第 4 讲　MATLAB 数据建模方法(下)：机器学习方法

作者：马文辉，MathWorks 中国

近年来，全国赛的题目中，多多少少都有些数据，而且数据量总体来说呈不断增加的趋势， 这是由于在科研界和工业界已积累了比较丰富的数据，伴随大数据概念的兴起及机器学习技术的发展， 这些数据需要转化成更有意义的知识或模型。 所以在建模比赛中， 只要数据量还比较大， 就有机器学习的用武之地。

1 MATLAB 机器学习概况

机器学习(Machine Learning)是一门多领域交叉学科，它涉及到概率论、统计学、计算机科学以及软件工程。机器学习是指一套工具或方法，凭借这套工具和方法，利用历史数据对机器进行“训练”进而“学习”到某种模式或规律，并建立预测未来结果的模型。

机器学习涉及两类学习方法（如图 1）：有监督学习，主要用于决策支持，它利用有标识的历史数据进行训练，以实现对新数据的标识的预测。有监督学习方法主要包括分类和回归；无监督学习，主要用于知识发现，它在历史数据中发现隐藏的模式或内在结构。无监督学习方法主要包括聚类。

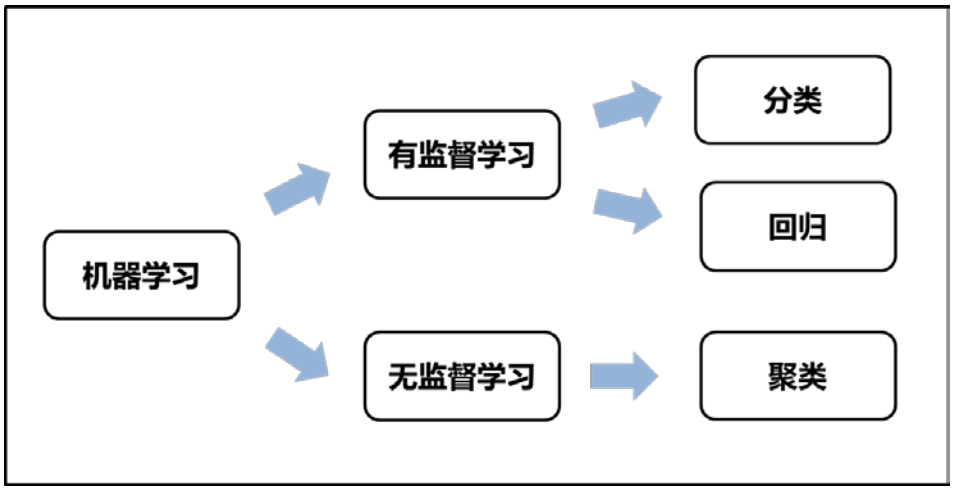

图 1　机器学习方法

MATLAB 统计与机器学习工具箱（Statistics and Machine Learning Toolbox）支持大量的分类模型、回归模型和聚类的模型，并提供专门应用程序（APP），以图形化的方式实现模型的训练、验证，以及模型之间的比较。

- 分类

分类技术预测的数据对象是离散值。例如，电子邮件是否为垃圾邮件，肿瘤是癌性还是良性等等。 分类模型将输入数据分类。 典型应用包括医学成像，信用评分等。MATLAB 提供的分类算法包括：

图 2 分类算法家族

- 回归

回归技术预测的数据对象是连续值。 例如，温度变化或功率需求波动。 典型应用包括电力负荷预测和算法交易等。回归模型包括一元回归和多元回归，线性回归和非线性回归，MATLAB 提供的回归算法有：

图 3 回归算法家族

- 聚类

聚类算法用于在数据中寻找隐藏的模式或分组。聚类算法构成分组或类，类中的数据具有更高的相似度。聚类建模的相似度衡量可以通过欧几里得距离、概率距离或其他指标进行定义。MATLAB 支持的聚类算法有：

图 4 聚类算法家族

以下将通过一些示例演示如何使用 MATLAB 提供的机器学习相关算法进行数据的分类、回归和聚类

2 分类技术

- 支持向量机（SVM）

SVM 在小样本、非线性及高维数据分类中具有很强的优势。在 MATLAB 中，可以利用 SVM 解决二分类问题。同时也可以使用 SVM 进行数据的多分类划分。

1） 二分类

以下示例显示了利用 MATLAB 提供的支持向量机模型进行二分类，并在图中画出了支持向量的分布情况(图 1 中圆圈内的点表示支持向量)。MATLAB 支持 SVM 的核函数（KernelFunction 参数）有：线性核函数（Linear），多项式核函数（Polynomial）、高斯核函数（Gaussian）。

```
%% 支持向量机模型
load fisheriris;
```

```
% 数据只取两个分类：'versicolor'和'virginica'
inds = ~strcmp(species,'setosa');
% 使用两个维度
X = meas(inds,3:4);
y = species(inds);
tabulate(y)
```

```
       Value    Count   Percent
  versicolor       50     50.00%
   virginica       50     50.00%
```

```
%% SVM模型训练，使用线性核函数
SVMModel = fitcsvm(X,y,'KernelFunction','linear');
%% 查看进行数据划分的支持向量
sv = SVMModel.SupportVectors;
figure
gscatter(X(:,1),X(:,2),y)
hold on
plot(sv(:,1),sv(:,2),'ko','MarkerSize',10)
legend('versicolor','virginica','Support Vector')
hold off
```

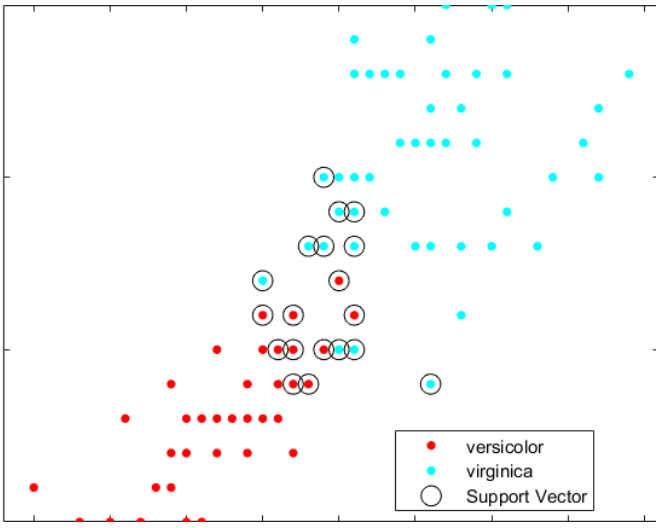

图 5 支持向量分布

2) 多分类

MATLAB 多分类问题的处理是基于二分类模型.下面的示例演示如何利用 SVM 的二分类模型并结合 fitcecoc 函数解决多分类问题。

```
% 导入 Fisher's iris 数据集.
load fisheriris
X = meas;
Y = species;
tabulate(Y)
```

```
       Value    Count   Percent
      setosa       50     33.33%
  versicolor       50     33.33%
   virginica       50     33.33%
```

```
% 创建SVM模板（二分类模型），并对分类变量进行标准化处理
% predictors.
t = templateSVM('Standardize',1);
```

```
% 基于SVM二分类模型进行训练并生成多分类模型
Mdl = fitcecoc(X,Y,'Learners',t,...
    'ClassNames',{'setosa','versicolor','virginica'})
```

```
Mdl =
  ClassificationECOC
             ResponseName: 'Y'
    CategoricalPredictors: []
               ClassNames: {'setosa'  'versicolor'  'virginica'}
           ScoreTransform: 'none'
           BinaryLearners: {3×1 cell}
               CodingName: 'onevsone'
```

MATLAB 的 fitcecoc 函数支持多种二分类模型，例如，templateKNN，templateTree，templateLinear，templateNaiveBayes，等等。

3 回归

回归模型描述了响应（输出）变量与一个或多个预测变量（输入）变量之间的关系。 MATLAB 支持线性，广义线性和非线性回归模型。以下示例演示如何训练逻辑回归模型。

- 逻辑回归

在 MATLAB 中，逻辑回归属于广义线性回归的范畴，可以通过使用 fitglm 函数实现逻辑回归模型的训练。

```
% 判定不同体重、年龄和性别的人的吸烟的概率
load hospital
dsa = hospital;
```

```
% 指定模型使用的计算公式。
% 公式的书写方式符合Wilkinson Notation，详情请查看：
% http://cn.mathworks.com/help/stats/wilkinson-notation.html

modelspec = 'Smoker ~ 1 + Age + Weight + Sex + Age:Weight + Age:Sex + Weight:Sex';
```

```
% 通过参数'Disribution'指定'binomial'构建逻辑回归模型。
mdl = fitglm(dsa,modelspec,'Distribution','binomial')
```

```
mdl =
Generalized linear regression model:
    logit(Smoker) ~ 1 + Sex*Age + Sex*Weight + Age*Weight
    Distribution = Binomial

Estimated Coefficients:
                        Estimate                SE                tStat               pValue
                    ____________________    ____________________    ____________________    ____________________
    (Intercept)          -6.04922294807952       19.7494553991791    -0.306298215612111    0.759377597706477
    Sex_Male             -2.28589740940292         12.423950346836     -0.1839911900473    0.854020366262019
    Age                  0.116908396522182      0.509769853043979     0.229335642004111    0.818608053266245
    Weight               0.0311093897007663      0.152084663337829     0.204553102318163    0.837921299452791
    Sex_Male:Age         0.0207335744147806      0.206813541990096     0.100252499015628    0.920143867723874
    Sex_Male:Weight      0.0121599715158412      0.053168447969146     0.228706535178489    0.819097015071498
    Age:Weight         -0.000719593684094343    0.00389639473279403    -0.184681926098985    0.853478523194913

100 observations, 93 error degrees of freedom
Dispersion: 1
Chi^2-statistic vs. constant model: 5.07, p-value = 0.535
```

4 聚类

聚类是将数据集分成组或类。 形成类，使得同一类中的数据非常相似，而不同类中的数据差异非常明显。

- 层次聚类

下面以层次聚类方法为例，演示如何利用 MATLAB 进行聚类分析。

```
% 数据导入
load fisheriris
```

```
% MATLAB中层次聚类是通过linkage函数实现。
% 通过参数可以配置距离计算方法
% 类内距离的计算方法:'euclidean',欧几里得距离;
eucD = pdist(meas,'euclidean');
% 类间距离的计算方法：'ward',最小化两个类内点之间聚类平方和;
Z = linkage(eucD,'ward');
```

```
% 使用cophenetic相关系数评价聚类计算过程(类内距离最小，类间聚类最大)
% 值越大表明距离计算结果越好
cophenet(Z,eucD)
```

```
ans =    0.872828315330562
```

```
% 生成4个类别的聚类结果
c = cluster(Z,'maxclust',4);
```

可以显示层次聚类生成的聚类树，使用 dendrogram 函数：

```
% 查看层次聚类树
dendrogram(Z)
```

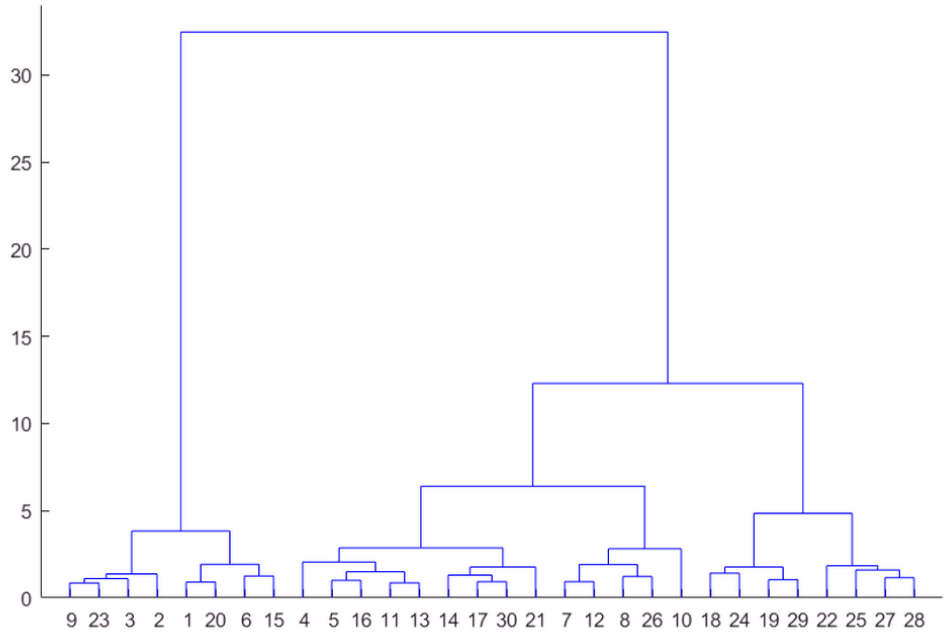

图 6 层次聚类

关于作者

马文辉，MathWorks 中国应用工程师， 南开大学工学博士，在大数据处理与分析领域有多年研究与开发经验；曾就职于 Nokia 中国研究院，Adobe 中国研发中心以及 IBM 中国。

相关资源

利用 MATLAB 进行机器学习：https://cn.mathworks.com/solutions/machine-learning.html?s_tid=hp_solutions_machine

Statistics and Machine Learning Toolbox 帮助文档：https://cn.mathworks.com/help/stats/index.html?s_cid=doc_ftr