Homework6

司可 经济系 15320171151903

特征工程与特征学习

在神经网络中，隐藏层中的神经元可以被看作是从数据中学习到的特征，神经网络可以被看作是支持特征学习或表示学习的。相比之下，在“经典”机器学习中，特征是由人类专家手工构建的——这一实践称为特征工程。

机器学习算法的成功通常依赖于数据表示，我们假设这是因为不同的表示可以或多或少地混乱和隐藏数据背后不同的解释变量，学习数据的表示形式，以便在构建分类器或其他预测器时更容易提取有用信息。

在机器学习的过程中，选择一组具有代表性的特征来构建模型和算法是非常重要的问题。特征选择通常选择与类别相关性强、且特征彼此间相关性弱的特征子集。然而在现实世界实际操作中，数据通常是非常复杂多变的，因此，我们可能面临着要从原始数据中来发现有用的特性。通过人工选取出来的特征依赖特定的人力和专业知识，不利于推广。于是我们需要通过机器来学习和抽取特征。

特征学习可以分为监督特征学习和非监督特征学习，监督特征学习包括监督字典学习、神经网络、多层感知机。监督字典学习是从输入数据中学习一组代表元素的字典，其中每个数据都可以表示为代表元素的加权和。通过最小化带有 L1 正则项的平均误差来确定字典元素和权重，并保证权重稀疏。监督字典学习利用输入数据和标签的隐含结构来优化字典元素。

神经网络是用来描述一系列学习算法，通过相互关联的节点构成的多层网络。它是受神经系统的启发，其中节点可以看做是神经元，边可以看成是突触。每个边都有相对应的权重，网络定义了计算规则，将数据从输入层传递到输出层。多层神经网络可以用来进行特征学习，因为它们可以学习在隐藏层中的输出的表示。

在一个二元分类问题中，令 $y \in \{-1,1\}$，$x = (1,x_1,\ldots,x_p)$，感知器（Perceptron）是一个能最小化错误分类的损失 $l(y, h(x)) = I(y \neq h(x))$的线性分类器 $h(x) = \text{sign}(w'x)$。其中 $w = (w_0, w_1, \ldots, w_p)$是权重。多层感知机（multilayer perceptron ；MLP)与简单的感知器相比，具有更多的层。额外的层称为隐藏层。任何可以分解成线性分割器的目标函数 f 都可以用三层 MLP 实现。

特征工程是利用数据的领域知识来创建使机器学习算法有效的特征的过程。特征工程在机器学习中占有相当重要的地位。在实际应用中，特征工程是机器学习成功的关键。那么下什么是特征工程呢？特征工程要做的事，就是为预测模型获更好的训练数据。简而言之，特征工程就是一个把原始数据转变成要选取的特征的过程，这些特征可以最好的描述

这些数据，并且利用它们建立的模型的表现性能可以接近最优。选择不同的数据特征，会直接影响我们模型的预测性能。选择适合的特征，能更好的进行预测。

特征工程的过程主要包括：特征选择（Feature Selection）、特征提取（Feature Extraction）和特征构造（Feature construction）。

特征选择的目标是寻找最优特征子集。特征选择能剔除不相关或多余的特征，从而达到减少特征个数，提高模型精确度，减少运行时间的目的。另一方面，选取出真正相关的特征简化模型，协助理解数据产生的过程。它的目的是从特征集合中挑选一组最具统计意义的特征子集，从而达到降维的效果。特征选择的过程包括：产生过程、评价函数、停止准则、验证过程。产生过程是按照一定的搜索策略产生候选特征子集；评价函数是子集评估，通过某个评价函数评估特征子集的优劣；停止准则是停止条件，决定特征选择算法什么时候停止；验证过程是用于验证最终所选的特征子集的有效性。

特征提取应该在特征选择之前。特征提取的对象是原始数据，它的目的是自动地构建新的特征，将原始特征转换为一组具有明显物理意义或统计意义或核的特征。比如通过变换特征取值来减少原始数据中某个特征的取值个数等。对于表格数据，你可以在你设计的特征矩阵上使用主要成分分析、独立成分分析、线性判别分析等来进行特征提取从而创建新的特征。对于图像数据，可能还包括了线或边缘检测。

在实际应用中，需要我们手工去构建特征。特征构造就是从原始数据中人工的构建新的特征。我们需要人工的创建它们。需要我们根据真实的数据样本的数据结构和预测模型，从原始数据中找出一些具有物理意义的特征。

对于数据挖掘和处理类的问题，使用一般的机器学习方法，需要提前做大量的特征工程工作，而且特征工程的好坏会在很大程度上决定最后预测结果的优劣。

然而在深度学习中，这一过程在某种程度上可以自动完成。深度学习（Deep Learning）模型是具有许多隐含层的神经网络。这允许构建层次特性，并支持使用与多个抽象级别对应的多个级别的表示学习。深度学习是机器学习领域之一，可以实现端到端的监督学习和非监督学习。

使用深度学习的话，特征工程相比之下就没那么重要了，特征只需要做些预处理就可以了，因为它可以自动完成传统机器学习算法中需要特征工程才能实现的任务，特别是在图像和声音数据的处理中更是如此。深度学习让我们可以省去特征工程这一较为繁琐的过程。但模型结构会比较复杂。