数据挖掘算法与机器学习库

张敏

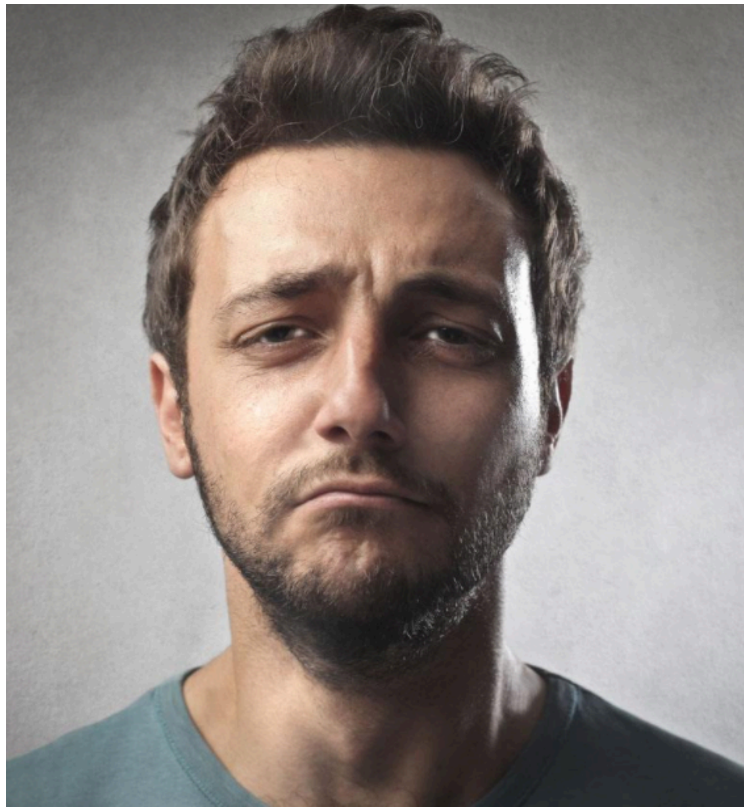
18/1/2

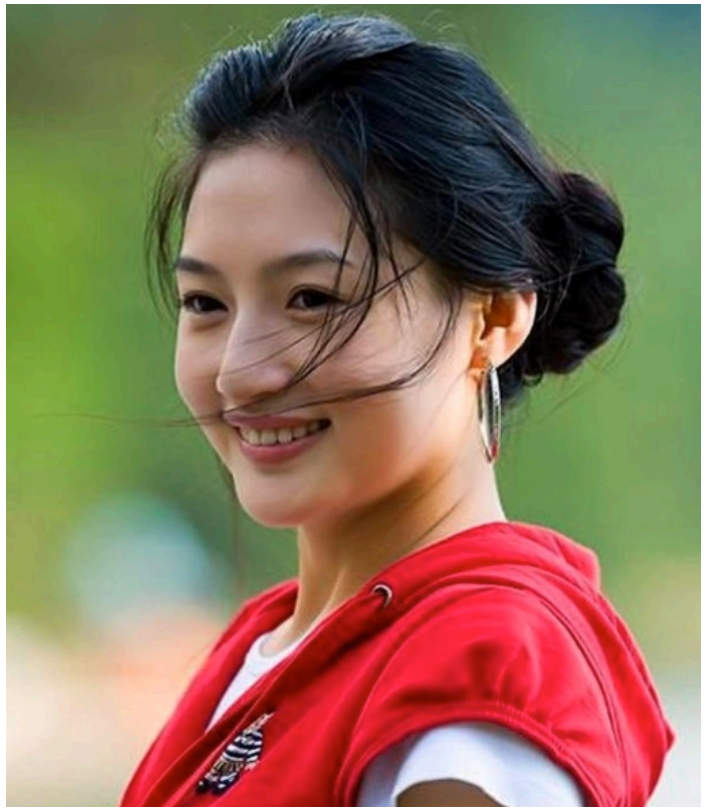
目录

数据挖掘介绍

引例

男、女？





数据挖掘介绍

引例

请分别讨论下列各组数据的内部关系，并填空。

x_1	3	1	7	2	4
y_1	4.5	2.5	8.5	3.5	?

$$y_1 = x_1 + 1.5$$

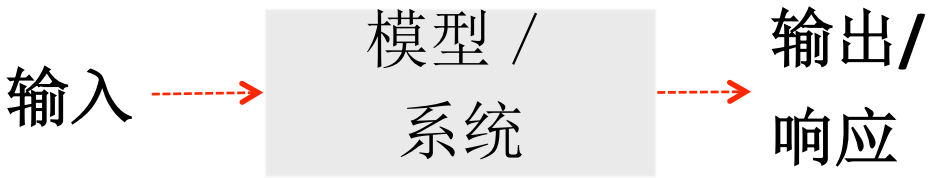
x_2	3	6	8	1	2
y_2	10.5	37.5	65.5	2.5	?

$$y_2 = x_2^2 + 1.5$$

x_1	2.0	6.0	5.0	1.0	4.0
x_2	7.0	9.0	3.0	2.0	5.0
y	52.8	96.7	21.2	6.0	?

$$y = x_1^{3/2} + x_2^2 + 1$$

数据挖掘介绍



输入：发型、喉结、胡须，已知数据对

输出 / 响应：男 / 女，对应y 值

$$y = f(x_1, x_2, x_3, x_4, x_5 \ldots\ldots)$$

数据挖掘介绍

数据挖掘企业项目

- 广播电视系统大数据营销推荐
- 网络舆情分析
- 电商产品评论数据情感分析
- 电力窃漏电用户自动识别
- 基于水色图像的水质评价
- 家用电器用户行为分析及事件识别
- 应用系统负载分析与容量预测
- 中医证型的关联规则挖掘

历届赛题

- 航空客运信息挖掘(2013年)
- 道路缺陷自动识别(2013年)
- 小车压双黄线自动检测(2014年)
- 面向网络舆情的关联度分析(2014年)
- 基于数据挖掘技术的市财政收入分析预测模型(2015年)
- 电商平台图片中文字的识别(2016年)
- 铁路旅客流量预测(2016年)
- 中央空调系统的数据分析与控制策略(2017年)

数据挖掘介绍

电力窃漏电用户自动识别

背景

- 全国每年因窃电造成的损失在200亿元左右，被查获的不足30%
- 深圳龙岗工业区一家只有两条生产线的小塑料包装厂，一年窃电折价就30 - 40万元
- 某市06年因窃电损失达4亿元
- 传统打击手段：突击检查

目标

- 如何通过监测数据自动识别偷漏电行为？

数据挖掘介绍

电子商务网站用户行为分析及服务推荐

背景

- 某法律网站是一家大型的法律资讯信息网站，它一直致力于为用户提供丰富的法律资讯信息与专业法律咨询服务，并为律师与律师事务所提供卓有成效的互联网整合营销解决方案。
- 大量的访问用户，每天上千万次的点击量，为其带来发展也带来瓶颈。
- 如何留住需要帮助的用户，快速找到其感兴趣的页面？并进一步为其推荐律师？

目标

- 客户行为分析
- 用户精准画像
- 智能推荐

数据挖掘介绍

数据挖掘常见任务 [R语言与数据挖掘]

- 分类与回归
- 聚类分析
- 智能推荐
- 自然语言处理 / 文本挖掘
- 关联规则
- 时间序列

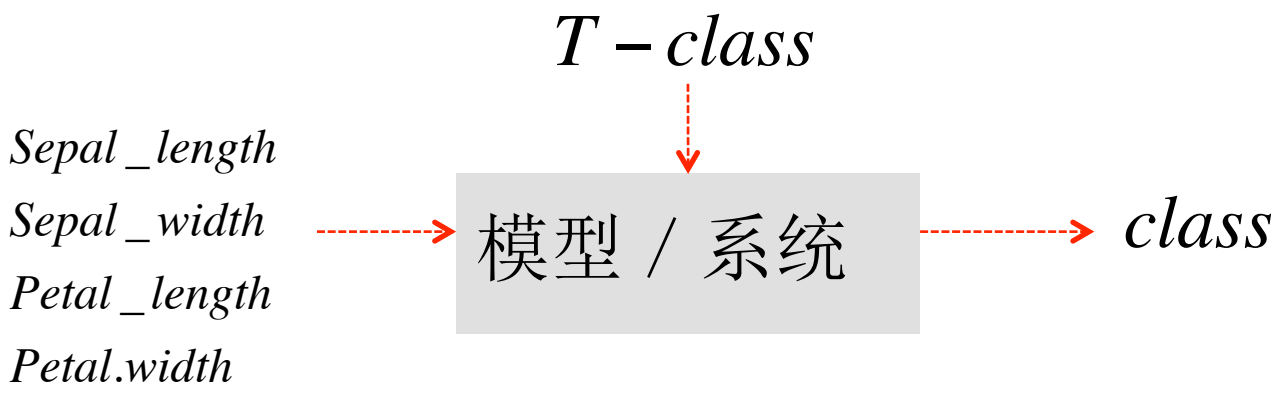
数据挖掘介绍

有监督与无监督

分类：学习 / 训练过程

有监督，训练样本有明确标签

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	T-class
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
7	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.3	3.3	6	2.5	virginica
5.8	2.7	5.1	1.9	virginica
6.5	3	5.8	2.2	?
6.2	2.9	4.3	1.3	?



数据挖掘介绍

有监督与无监督

聚类：学习 / 训练过程

无监督，样本无明确标签

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width
5.1	3.5	1.4	0.2
4.9	3	1.4	0.2
7	3.2	4.7	1.4
6.4	3.2	4.5	1.5
6.3	3.3	6	2.5
5.8	2.7	5.1	1.9
6.5	3	5.8	2.2
6.2	2.9	4.3	1.3

数据挖掘介绍

数据挖掘与机器学习算法

数据挖掘算法 [数据挖掘十大算法]

- C4.5、K-means、SVM、Apriori、EM、PageRank、AdaBoost、kNN、Naive Bayes、CART

机器学习算法 [机器学习]

- 线性回归、决策树、神经网络、支持向量机、贝叶斯分类器、集成学习、聚类、降维与度量学习、特征选择与稀疏学习、计算学习理论、半监督学习、概率图模型、规则学习、强化学习

目录

Scikit-learn

机器学习算法库scikit-learn简介

- scikit-learn是在NumPy，SciPy和matplotlib三个模块上编写的，是数据挖掘和数据分析的一个简单而有效的工具。
- 在其官方网站上我们可以看到scikit-learn有6大功能：
- 学习问题主要可以归为2类：

有监督学习	分类：样本属于两个或多个类别 回归：输出是一个或多个连续的变量
无监督学习	无监督学习的训练数据包括了输入向量X的集合，但没有相应的目标变量

Scikit-learn

安装和使用

- pip3 install sklearn
- **from** sklearn **import** …

API文档：

- http://scikit-learn.org/stable/modules/classes.html

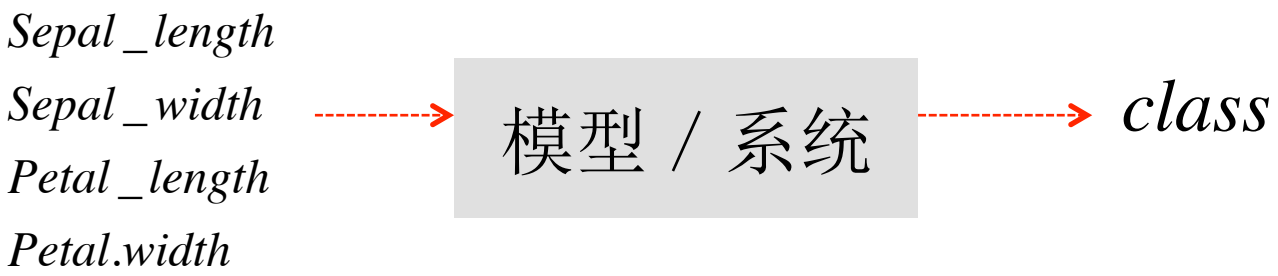
Scikit-learn

主要函数库

- sklearn.datasets: Datasets（数据集）
- sklearn.tree: Decision Trees（决策树）
- sklearn.neural_network: Neural network models（神经网络）
- sklearn.svm: Support Vector Machines（支持向量机）
- sklearn.cluster: Clustering（聚类分析）
- sklearn.naive_bayes: Naive Bayes（朴素贝叶斯）
- sklearn.neighbors：KNeighborsClassifier（KNN）
- sklearn.covariance: Covariance Estimators（协方差估计）
- sklearn.model_selection: Model Selection（模型选择）

Scikit-learn

Sepal.Length	Sepal.Width	Petal.Length	Petal.Width	class
5.1	3.5	1.4	0.2	setosa
4.9	3	1.4	0.2	setosa
7	3.2	4.7	1.4	versicolor
6.4	3.2	4.5	1.5	versicolor
6.3	3.3	6	2.5	virginica
5.8	2.7	5.1	1.9	virginica
6.5	3	5.8	2.2	?
6.2	2.9	4.3	1.3	?



Scikit-learn

任务一：实现对鸢尾花样本聚类、分类操作

神经网络

代码实现（sklearn）

- from sklearn.neural_network import MLPClassifier #导入神经网络包
- Net = MLPClassifier(hidden_layer_sizes=10,max_iter=1000).fit(tr_data.ix[:,0:6],tr_data.ix[:,6])
- res = Net.predict(te_data.ix[:,0:6])

决策树

代码实现（sklearn）

- from sklearn.tree import DecisionTreeClassifier
- modle = DecisionTreeClassifier(criterion='gini').fit(tr_data.ix[:,0:6],tr_data.ix[:,6]) #模型训练
- res = modle.predict(te_data.ix[:,0:6])

支持向量机

代码实现（sklearn）

- from sklearn.svm import LinearSVC　#导入支持向量机函数
- clf = LinearSVC(random_state=1).fit(tr_data.ix[:,:9],tr_data.ix[:,9])　#构建模型
- res = clf.predict(te_data.ix[:,:9])　　#模型预测

Thank you!

泰迪科技：www.tipdm.com
热线电话：40068-40020