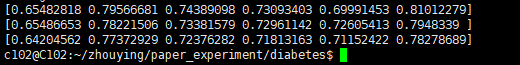
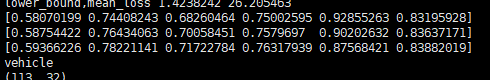
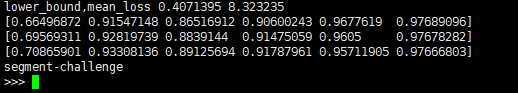
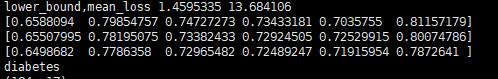
Diabetes 10次10折交叉验证

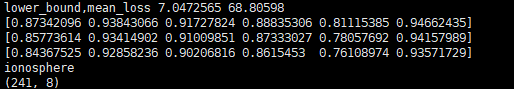


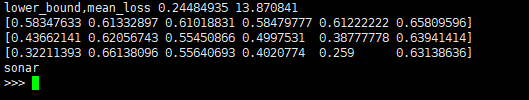








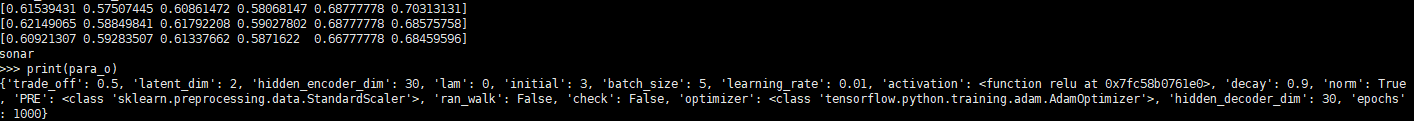


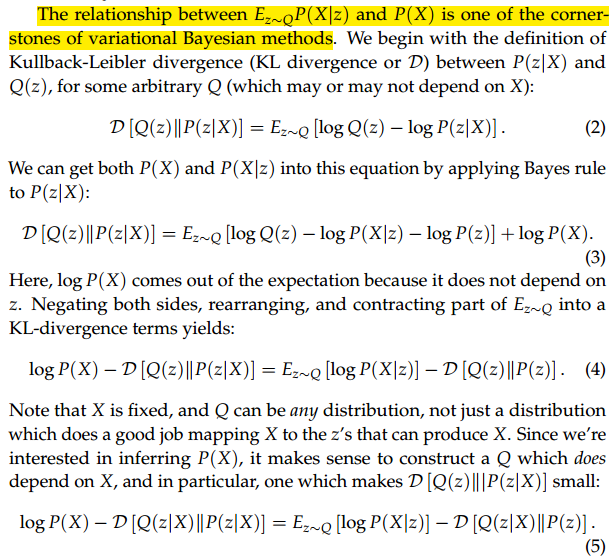


breastw  
最终版

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| 0.94229402 | 0.9677326 | 0.95863004 | 0.96044549 | 0.967 | 0.98552355 |
| 0.94997913 | 0.97224899 | 0.96432421 | 0.96476382 | 0.967 | 0.98579167 |
| 0.94997913 | 0.97224899 | 0.96432421 | 0.96476382 | 0.967 | 0.98727899 |

Sonar 训练过程





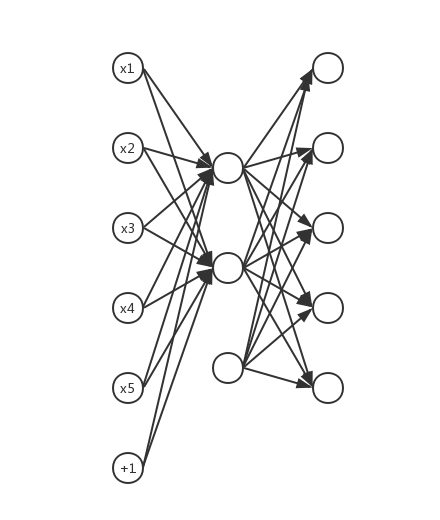
从等式（5）中可以看到，只要趋于0，则达到了我们最大化P（x）的目的。

但是我们对潜在变量的形式和分布都不清楚，需要利用神经网络的拟合能力对其真正的分布进行拟合，因此vae的网络结构需要多隐藏层，以对z的真实分布进行拟合。

利用神经网络生成样本，假设样本的最后表现形式由某些潜在的变量决定，设其为z，并对其先验分布和后验分布用神经网络进行模拟。基于神经网络的生成模型的优势在于其输出维度任意，因此几乎可以生成任意维度的数据；同时，由于神经网络强大的拟合能力，理论上神经网络可以将标准高斯分布映射成任意分布的形式，而不需要提前设定未知分布的概率分布函数。

3.12.2018

自编码器[13]用于对数据的降维或者特征学习，网络架构如图：



自编码器由两部分组成，编码部分和解码部分，假设编码部分构成函数f(x)，而解码部分构成函数g(x)，隐层空间为z，则自编码器的目标函数为

自编码器中的隐层变量z可以认为是原始空间中变量的压缩编码，因为z为实数范围中的任意数字，根据实数的无穷性，理论上该类形式的编码可以存储无限信息。但这些信息在原始的自编码器中，都是由原始的X决定的，无法随意生成，为了能够生成原始样本中没有出现的样本，

在合成样本过程中，因为传统的线性插值方法并未考虑到特征的现实意义，而对已有的特征值直接进行插值，SMOTE的公式如下：

在这种计算下，如果现实意义中的特征为离散型特征，例如性别等，线性插值会产生无意义的合成样本，因此论文[14]中在合成样本时，对该类特征进行筛选，采取轮盘赌或者random walk的形式确定合成样本的离散特征。

分类问题是机器学习中非常重要的一部分，在传统的分类问题中，我们都假设不同类别的样本是均匀分布的，因而每个样本的分类代价是一致的。但是在现实的数据集中，一方面我们对不同类别的感兴趣程度不一样，另一方面数据的分布也不相同，因而出现了不平衡分类问题：即在分类问题中，由于不同类别的样本数量和分布不同，不同的类别样本的分类代价不一致，而导致的对感兴趣的样本识别率过低。在这类问题中，传统的机器学习方法中，假设每类样本是均匀分布且分类代价一致，分类器的目标是追求全局的准确率，但是在现实的数据中，少数类数量过少，为了全局准确率，传统分类器容易忽视少数类样本的识别，造成少数类识别率很低。

从不同的过采样算法的对比结果来看，本文中所提出的过采样算法能够产生更加合理的样本，和以往的过采样算法会牺牲多数类的识别率不同，本文所提出的算法能够帮助分类器产生更合理的分类结果，提高整体的识别效果。

英文报告

在这个研究中，我提出了一种新的基于分布的过采样方法，利用变分自编码器对少数类样本的分布进行建模，并对该模型随机采样，生成更加合理的合成样本，实验结果证明了该算法的有效性；同时，还提出了一种新的数据集不平衡度的衡量指标，从样本分类难度的不一致考虑数据集的不平衡程度，改进传统的基于样本集合大小的不平衡率计算方法，该指标与最终的分类结果具有高度的线性相关性，利用该指标对前一步骤中随机采样的样本进行筛选，提高最终的少数类分类准确度。