模型选择---机器学习方法及判别方法应用

机器学习致力于建立计算模型，以发现并描述已知数据集中的规律并进行预测推断。此次对机器学习领域的学习及研究主要包括使用机器学习方法进行数据及模型类型选择分析、变量预处理、变量筛选、训练建模、预测及模型比较，以下对研究结果进行展示：

1、应用软件

R

2、数据源：

- ✧ 使用已建立的行为评分卡宽表（总数量级为 20 万，坏客户占比约 11%，共 27 个变量均已为筛选后的强变量）
- ✧ 随机抽样：20000 个数据（总数据量为 20 万左右--主要考虑运行速度且目标为模型比较，所以仅取数据量的 10%左右）

3、数据及模型类型选择分析：

- ✧ 因变量 Target（好客户为 0，坏客户为 1）；自变量 26 个（均为客户信息），包含文本型、哑变量、数值型、整数型
- ✧ 变量分析-模型选择：因变量为典型的二元分类变量，针对此类因变量，传统统计一般采用 Logistic 回归进行分析；但考虑到目前研究的发展，多种方法已经可以较好的处理此类数据，并能够突破一些正态总体的假定条件，机器学习方法就是其中之一。
- ✧ 模型类型选择分析：由于因变量为二元分类变量，应用机器学习模型做回归分析，并不合理，应属于分类问题，即各个变量根据目标变量（好坏客户进行分类）。本文为证实此类问题的应用准确性，分别进行了回归分析和分类分析，并进行对比，以便做最好的推断选择。
- ✧ 模型评价标准选择分析：传统的模型评价指标有 AIC、K-S 曲线等，但这些方法都建立在误差服从正态分布的假设条件下进行的，并不适用机器学习的评价，针对机器学习方法，交叉验证方法是较为实用且可靠的评价指标，也是评价预测效果的较好评价手段；此外为方便理解，本文还应用 ROC 曲线对部分模型预测结果进行了评价，详见下面介绍。
- ✧ 备注：

  1）除机器学习方法外，为方便对比，同时应用了多种优秀的分类方法。

  2）因为所选取的变量均为筛选后的强变量，模型对比效果并不明显，日后如有应用可直接将全部变量放入，进行更为综合的模型选择。

4、数据预处理：

✧ 进行空值填补：根据业务知识对变量空值进行填补。

✧ 文本型及因子型变量转化：WOE 替换/哑变量法。

5、模型选择：

以下采用常用的几种机器学习算法和判别方法分别对数据源进行建模及预测推断（决策树；随机森林；SVM 支持向量机；Boosting；Bagging；神经网络；K 近邻；线性判别分析；混合判别分析；灵活判别分析）

使用的分析工具为 R 语言（由于 R 语言有许多开源的机器学习函数包，下面会简单介绍各种机器学习用到的具体函数）：

1） 决策树(Decision Tree)

R 软件包-rpart.plot(函数：rpart)

2） 随机森林(Random Forest)

R 软件包-randomForest(函数：randomForest)

3） 自适应助推法(Adaboost)

R 软件包-mboost(函数：boosting)

4） 自助整合法(Bootstrap Aggregating,Bagging)

R 软件包-ipred、adabag(函数:bagging)

5）支持向量机(Support Vector Machine,SVM)

a.R 软件包-e1071(函数:svm)

b.R 软件包-kernlab(函数:ksvm)

6）神经网络(Neural Network)

R 软件包-nnet(函数：nnet)

7）线性判别分析(Linear Discriminant Analysis,LDA)

R 软件包-MASS(函数：lda)

8）混合判别分析(Mixture Discriminant Analysis,MDA)

R 软件包-mda(函数：mda)

9）灵活判别分析(Flexible Discriminant Analysis,FDA)

R 软件包-mda(函数：fda)

10）K 近邻法(K-Nearest Neighbour,KNN)

R 软件包-kknn(函数：kknn)

4、推断预测与模型比较方法

1） 交叉验证方法：

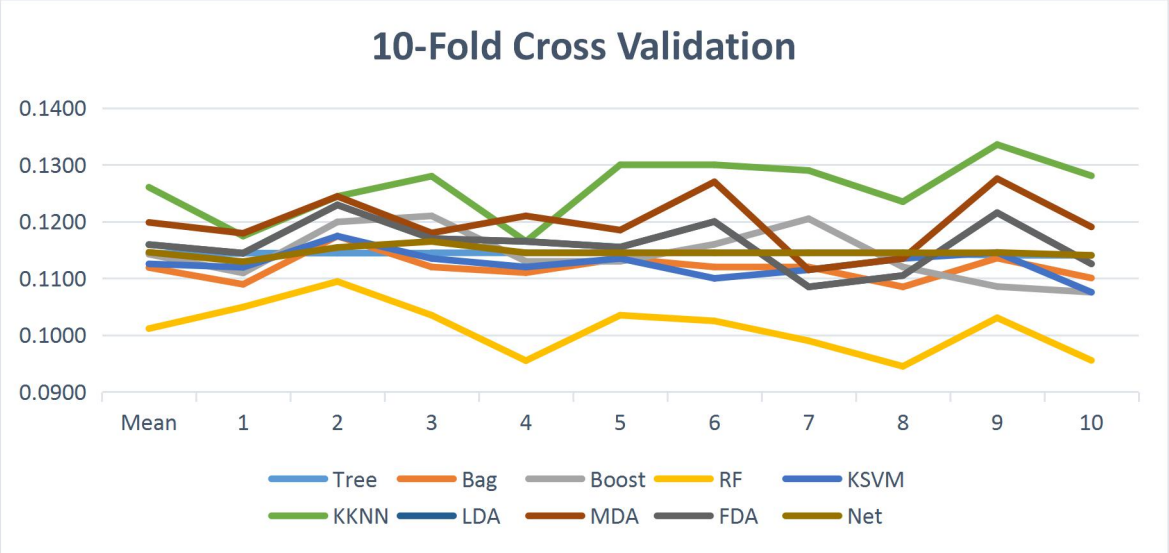
把数据随机分成 n 份(即:n 折)，依次取 1 份作为测试集，其余(n-1)份一起作为训练集,用各种方法拟合训练集数据得到具体模型,再用未参加拟合的测试集代入模型得到误差。这样做 n 次，得到 n 个误差，平均起来作为该方法的交叉验证误差。这一套训练集-测试集对不同方法做交叉验证，可得到它们的交叉验证误差。最后比较这些误差来确定哪种方法最合适。

对于分类问题，我们使用的是在训练集训练出来的模型拟合测试集所得到的预测误差。

在 n 折交叉验证中，把得到的测试集的误差平均起来。

2）11 类分类模型交叉验证结果展示

基于分类模型的机器学习及判别方法应用-10 折交叉验证评价结果											
	Tree	Boost	Bag	RF	KSVM	KKNN	LDA	MDA	FDA	Net	SVM
Mean	0.1144	0.1142	0.1119	0.1011	0.1125	0.1261	0.1159	0.1199	0.1159	0.1146	0.3335
1	0.1144	0.1109	0.1089	0.1049	0.1119	0.1174	0.1144	0.1179	0.1144	0.1129	0.3383
2	0.1144	0.1199	0.1174	0.1094	0.1174	0.1244	0.1229	0.1244	0.1229	0.1154	0.3313
3	0.1145	0.1210	0.1120	0.1035	0.1135	0.1280	0.1170	0.1180	0.1170	0.1165	0.3240
4	0.1145	0.1130	0.1110	0.0955	0.1120	0.1165	0.1165	0.1210	0.1165	0.1145	0.3380
5	0.1145	0.1130	0.1135	0.1035	0.1135	0.1300	0.1155	0.1185	0.1155	0.1145	0.3325
6	0.1145	0.1160	0.1120	0.1025	0.1100	0.1300	0.1200	0.1270	0.1200	0.1145	0.3210
7	0.1145	0.1205	0.1120	0.0990	0.1115	0.1290	0.1085	0.1115	0.1085	0.1145	0.3455
8	0.1145	0.1120	0.1085	0.0945	0.1135	0.1235	0.1105	0.1135	0.1105	0.1145	0.3460
9	0.1141	0.1086	0.1136	0.1031	0.1146	0.1336	0.1216	0.1276	0.1216	0.1146	0.3422
10	0.1141	0.1076	0.1101	0.0955	0.1076	0.1281	0.1126	0.1191	0.1126	0.1141	0.3162



3）上图结论：

- ✧ 模型稳定性：KSVM 和 Bagging 算法得出的模型稳定性较好；KKNN 稳定性较差。
- ✧ 预测准确性：从预测误差的数值来看，拟合预测效果非常好，其中相对来说随机森林、KSVM 和 Bagging 算法得出的模型误差较低；KKNN 模型误差较高。
- ✧ 算法速度：随机森林、支持向量机计算速度较慢；Boosting、Bagging 计算速度也较慢，但在数据量较大时，Bagging 算法较 Boosting 算法快；其余算法在数据量一定时计算速度相当。

4）交叉验证--模型选择评价标准与优势

- ✧ 模型稳定性判定：可以通过多折交叉验证结果变化趋势来考查模型稳定性
- ✧ 预测准确性考查：根据交叉验证最终预测误差取值来评价拟合及预测效果 （即：误差越小越好）
- ✧ 算法运行速度检测：主要结合数据量及变量数量进行测试考查

5）8 种机器学习回归模型进行交叉验证结果展示

（回归验证使用的程序包、函数有所不同，具体将在较优应用中给出，在此不做详解）

交叉验证方法差异

这里采用的是标准化均方误差 NMSE 来计算交叉验证误差，这里当 NMSE>1，意味着模型拟合效果差，不建议使用。

10 折交叉验证--基于 8 种机器学习模型的回归结果对照								
	tree	boost	bboost	bagging	RF	svm	ksvm	net
Mean	0.9006	0.8693	0.8531	0.8808	0.8459	0.9689	0.9834	0.3954
1	0.8772	0.8507	0.8345	0.8619	0.8354	0.9772	0.9948	0.2497
2	0.9275	0.8917	0.8863	0.9037	0.8771	1.0022	1.0177	0.2822
3	0.8557	0.8306	0.8054	0.8408	0.7989	0.9448	0.9652	0.4402
4	0.8770	0.8646	0.8389	0.8470	0.8509	0.9666	0.9781	0.2402
5	0.8681	0.8361	0.8152	0.8492	0.8132	0.9305	0.9536	0.3739
6	0.9130	0.8771	0.8616	0.8963	0.8745	0.9927	1.0004	0.7727
7	0.9252	0.9023	0.8890	0.9051	0.8652	0.9420	0.9593	0.5574
8	0.8752	0.8535	0.8285	0.8447	0.8236	0.9728	0.9848	0.2330
9	0.9468	0.8941	0.8826	0.9325	0.8728	0.9677	0.9784	0.6149
10	0.9404	0.8927	0.8894	0.9268	0.8470	0.9920	1.0015	0.1897

上图结论：从表中数据结果可以看出，除神经网络外，模型拟合预测的效果均较差，都较接近 1，但神经网络由于数据未进行归一化处理，可能存在过度拟合的情况，拟合结果并不稳定。所以应用机器学习的回归模型拟合该数据并不合理，同时应用回归交叉验证的方法来验证模型也不合适。即验证，当因变量为分类变量时，应选择机器学习分类模型进行拟合。

5） ROC 曲线评价模型预测效果（数据均做了 WOE 替换）

考虑到传统统计在因变量为哑变量时一般采用 Logistic 回归的方法来进行回归分析，下面我们将几种常用的机器学习方法与 Logistic 回归分别做 ROC 曲线图，进行下预测效果对比。（注：变量均为强变量且数据均做了数值化处理，预测效果差异并不明显）