机器学习 期末作业

王凯祺 16337233

第一题：怎么理解机器学习模型的泛化能力（Generalization Capability）？如何提升分类模型的泛化能力？

学得模型适用于新样本的能力，称为泛化能力。具有强泛化能力的模型能很好地适用于整个样本空间。尽管训练集只有样本空间的一个很小的采样，我们仍希望它能很好地反映出样本空间的特性。

要提升分类模型的泛化能力，我们可以从以下几个方面考虑：

一、增加训练样本。一般而言，训练样本越多，我们得到的关于未知分布 D 的信息越多，这样就越有可能通过学习获得具有强泛化能力的模型。

二、对数据做缩放。将数据缩放倒激活函数的阈值范围，比如归一化到 0 ~ 1，或者归一化到-1 ~ 1。

三、用增强学习的方法来优化。比如使用 AdaBoost 来提高未被准确分类的样本权重，从而达到更好的分类效果。

第二题：请阐述在聚类任务当中，聚类算法重要，还是样本点间距离（相似度）定义更重要？

我认为聚类算法更重要。

衡量样本点间距离时，一般有这么几种常见的距离公式：欧式距离$d(x,z) = \left||x - z|\right| = \sqrt{\sum_{d=1}^{D}(x_d - z_d)^2}$、曼哈顿距离$d(x,z) = \sum_{d=1}^{D}|x_d - z_d|$、以及核函数映射后的距离 $d(x,z) = ||\phi(x) - \phi(z)||$。在多数情况下，欧式距离与曼哈顿距离是正相关的。因此，以它们之中的任意距离定义对聚类结果来说，显得不那么重要。另外，在实际操作中，我们会更偏向于选择易于求导（求梯度）的距离定义，而不是效果更好的距离定义。

而在聚类算法上，用不同的聚类算法可能会导致聚类的结果完全不一样。每种聚类算法都有它的优点和缺点。例如：K-means 算法的优点是计算复杂度低、原理简单、容易解释，缺点是对噪声敏感、引入超参数 K、聚类结果依赖于分类中心的初始化；高斯混合模型的优点是投影后的样本点能得到每个类的概率，缺点是计算量大、聚类结果依赖于初始值的选取。对于特定的数据，我们应该对数据进行分析，然后结合每种算法的优缺点，选择最适于这种数据的聚类算法。在选定算法后，我们再根据求梯度的难易程度、效果来决定距离定义。

第三题：在文本分类、问答和翻译等 NLP 任务中，分析、比较你学过的和已知的文本建模算法。

解决 NLP 问题一般可以采用 CNN（Convolutional Neural Networks）、RNN（Recurrent Neural Networks）、LSTM（Long Short-Term Memory）这些神经网络模型。

CNN 能够从数据中自动学习特征，从而代替手工设计的特征，且深层的结构使它具有很强的表达能力和学习能力，擅长空间特征的学习和捕获。CNN 的运行速度快，适合做短文本分类。

RNN 将状态在自身网络中循环传递，因此可以接受更广泛的时间序列结构输入。它能

处理前后有关系的输入，也就是说，它不会孤立地理解这句话的每个词，而是处理这些词连接起来的整个序列。它更适合用于做翻译、生成文本等任务。

LSTM 适合于处理和预测时间序列中间隔和延迟非常长的重要事件。单纯的 RNN 因为无法处理随着递归，权重指数级爆炸或梯度消失问题，难以捕捉长期时间关联；而结合不同的 LSTM 可以很好解决这个问题。具体到语言处理任务中，LSTM 非常适合用于处理与时间序列高度相关的问题，例如机器翻译、对话生成、编码解码等。