编号	方法	适用数据	适用问题	模型类型	模型特点	学习策略	目标（损失）函数	正则项（惩罚函	学习算法	缺失值敏感性	异常值敏感性	数据处理	模型实现（R）	模型实现（python）	方法特点	方法优点	方法缺点
1	线性回归	y服从正态分布，X与y线性相关	连续值预测问题	判别模型	可以对样本非线性，只要对参数θ线性	最小二乘（OLS）	平方和损失	L1范式/L2范式	参数解析/SGD(随机梯度下降)	敏感	敏感	去除多重共线性，y正态性转换	MASS包glm函数族	sklearn库，sklearn.linear_model import LinearRegression	自动特征选择/特征相关性	1.可用于较小样本量；2.实现简单	1.局限性大
2	Logistic回归	y服从二项分布或k分类	二分类或有序（无序）多分类问题	判别模型	特征条件下类别的条件概率分布，对数线性模型	极大似然估计	Logistic损失		改进的迭代尺度算法，SGD,拟牛顿法	不敏感	不敏感		MASS包glm函数族，glm(y_train ~ ., data = x,family='binomial')	sklearn库，sklearn.linear_model import LogisticRegression，model = LogisticRegression()	特征选择/危险因素评价	1.易解释；2.可用于危险因素评价；3.特征选择	1.需要较大样本量
3	决策树	任意分布数据或不平衡数据	分类或回归	判别模型	应用信息熵下降最快选择特征作为结点	正则化的极大似然估计		树节点个数及树的深度，树剪枝	C4.5\CART\Gini系数，递归生成，剪枝	不敏感	不敏感		rpart包rpart函数，rpart(y_train ~ ., data = x,method="class")	sklearn库，from sklearn import tree，tree.DecisionTreeClassifier(criterion='gini')/tree.DecisionTreeRegressor()	特征选择/特征重要性排序	1.易于理解，解释直观；2.适用于数据探索阶段；3.对数据缺失或异常不敏感；4.对数据类型不敏感	1.易产生过拟合；2.连续变量截断分类，易丢失信息
4	随机森林	任意分布数据或不平衡数据	分类或回归	判别模型	集成思想，样本集中有放回重复采样获取m个分类器，依据投票结果选择分类	booststrap抽样，bagging集成投票获得分类			Bagging集成提升算法	不敏感	不敏感		randomForest包，randomForest(Species ~ ., x,ntree=500)	sklearn库，from sklearn.ensemble import RandomForestClassifier，RandomForestClassifier()	特征选择/特征重要性排序/降维，通过减少模型方差提高性能	1.同决策树；2.可以处理高维数据，有降维功能；3.可用于不平衡分类问题；4.可延伸到非标签数据，用于无监督聚类	1.同决策树；2.不易于解释，黑箱效应

5	提升方法（GBDT，XGBoost）	数据无缺失，进行one-hot编码，转换为数值型向量	分类或回归	判别模型	弱分类器的线性组合，弱预测模型的加权累加，每一步都依据损失函数的梯度方向进行	对目标函数分解为若干基函数的加权和	使用线性搜索计算学习率，最小化损失函数	决策树剪枝，叶子节点数目，叶子节点包含的最小样本数量（如XGBoost使用叶节点总数和叶权值平方和的加权），梯度提升迭代次数	SGD，二阶Taylor展开（XGBoost）	敏感	敏感	数据缺失填补，进行one-hot编码，转换为数值型稀疏矩阵	gbm包或xgboost包或caret包，fitControl <- trainControl(method = "repeatedcv", number = 4, repeats = 4), fit <- train(y ~ ., data = x, method = "gbm", trControl = fitControl,verbose = FALSE), predicted= predict(fit,x_test,type= "prob")[,2]	sklearn库，from sklearn.ensemble import GradientBoostingClassifier，clf = GradientBoostingClassifier()/import xgboost as xgb，xgb.train(param, data_train, num_boost_round=n_round, evals=watchlist)	通过减少模型偏差提高性能	主要指XGBoost优点：1.有正则化项以防止过拟合；2.可并行；3.内建缺失值处理及CV处理；4.方便的树剪枝；5.速度更快捷	1.需要大量调参
6	提升方法（主要是AdaBoost）	数值型或二分类数据	二分类	判别模型	弱分类器的线性组合	对训练失败的训练例赋以较大的权重，对分类问题采用有权重的投票方式，对回归问题采用加权平均的方法进行判别	极小化加法模型的指数损失		前向分步加法算法	敏感	敏感	数据缺失填补，分类标签修改为1或-1	boost包，adaboost (xlearn, ylearn, xtest, presel = 200, mfinal = 100)或adabag 包	创建弱分类器，AdaBoost的训练函数，创建分类函数，分类加权		1.属于高精度分类器；2.不用做特征筛选；3.不易过拟合	1.对异常值敏感
7	SVM	数值型无缺失	二分类	判别模型	分离超平面，核函数技巧	极小化正则化合页损失，软间隔最大化	对特征空间划分的最优超平面是SVM的目标,最大化分类边际的思想是SVM方法的核心	L2范式，加入松弛因子，惩罚因子	序列最小最优化算法（SMO），约束条件下的最优化问题（拉格朗日对偶	敏感	不敏感		e1071包，svm(y_train ~ ., data = x)	sklearn库，from sklearn import svm，svm.svc(kernel='linear', c=1, gamma=1)		1.可以解决高维问题，即大型特征空间；2.能够处理非线性特征的相互作用；3.无需依赖整个数据；4.泛化能力强；	1内存消耗大；2.难以解释；3.需要调参；4.对非线性问题没有通用解决方案；5.对缺失数据敏感；6.核函数难找
8	朴素贝叶斯		二/多分类问题	生成模型	基于贝叶斯定理和特征条件独立假设	基于特征条件独立假设学习输入输出联合概率，利用贝叶斯定理求后验概率最大			概率计算公式，最大似然估计算法，贝叶斯估计（拉普拉斯平滑估计）	不敏感	不敏感	1.提取特征相关性分析；2.手动选择特征；3.连续值转换为高斯分布	e1071包，naiveBayes(y_train ~ ., data = x)	sklearn库，from sklearn.naive_bayes import GaussianNB		1.算法简单，易于实现；2.算法稳定，健壮性好；3.计算速度快	1.特征条件独立假设现实很难实现（可用贝叶斯网络）；2提前进行特征选择困难；3.防止后验概率可能为0

9	K近邻法		分类或多回归问题	判别模型	距离度量，k值选择和分离决策规则确定	首先确定输入实例点的k个最近邻训练实例点，利用这k个训练实例			多数表决	不敏感	不敏感	归一化处理	knn包，knn(y_train ~ ., data = x,k=5)	sklearn库，from sklearn.neighbors import KNeighborsClassifier,		1.可用于非线性分类；时间复杂度低	1.样本不平衡问题；2.消耗内存（用kd树优化）；3.k值难以确定，
10	人工神经网络（深度学习）		分类/预测/识别	判别模型	由大量节点（神经元）连接的运算模型	模拟生物神经信号传递网络		正则化或dropout	BP算法，SGD算法等	需要填补	不敏感	去均值；归一化；PCA/白化等	Mxnet包等	caffe、Mxnet、tensorFlow等开源框架	可用于语音识别、图像识别、NLP等多个领域功能强大	1.分类的准确度高；2.并行分布处理能力强,分布存储及学习能力强；3.对噪声神经有较强的鲁棒性和容错能力，能充分逼近复杂的非线性关系；4.具备联想记忆的功能；5.应用	1.需要大量的参数；2.不能观察之间的学习过程，输出结果难以解释
11	EM	可处理缺损数据，截尾数据，带有噪声等不完全数据参数估计	概率模型参数估计	生成模型	含隐变量概率模型	极大似然估计，极大后验概率估计	对数似然损失		迭代算法	不敏感	不敏感		mclust包实现高斯混合模型GMM聚类分析	构建EM算法函数	应用于处理缺损数据，截尾数据，带有噪声等不完全数据参数估计，也可用于数据聚类	1.简单和稳定	1.容易陷入局部最优
12	kmean聚类	数值类型数据	聚类问题	非监督学习		按数据内在相似性划分类别，使类别内数据相似性大，而类别间的数据相似性小				敏感	敏感	归一化处理	library (cluster) fit <- kmeans(X, 3)	sklearn库，from sklearn.cluster import KMeans, KMeans(n_clusters =3, random_state=0)		1.经典、简单、快速；2.当簇接近高斯分布时效果较好	1.可能收敛到局部最小值，在大规模数据上收敛较慢；2.K值比较难以选取；3.对初值的簇心值敏感；4.不适合于发现非凸面形状的簇，或者大小差别很大的簇。5.对于”噪声”和孤立点数据敏感，少量的该类数据能够对平均值产生极大影响

13	降维算法（PCA，因子分析）	高维数据，数值型数据	维度规约	判别模型		把原来有多个指标转化成少数几个代表性好的综合指标，以反映原来指标大部分信息				敏感	敏感	数据填补，数值归一化	library(stats) pca <- princomp(train, cor = TRUE)	from sklearn import decomposition, pca= decomposition.PCA(n_components=k), train_reduced = pca.fit_transform(train), test_reduced = pca.transform(test)	具有较强假设条件：线性、大方差对应主要数据结构、主成分之间正交	1.方法简单，易于实现；2.消除评价指标见的相关性；3.	1.主成分难以解释；2.当样本具有非线性性质时，降维结果无法反映其特性；3.主成分个数难以确定