本周主要工作：阅读和了解自我学习

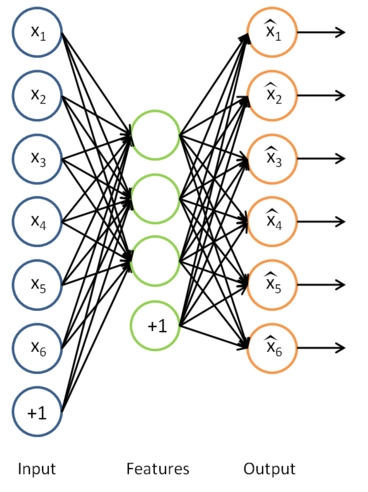
自我学习

自我学习的概念：

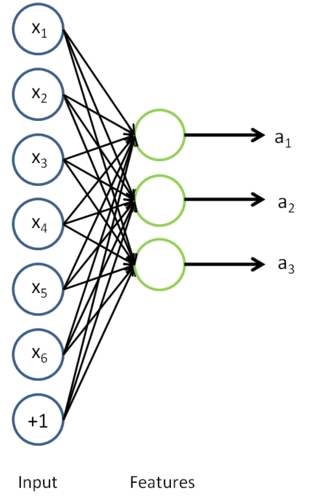
在现实生活中存在着大量的无标注的数据，即数据中只存在着特征属性，而没有类别标签，要想有效地利用这些无标注的数据，一种办法是人工标注这些数据，然而，人工标注这些数据需要花费大量的人力和物力。另一种方法是利用[算法](http://lib.csdn.net/base/datastructure" \o "算法与数据结构知识库" \t "http://blog.csdn.net/google19890102/article/details/_blank)对这些无标注的数据进行学习，学习出更好的特征描述，自我学习便是这样一个概念，即从大量的无标注的数据中学习到特征的更好表示。

特征学习：

在前面学习了使用自编码器学习到特征的表示方法，具体过程为，对于m无标注的训练样本{x(1),x(2),⋯,x(m)}，可以学到如下的稀疏自编码器模型：



其中，稀疏自编码器的模型参数为：W(1),b(1),W(2),b(2)。对于稀疏自编码器模型来说，隐含层代表的是特征的另一种表示，可能是一种更好的表示，如下图所示：



假设有ml个标注的样本{(x(1),y(1)),(x(2),y(2)),⋯,(x(m1),y(ml))}，利用稀疏自编码模型得到更好的特征表示。如对于输入x(1)，经过稀疏编码器得到其特征表示为a(1)，通常有两种方法处理这两种特征表示：

“替代表示(Replacement Representation)”，即直接将a(1)替代x(1)。

“级联表示(Concatenation Representation)”，即将上述两种特征组合在一起，使用新向量(x(1),a(1))代替原先的特征x(1)。

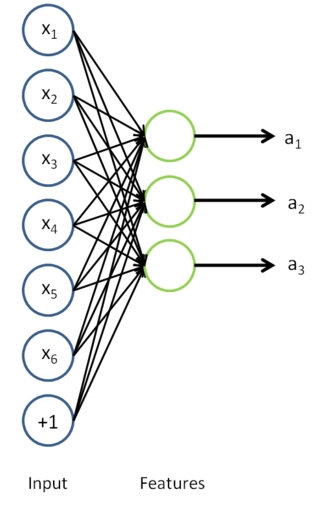
经变换后的特征变成了替代表示{(a(1),y(1)),(a(2),y(2)),⋯,(a(m1),y(ml))}或者是级联表示{((x(1),a(1)),y(1)),((x(2),a(2)),y(2)),⋯,((x(m1),a(m1)),y(ml))}。

通常采用级联表示可以或得更好的效果。

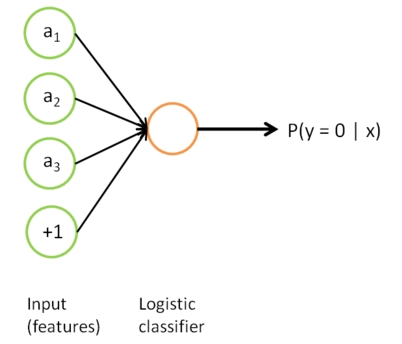
对于上述的特征表示，可以对其使用常用的分类器，如SVM，Logistic回归等等对其进行学习和预测。

从自我学习到深层网络：

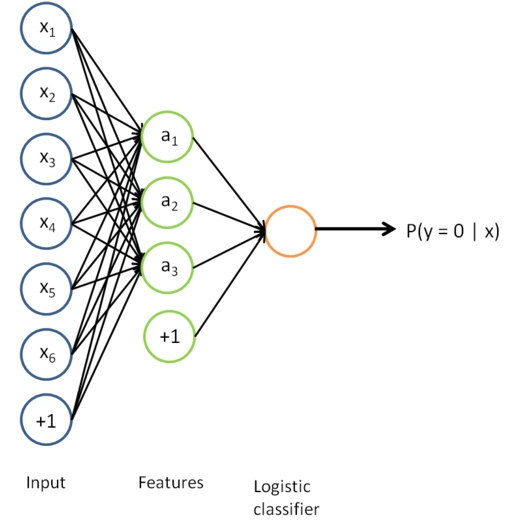
在自我学习的过程中，首先通过未标注的数据训练一个稀疏自编码器，即对于给定的未标注的样本{x(1),x(2),⋯,x(m)}，训练过程如下图所示：



对于一个新的样本x，我们可以通过训练好的稀疏自编码器提取出其特征a。然而对于一个分类问题，我们关心的是样本的标签y，对于有标注的样本{(x(1),y(1)),(x(2),y(2)),⋯,(x(m1),y(ml))}，利用训练好的稀疏自编码器得到新的特征a，利用替代表示，得到新的数据集表示形式{(a(1),y(1)),(a(2),y(2)),⋯,(a(m1),y(ml))}，我们需要训练的是从特征a到标签y的映射，可以使用logistic回归分类器，其具体形式如下：



将上述两个过程结合在一起，就可以表示为如下的图：



该模型首先通过一些未标注的样本训练出稀疏自编码器的权重W(1),b(1)，然后利用对标注的样本，利用稀疏自编码器对其进行特征学习，学习到新的特征表示a，再利用logistic回归得到其第二层的权重和偏置W(2),b(2)。

技巧：

在上述得到初步的模型之后可以通过微调，通常是在有大量的标注数据的条件下利用梯度下降法对整个模型进行微调。

参考资料：

[1]斯坦福大学深度学习教程中文版

[2] CSDN博主zhiyong\_will的自我学习文章