数据挖掘-支持向量机

- 概述
- 示例

概述

支持向量机，英文为Support Vector Machine，简称SVM。它是一种监督式学习的方法，它广泛的应用于统计分类分析中。支持向量机将向量映射到一个更高维的空间里，在这个空间里建立有一个最大间隔超平面。在分开数据的超平面的两边建有两个互相平行的超平面，分隔超平面使两个平行超平面的距离最大化。

支持向量机算法是以极大化类间间隔为目标，并以之作为最佳分类超平面，其中定义的类间间隔为两类样本到分类超平面的最小距离，通过引入松弛变量，使支持向量机能够解决类间重叠问题，并提高泛化能力。

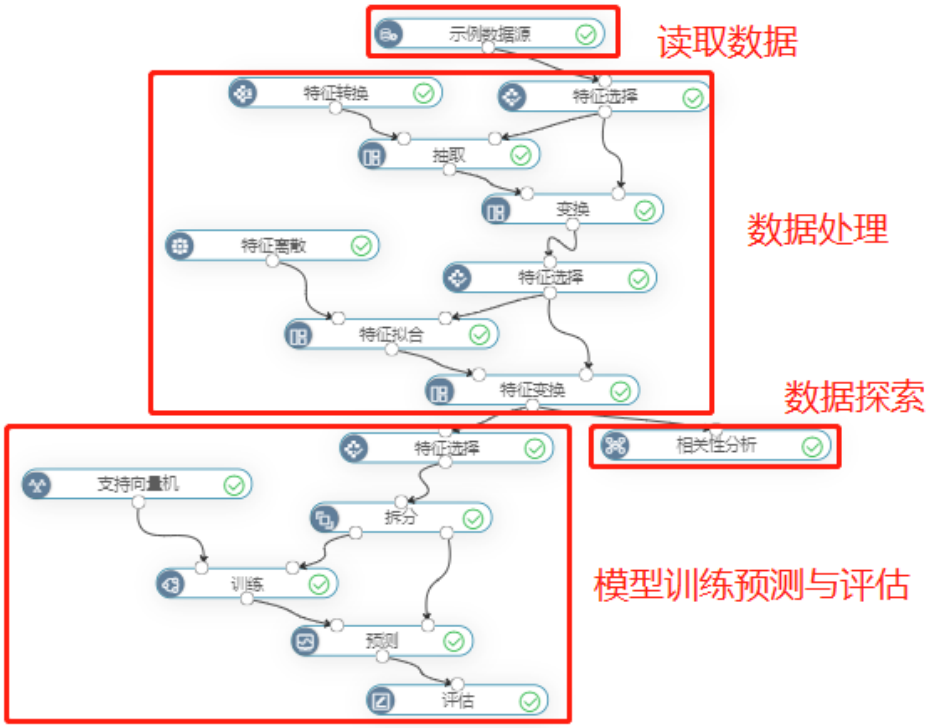
它的主要优势为：

1）特征映射，有效区分。

2）约束性求优，更加准确区分类别。

示例

使用“银行信用贷款预测”案例数据，共12个特征和1个二类的目标标签，需要预测是否贷款。通过数据预处理及模型训练，如下图：



其中，数据探索是为了解各变量之间的相关关系，方便之后数据分析中参数特征的设定；特征转换是为了将各变量中的类别型变量变换成数值型变量，类别型无法进入模型，转换后方便算法模型学习；特征离散是把连续特征分段，每一段内的原始连续特征无差别的看成同一个新特征，是为了提高模型的准确度，提高运行速度。

模型构建中，支持向量机的各个参数：

参数名称	值	说明
归一化	正则化	详情请参考 归一化 介绍说明。
	标准化	
	最小最大值归一化	
	最大绝对值归一化	
最大迭代数	参数范围为：>=1的整数，默认值为10	算法的最大迭代次数，达到最大迭代次数即退出。 最大迭代次数的值越大，模型训练更充分，但会耗费更多时间。

正则参数	参数范围为：>=0的数，默认值为 ：0。	正则项系数，损失函数中的 。 正则化可以解决模型训练中的过拟合现象； 正则项系数越大，模型越不会过拟合。
收敛阈值	参数范围为：>=0的数，默认值为 ：0.000001。	收敛误差值。 收敛误差值，当损失函数取值优化到小于收敛阈值时停止迭代。
分类阈值	参数范围为：0~1。默认值为：0.5。	在二进制分类中设置阈值thresholds。 如果模型预测结果为分类标签1的估计概率>thresholds，则预测为1，否则为0。高阈值是鼓励模型更频繁地预测0，反之则预测为1。
自动调参设置	系统默认的各项参数值范围。	必须结合“启用自动调参”功能使用。系统将对这些范围内的参数值循环调参，匹配出最优的组合。若运行速度比较慢，可将参数范围调小点。 启用自动调参：勾选该项，则系统自动调参数，不需要用户手工设置参数。

右键单击评估节点，选择查看分析结果，如下图：

指标	值		
confusion matrix(混淆矩阵)	真实|预测	0	1
	0	12042	2
	1	1577	1
accuracy(准确率)	0.8840845690794303		
roc曲线	查看ROC曲线		
auc	0.8788538455742549		
ks	查看KS曲线		
weighted precision(加权精确率)	0.8203916990963109		
weighted F1 score(加权F1分数)	0.8299038113179478		
Class 0.0 precision(精确率)	0.8842058888317791		
Class 0.0 recall(召回率)	0.9998339422118897		

F1分数决定模型效果的程度，F1值越大，说明模型预测的效果最佳。这里f1值为0.82。说明该模型效果还是不错的。效果不理想的化还可以继续调优模型。