《数据挖掘》Task1

刘震 (31520211154070), 王莹 (31520211154089)

1 概述

我们基于scikit-learn机器学习工具包，分别使用决策树、支持向量机、KNN、多层感知机尝试对数据进行拟合。

2 实验过程

训练数据由 6270 个样本构成，每个样本具有 100 个属性 (均为 0-1 实数)，样本标签共由 0、1、2 三种类别标签组成。

每种分类方法的实验中，均使用十折交叉验证法，所述分析结果均为十折交叉验证的平均准确率。

2.1 决策树

2.1.1 概述

决策树归纳是从有类标号的训练元组中学习决策树。在决策树构造时，使用属性选择度量来选择将元组最好地划分成不同的类的属性。

属性度量是一种选择分裂准则，把给定类标记的训练元组的数据分区“最好地”划分成单独类的启发式方法。常用的属性选择度量有：信息增益、增益率和基尼指数 (Gini 指数)。

2.1.2 实验结果

我们分别使用基于信息熵、*Gini* 指数的属性度量方法构件决策树。实验结果如图 1、图 2 所示:

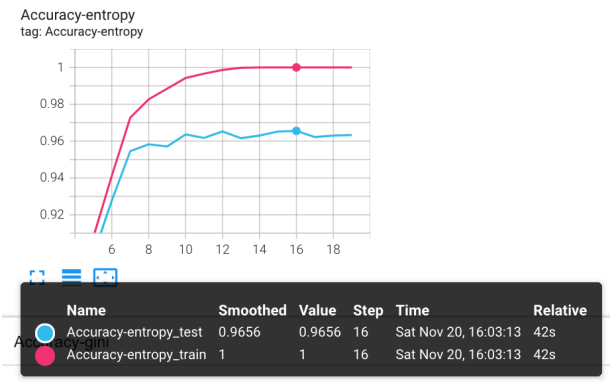


图 1: 基于信息熵的决策树

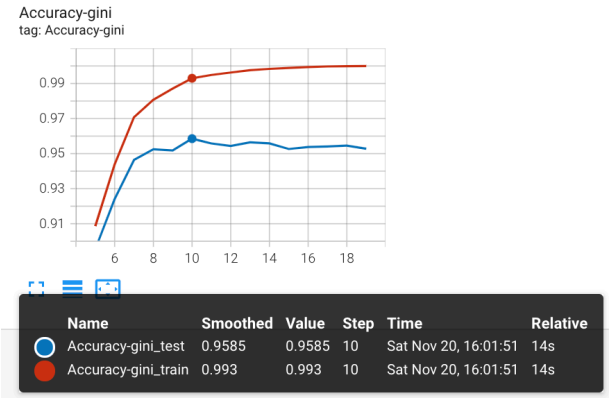


图 2: 基于 Gini 指数的决策树

	基于信息熵的决策树	基于 Gini 指数的决策树
最大深度	12	10
训练准确率	0.9987	0.9930
测试准确率	0.9652	0.9585

表 1: 决策树在验证集上的准确率及相关参数

	h 次多项式核函数	高斯径向基函数核函数	sigmoid 核函数
训练准确率	1	0.9893	0.3571
测试准确率	0.9520	0.9442	0.4012

表 2: 基于三种核函数的 SVM 在验证集上的表现

2.2 支持向量机

2.2.1 概述

支持向量机通过非线性映射，将原训练数据映射到较高维空间中，并在新的高维空间中搜索最佳分离超平面来对训练数据进行划分。

由于向高维空间映射时，对数据的运算都可以归结为对向量的点积运算，这完全等价于直接将核函数应用于原始低维数据，而不比在变换后的数据元组上进行点积运算。前述理论不仅简化了运算过程，也将寻找线性可分超平面问题归结为寻找核函数的问题。

常用的核函数有 *h 次多项式核函数*、*高斯径向基函数核函数*、*sigmoid 核函数*。

2.2.2 实验结果

我们基于 *h 次多项式核函数*、*高斯径向基函数核函数*、*sigmoid 核函数*分别进行了实验，实验结果表 2 所示。其中，基于 *h 次多项式核函数*的详细实验结果如图 3 所示。

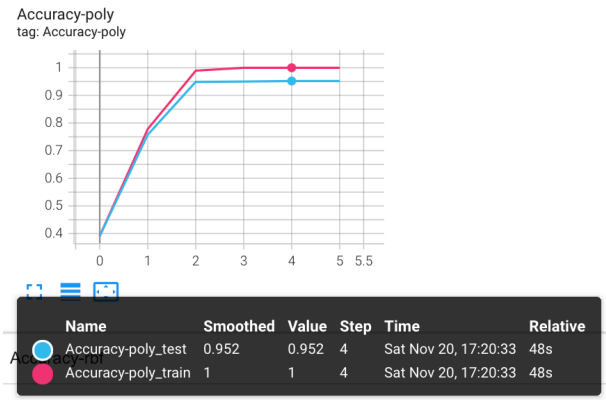


图 3: 基于 h 次多项式核函数的 SVM

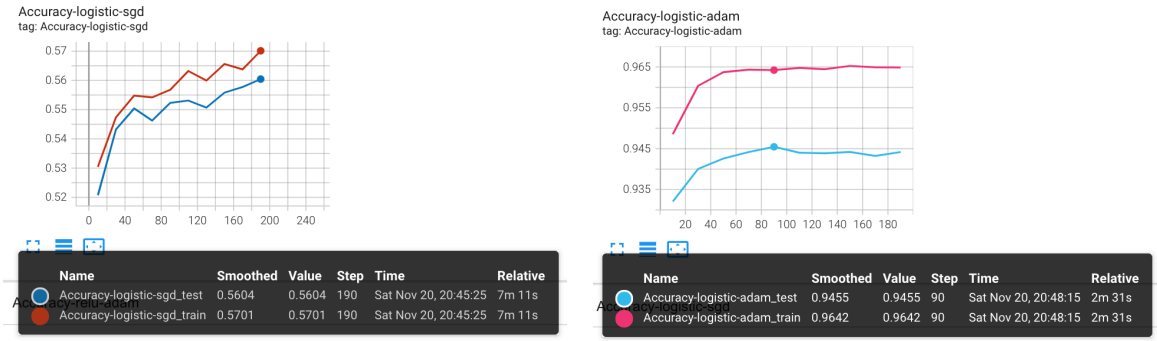
2.3 *k*-Nearest Neighbor

2.3.1 概述

k 近邻学习是一种*懒惰学习*方法，此类学习技术仅仅是把样本保存起来，训练时间开销为零。给定测试样本，基于某种距离度量找出训练集中与其最靠近的 *k* 个训练样本，然后基于这 *k* 个“邻居”的信息来进行预测。

搜索算法	权重函数	训练准确率	测试准确率
ballTree	uniform	0.8876	0.8737
	distance	0.8876	0.8737
kdTree	uniform	0.8876	0.8737
	distance	0.8876	0.8737
brute	uniform	0.8876	0.8737
	distance	0.8876	0.8737

表 3: 基于各种算法与权重函数的 k-NN



(a) 基于 *sgd* 优化算法的 *MLP*　　(b) 基于 *adam* 优化算法的 *MLP*

图 4: 基于 *logistic* 激活函数的 *MLP*

k 近邻学习的搜索过程主要有三种算法：*brute*、*kdTree*、*ballTree*。对于预测时邻居结点的权重主要有两种计算方式：*uniform*(即所有邻居结点的权重相同)、*distance*(即邻居结点的权重为距离的倒数)。

2.3.2 实验结果

我们基于 *brute*、*kdTree*、*ballTree* 三种算法，分别使用 *uniform*、*distance* 两种权重衡量方式进行了实验，各种算法的准确率如表 3 所示。

2.4 Multi-Layer-Perceptron

2.4.1 概述

多层感知机由一个输入层，一个输出层，多个隐藏层组合而成的前向结构的人工神经网络。*MLP* 模型的参数即是层与层连接的权重和偏置项。可以通 *sgd*、*Adam*，等方法通过迭代训练更新来优化权重。

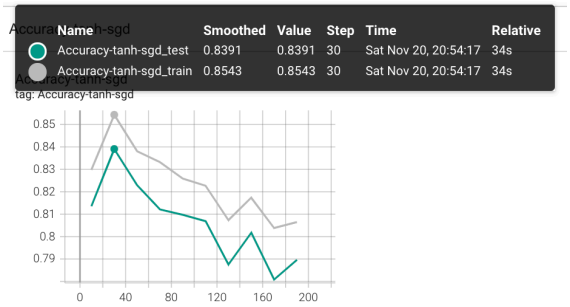
2.4.2 实验结果

我们分别使用激活函数 *logistic*、*tanh*、*relu*, 优化器 *Adam*、*sgd* 两两组合进行了训练与测试。实验结果分别如图 4、图 5、图 6 所示。

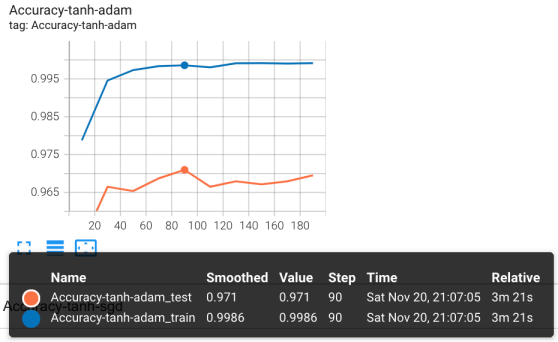
2.4.3 实验分析

2.5 实验结果

基于第二节的实验结果，我们最终选取 7 个模型在整个训练集上进行训练，并采用投票法得到测试集的最终分类结果。选取的 7 个模型及关键参数如表 5 所示。

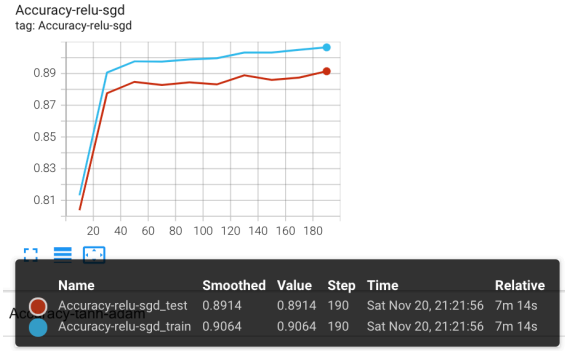


(a) 基于 *sgd* 优化算法的 *MLP*

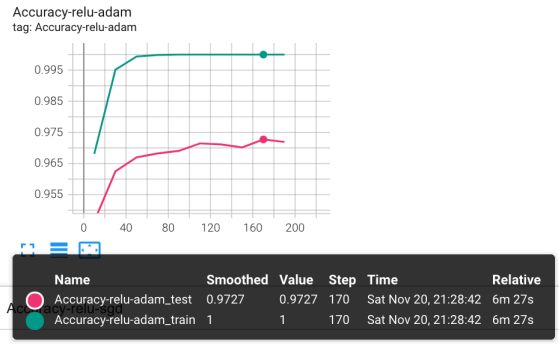


(b) 基于 *adam* 优化算法的 *MLP*

图 5: 基于 *tanh* 激活函数的 *MLP*



(a) 基于 *sgd* 优化算法的 *MLP*



(b) 基于 *adam* 优化算法的 *MLP*

图 6: 基于 *relu* 激活函数的 *MLP*

激活函数	优化算法	训练准确率	测试准确率
logistic	**sgd**	0.5676	0.5604
	adam	0.9649	0.9442
tanh	**sgd**	0.8065	0.7896
	adam	0.9992	0.9695
relu	**sgd**	0.9064	0.8914
	adam	1	0.9719

表 4: 基于各种激活函数与优化算法的 *MLP*

	最大深度	属性度量方法	核函数类型	激活函数	优化方法	隐藏层大小
决策树	10	Gini 指数	/	/	/	/
	12	信息熵	/	/	/	/
支持向量机	/	/	4 次多项式	/	/	/
	/	/	高斯径向基函数	/	/	/
多层感知机	/	/	/	logistic	adam	[90,]
	/	/	/	tanh	adam	[90,]
	/	/	/	relu	adam	[170,]

表 5: 选取的 7 个模型及关键参数