用适当方法进行模型构建，并评价模型效果。



下课前，把所建模型及结果保存后发给老师查看。

文件命名格式：45\_戴子博\_2220100519

所得结果或者图片的标题部分请加上考勤标号及姓名，方便老师登记成绩。

例如：图1 决策树示例（45\_戴子博）

首先，我们对得到的数据进行了详实的描述性统计

为明确数据在分布上的特征，我们对6种特征的样本分布进行了统计：

# 创建直方图

p1 <- ggplot(data\_raw, aes(*x* = SO2)) + geom\_histogram(*bins* = 20, *fill* = "blue", *color* = "black") +

  theme\_minimal() + ggtitle("SO2 Distribution")

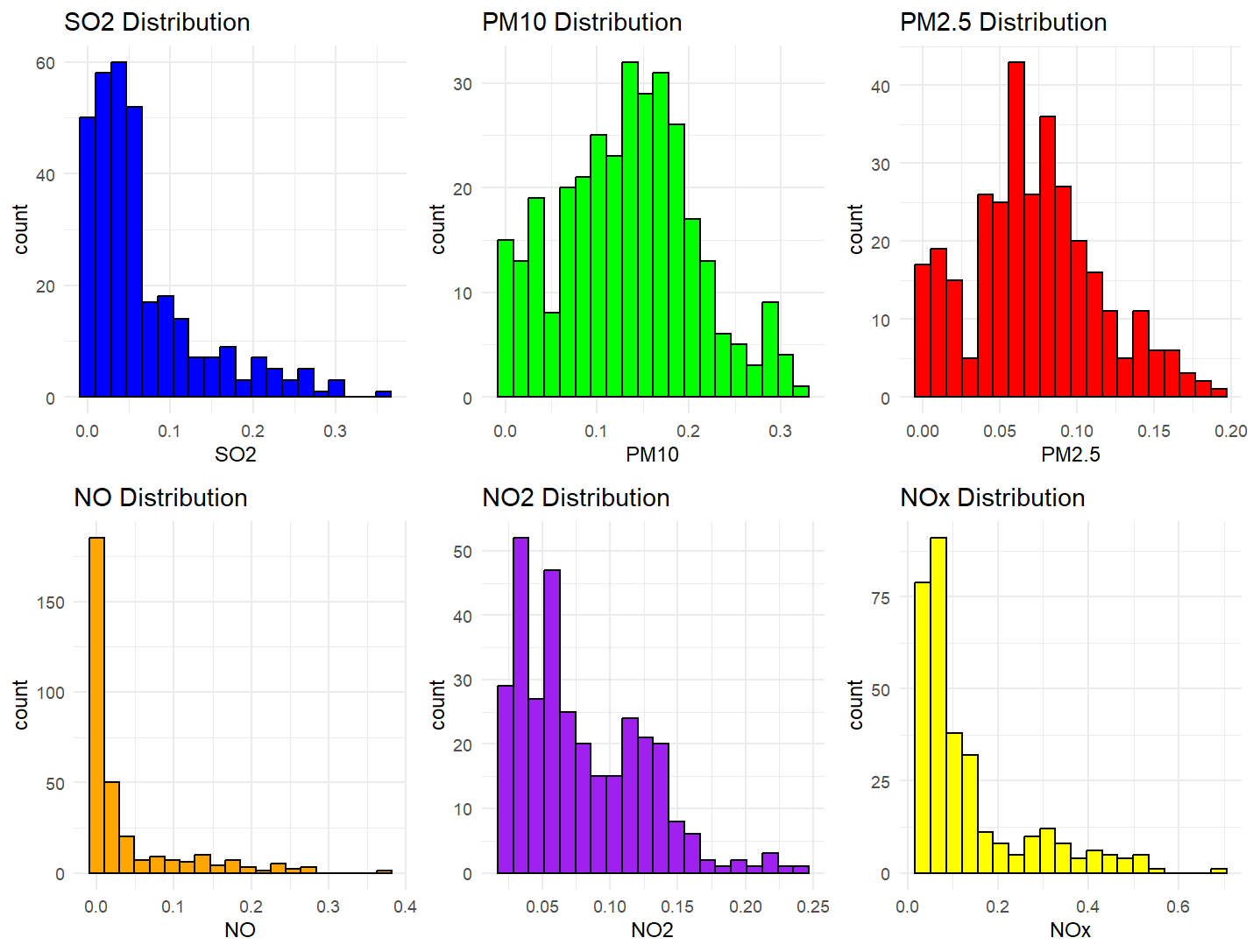
p2 <- ggplot(data\_raw, aes(*x* = PM10)) + geom\_histogram(*bins* = 20, *fill* = "green", *color* = "black") +

  theme\_minimal() + ggtitle("PM10 Distribution")

p3 <- ggplot(data\_raw, aes(*x* = `PM2.5`)) + geom\_histogram(*bins* = 20, *fill* = "red", *color* = "black") +

  theme\_minimal() + ggtitle("PM2.5 Distribution")

p4 <- ggplot(data\_raw, aes(*x* = NO)) + geom\_histogram(*bins* = 20, *fill* = "orange", *color* = "black") +

  theme\_minimal() + ggtitle("NO Distribution")

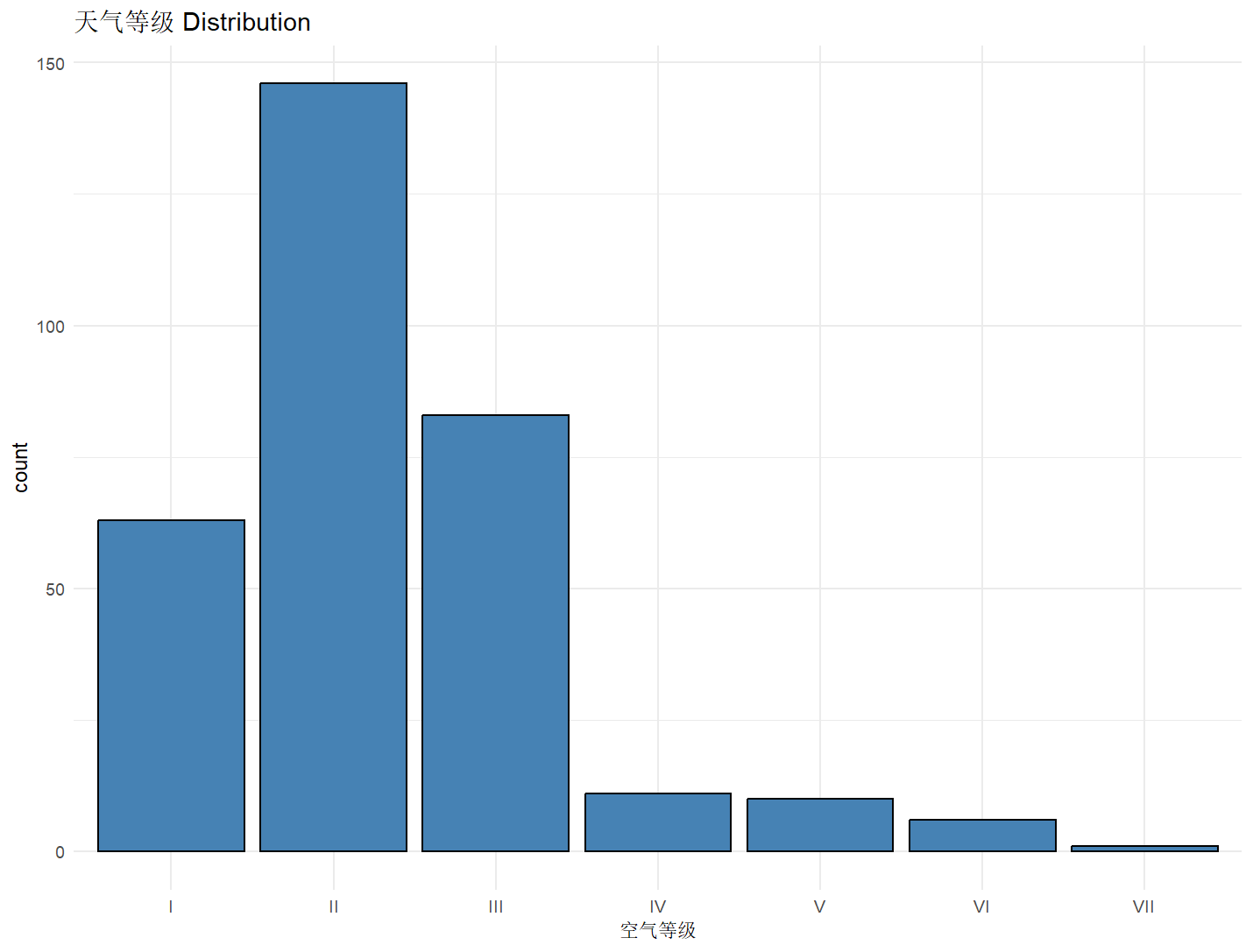
p5 <- ggplot(data\_raw, aes(*x* = NO2)) + geom\_histogram(*bins* = 20, *fill* = "purple", *color* = "black") +

  theme\_minimal() + ggtitle("NO2 Distribution")

p6 <- ggplot(data\_raw, aes(*x* = Nox)) + geom\_histogram(*bins* = 20, *fill* = "yellow", *color* = "black") +

  theme\_minimal() + ggtitle("NOx Distribution")

同时也对样本的标签分布进行了统计：



面对不均衡的样本，在机器学习任务上，我们首先使用全连接神经网络对所有特征变量直接进行建模：

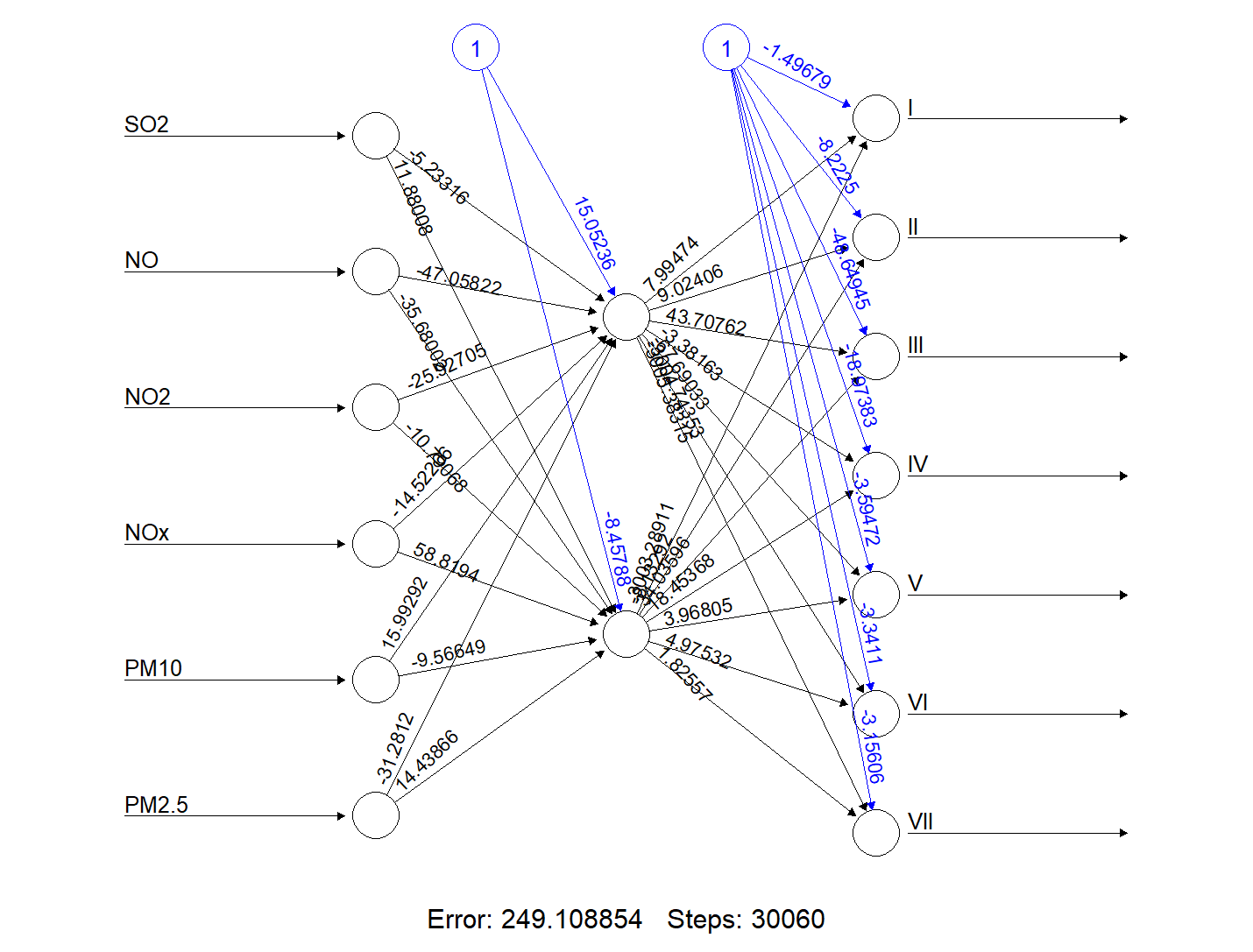
library("neuralnet")

set.seed(12345)

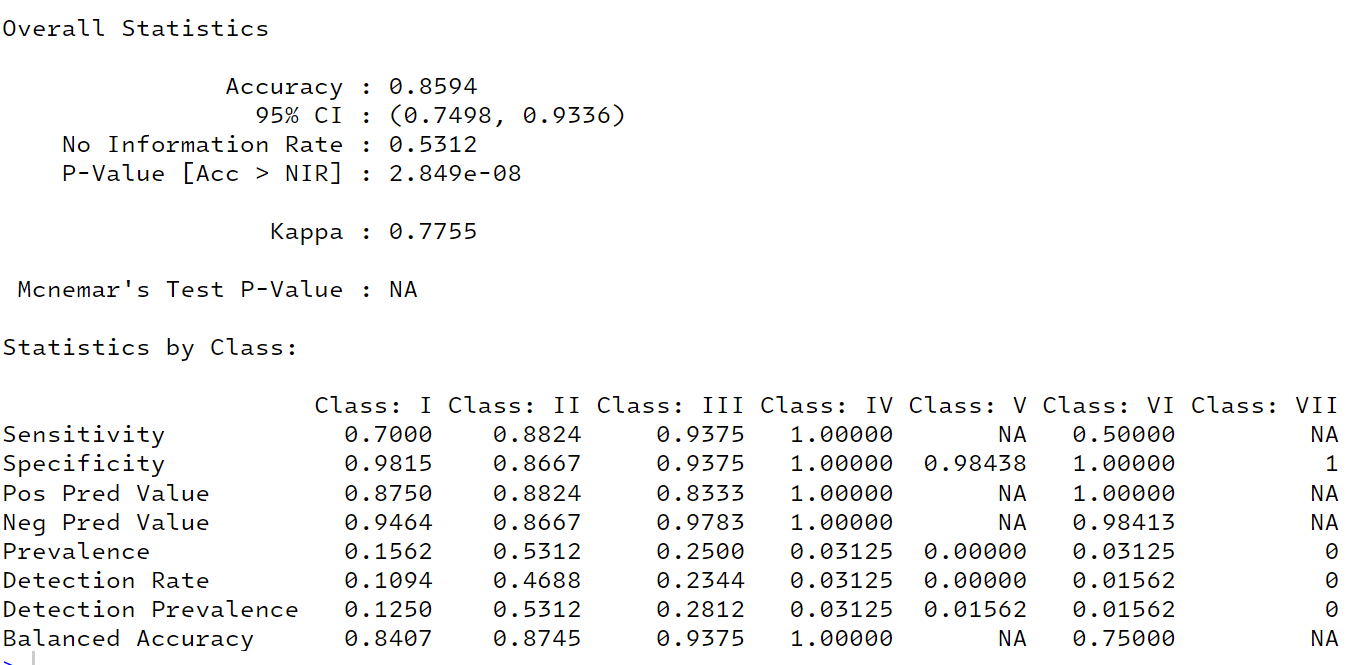
(BPnet1<-neuralnet(`空气等级`~.,*data*=data\_raw,*hidden*=2,*err.fct*="ce",*linear.output*=FALSE))

plot(BPnet1)

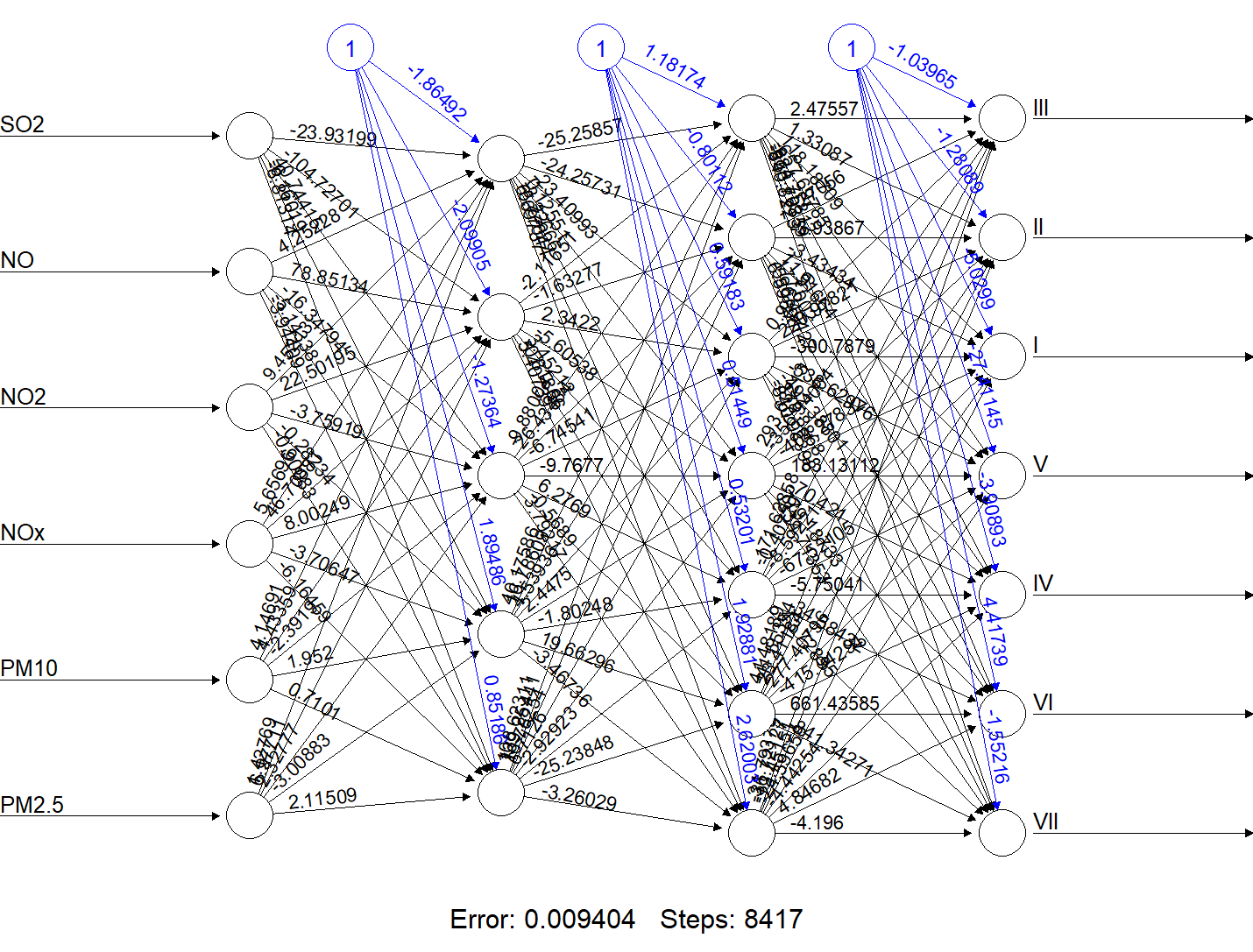
得到了如下的神经网络：



使用混淆矩阵对神经网络模型的拟合效果进行评估，我们得到了以下结果，准确率为85.94%，但模型的数值敏感性似乎还有待提升：



经过进一步的超参数优化，我们得到了一个相对更有效的双隐层神经网络：



从混淆句真和准确率来看，模型具有较小的扰动敏感性以及较好的泛化能力，在验证集上的准确率达到了89.06%。

