第8章　机器学习

Jupyter金融应用从入门到实践

1950年，阿兰·图灵（Alan Turing）发表了题名为《计算机与智能》的论文，讨论了创造一种智能机器的可能性，并提出著名的“图灵测试”。1956年的达特茅斯会议上“人工智能”被提出并作为本研究领域的名称，从此开启了人工智能跌宕起伏几十年的发展历程。1997年汤姆·米切尔（Tom Mitchell）定义机器学习时提到“机器学习是对能通过经验自动改进的计算机算法的研究”。机器学习作为一种实现人工智能的重要方法，帮助科学家们在弱人工智能领域取得了巨大的突破。

机器学习常用术语

1．数据集：数据是计算机对现实世界多个维度的观测，每一次观测都形成一条记录，记录也可被称为样本，样本的集合被称为“数据集”。根据不同的使用目的，数据集还可以进一步分为“训练集”“验证集”和“测试集”。

2．特征与标签：每一条样本数据通常都由输入和对应的输出组成。输入变量X（X表示从x_1至x_n的n维列向量，下文同理），也被称为特征或解释变量；而输出变量Y，也可以视为标签或被解释变量。特征用来从多个维度描述一条样本。

3．模型训练：计算机利用给定的训练集修正模型参数的过程，被称为训练。模型是机器学习训练数据的表现形式，可以表示为条件概率分布 $P(y|x)$或者决策函数 $y=f(x)$。训练是为了让模型不断逼近数据内在的真实规律。

4．超参数：模型本身有许多未知的变量：一种是通过训练可以求解的，被称为参数；一种要在训练之前设置，被称为超参数。

5．模型预测：使用模型对测试集进行预测的过程称为模型预测。对具体的输入进行相应的输出预测时，写作$P(y|x)$或$y=f(x)$。

机器学习算法分类

表 8-1　机器学习算法分类

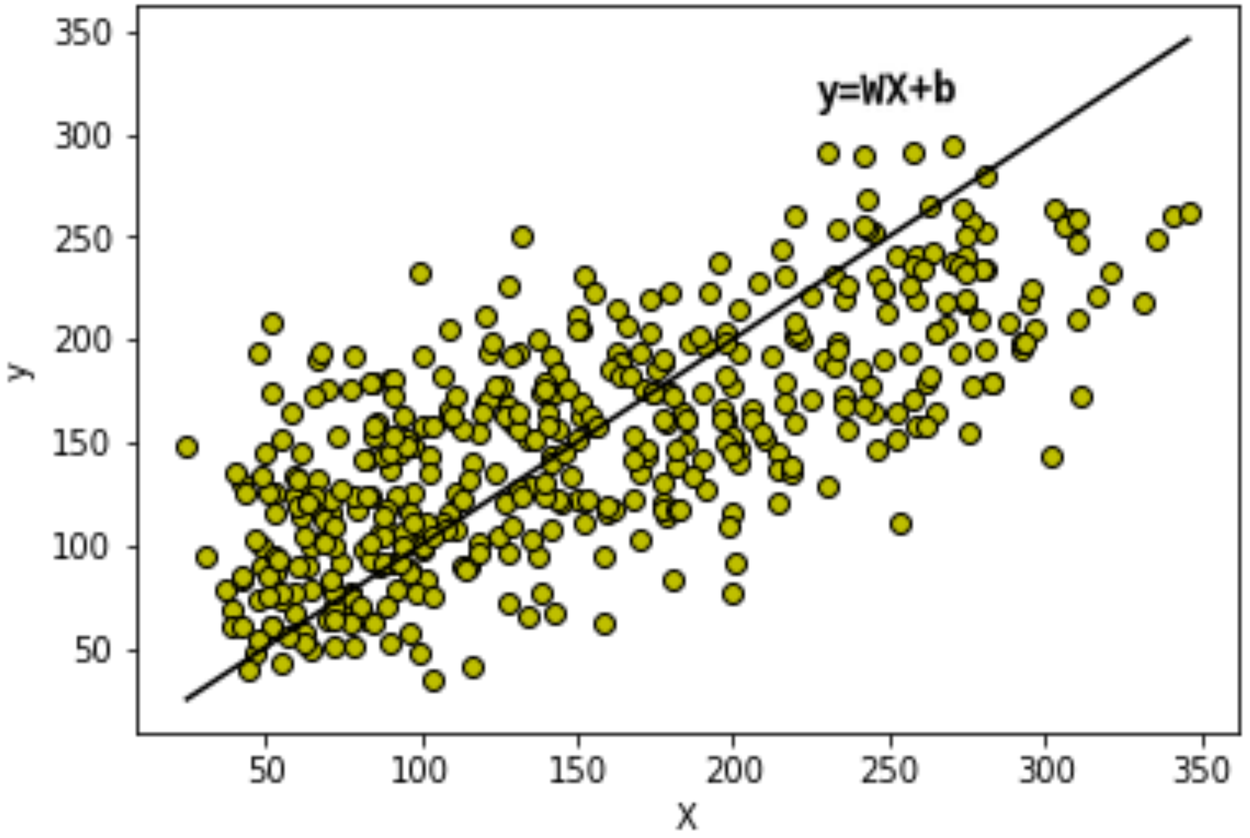
标签类型	监督学习	无监督学习
标签为连续数据	• 回归 ▪ 线性回归（Linear Regression） ▪ 非线性回归（Non-Linear Regression）	• 聚类 ▪ 基于距离 ▪ 基于密度 ▪ 基于树形结构 • 降维 ▪ 主成分分析（PCA） ▪ 线性判别分析（LDA） ▪ 最近邻
标签为离散数据	• 分类 ▪ 最近邻（KNN） ▪ 支持向量机（SVM） ▪ 逻辑回归（Logistic Regression） ▪ 朴素贝叶斯（Naïve Bayes）	• 关联分析 ▪ Apriori ▪ FP-Growth

线性回归

线性回归的数学表达式一般如下。

$$\boldsymbol{y} = \boldsymbol{W}^T\boldsymbol{X} + b$$

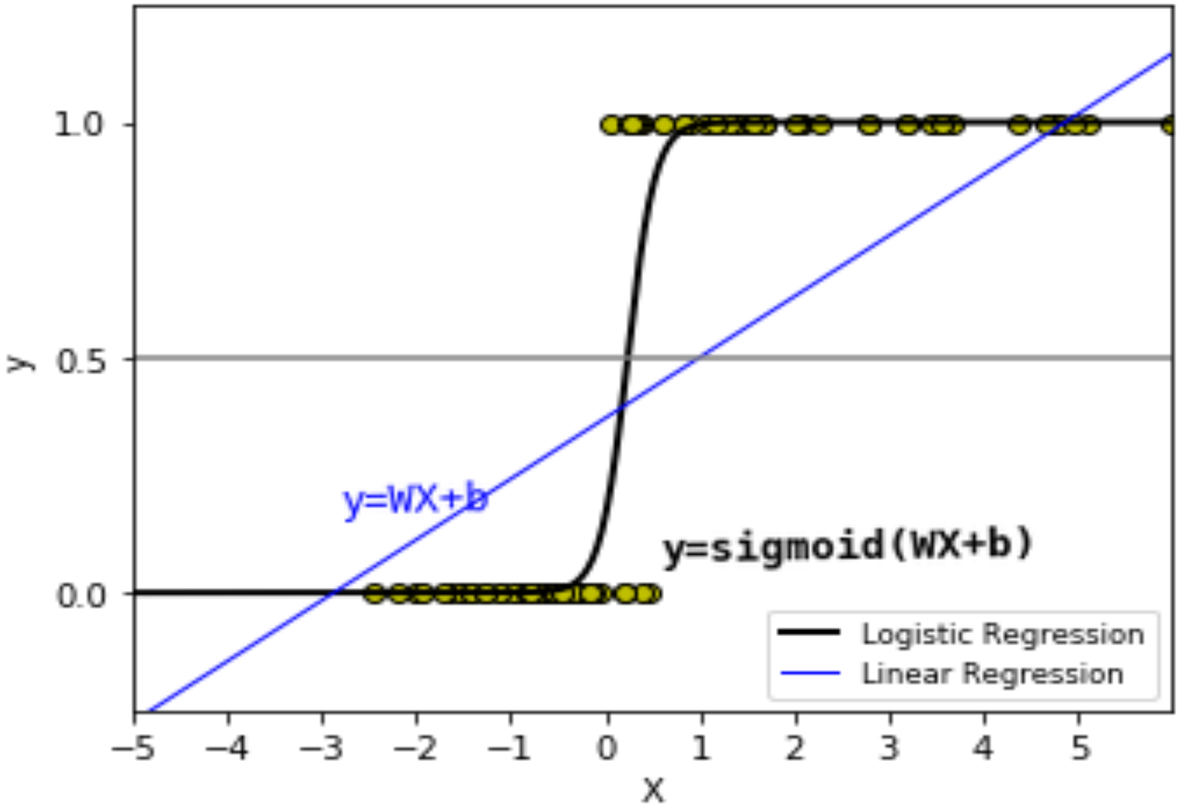
其中$\boldsymbol{X}$为输入变量，也是自变量；y是输出变量，也是因变量；$\boldsymbol{W}$叫作$\boldsymbol{X}$的系数（$\boldsymbol{W}$表示从w_1至w_n的n维列向量），b叫作偏置项，它们都是模型的参数。



逻辑回归

逻辑回归（Logistic Regression）。通过sigmoid函数将一般线性回归的结果映射到0和1之间：

$$h_{w,b}(x) = sigmoid(w^T x + b)$$



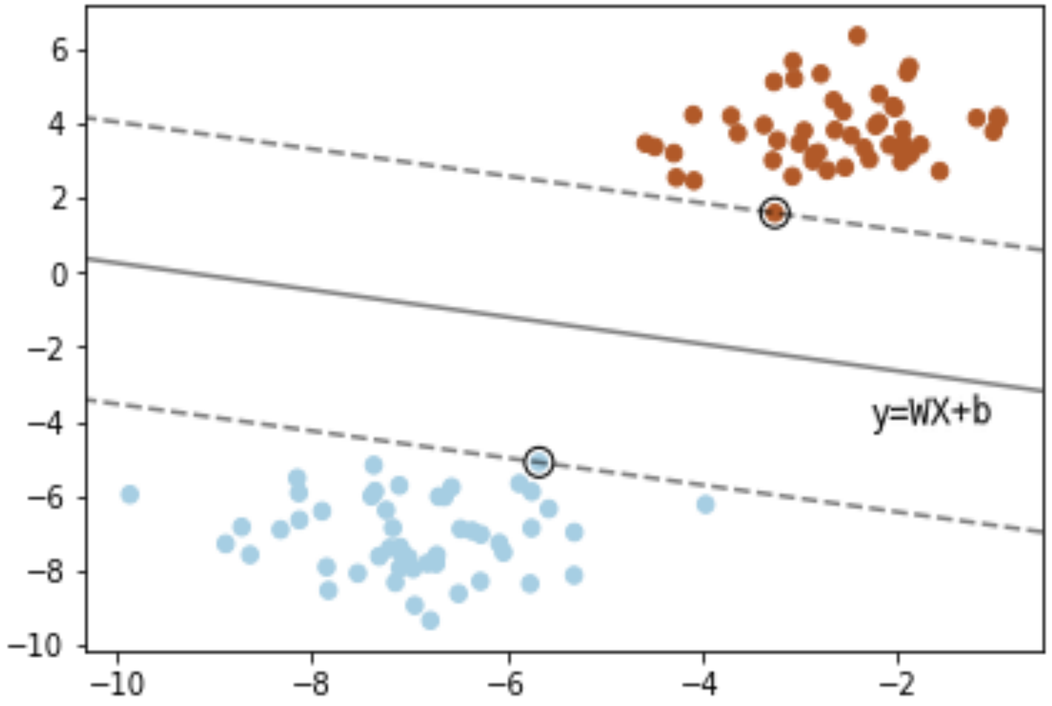
支持向量机

支持向量机（SVM）也是一种经典的二分类模型。其函数表达也是建立在一般线性回归上。支持向量机函数表达式如下：

$$f(\boldsymbol{X}) = \text{sign}\left(\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X} + b\right)$$

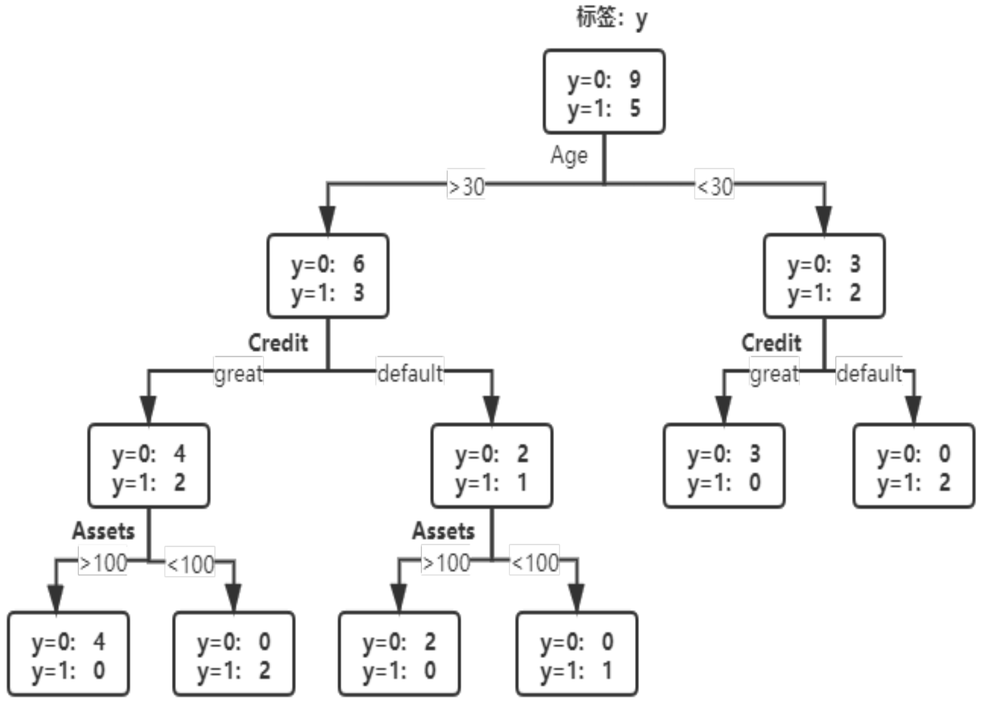
其中sign函数的含义是，对于一般线性回归所形成的直线，当点在直线右边为正例，在直线左边为反例：

$$f(\boldsymbol{X}) = \begin{cases} +1, \boldsymbol{W}^{\mathrm{T}}X + b \geqslant 0 \\ -1, \boldsymbol{W}^{\mathrm{T}}X + b < 0 \end{cases}$$



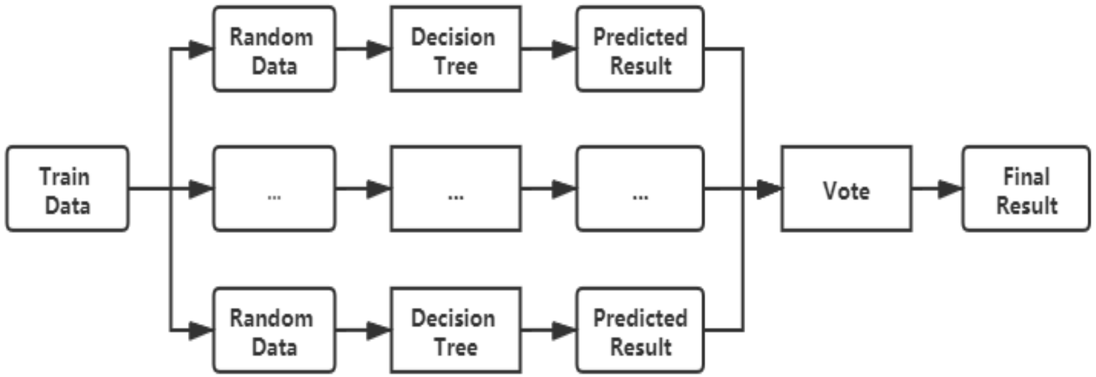
决策树

决策树（Decision Tree）是一种基本的分类和回归模型。决策树由节点和有向边组成，内部节点表示一个特征，叶结点表示类别标签。从根结点开始，对样本的每个特征进行分配计算，一步一步分配到子节点，直至到达叶节点，得到分类结果。



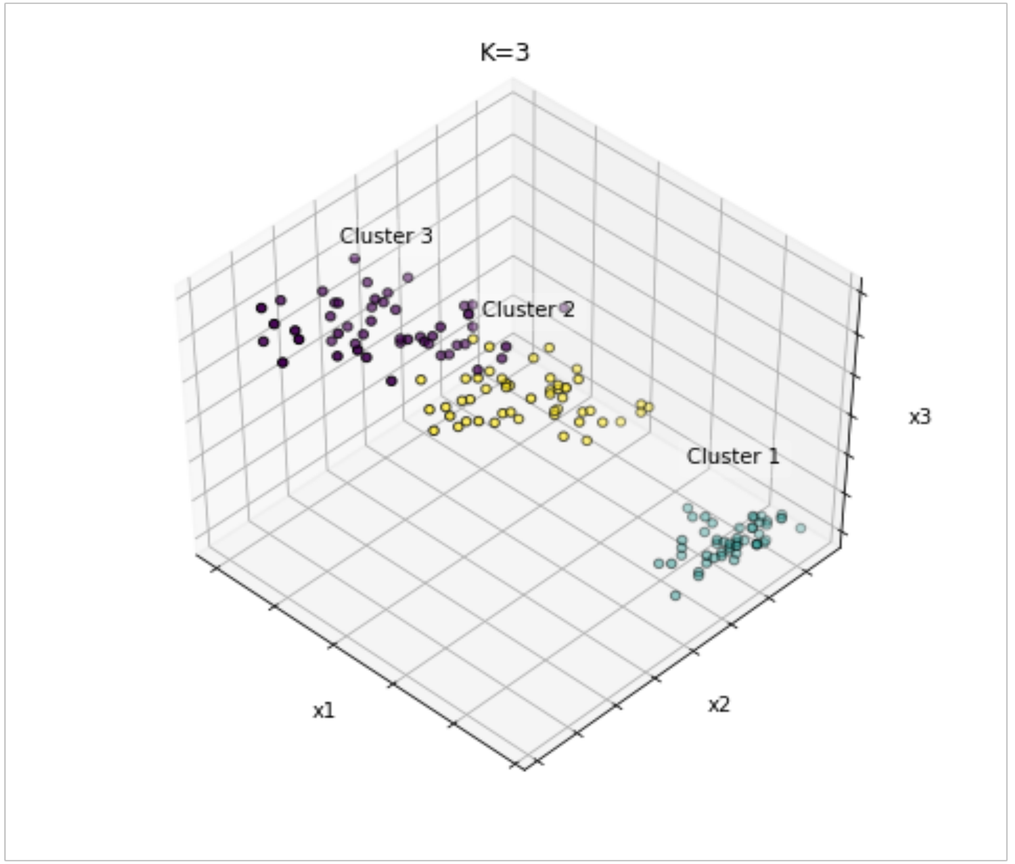
随机森林

随机森林（Random Forest）建立在决策树的基础上，以决策树为基学习器，将多个决策树集成起来，从而从“树”变成了“森林”。所谓的“随机”在于其在训练过程中随机划分特征，从候选特征中随机生成包含部分特征的子集；同时在训练机器学习时，采用有放回随机采样的方式训练，减少训练噪声。



K均值聚类

K均值聚类（K-means）是一种应用较广泛的聚类模型，属于无监督学习。K是人为设定的簇的个数，假设数据集合可以分为K类，那么其模型目标就是将样本划分到K个簇中，其中每个样本归属于距离最近的簇，利用训练数据来训练出这K个分类来。



谢谢！