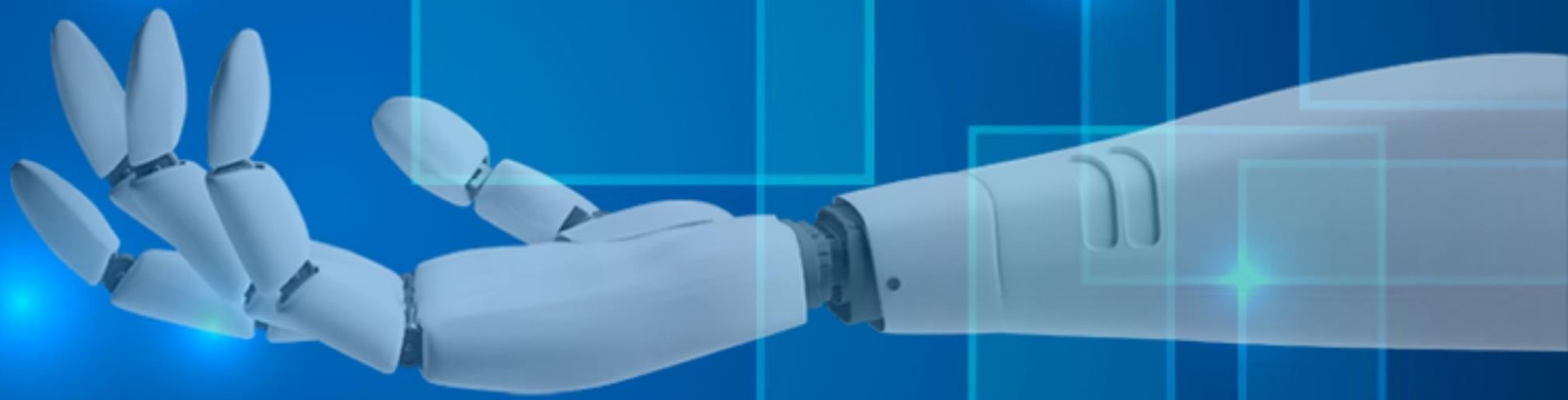
Tedu.cn 达内教育

达内教育研究院

# 数据分析建模

## 数据分析与建模



数据分析和数据挖掘领域，简单的分析可通过基本统计方法完成，需要复杂的建模就可以采用机器学习算法进行建模分析，机器学习就是通过学习来获得进行预测和判断的能力

机器学习方法的重要理论基础之一是统计学，基于统计学习理论，在自然语言处理、语音识别、图像识别、信息检索和生物信息等许多计算机领域获得了广泛应用。

分类：将实例数据划到合适的类别中，分类的目标变量一般是类别型的，一般分为二分类和多分类，

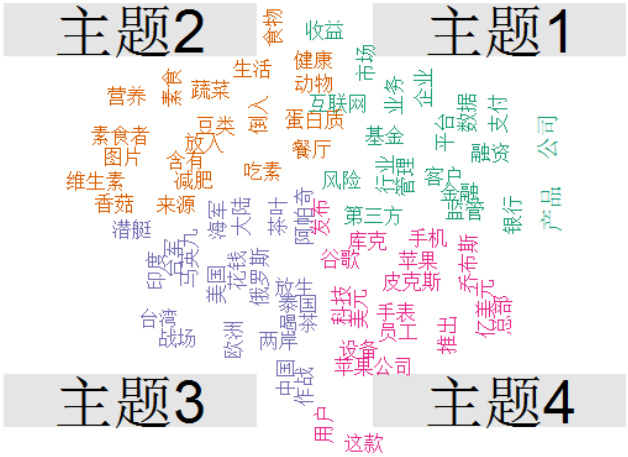
分类方法是机器学习领域使用最广泛的技术之一。分类是依据历史数据形成刻画事物特征的类标识，进而预测未来数据的归类情况。目的是学会一个分类函数或分类模型（也称作分类器），该模型能把数据集中的事物映射到给定类别中的某一个类。

例如：病人病情的判断，客户流失预测，邮件过滤，金融欺诈等

聚类：将对象的集合分成由类似的对象组成的多个类别的过程，这些对象与同一个簇中的对象彼此相似，与其他簇中的对象相异。在许多应用中，一个簇中的数据对象可作为一个整体来对待。

应用：细分客户、新闻聚类等

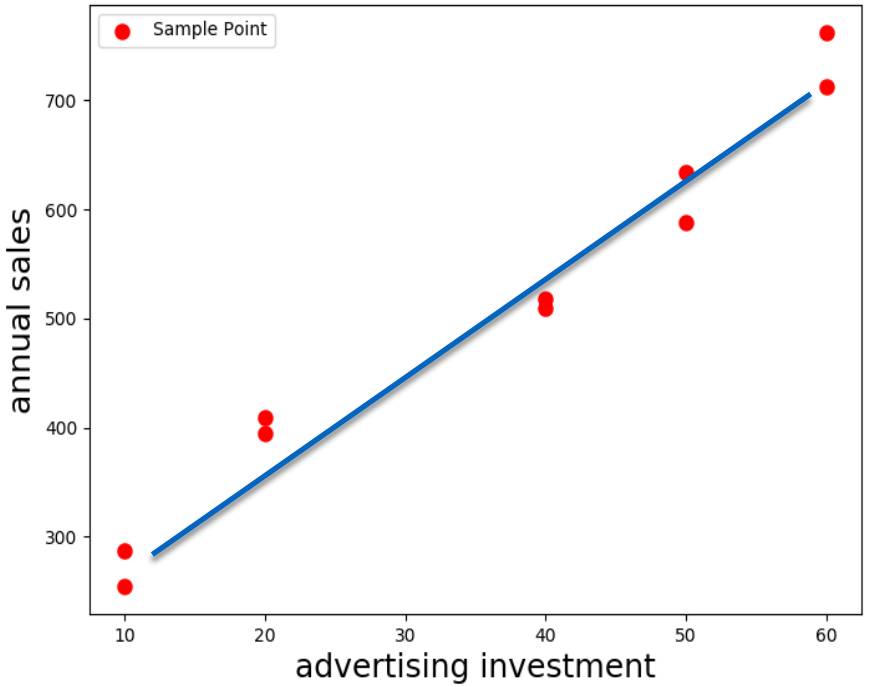




回归：回归是根据已有数值（行为）预测未知数值（行为）的过程，与分类模式分析不同，预测分析更侧重于“量化”。一般认为，使用分类方法预测分类标号（或离散值），使用回归方法预测连续或有序值

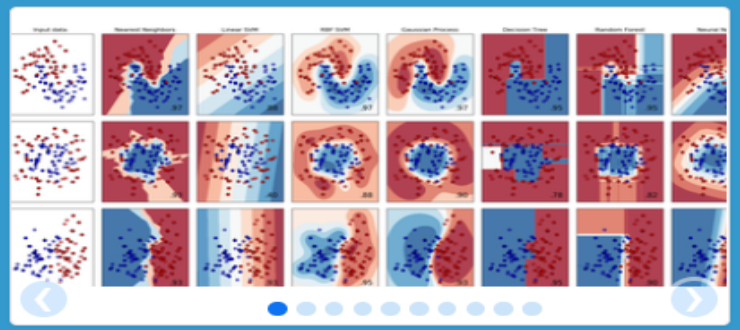
例如：股票走势，房价走势预测，

网站点击量预测等



scikit learn

Home Installation Documentation Examples



# scikit-learn

*Machine Learning in Python*

- Simple and efficient tools for data mining and data analysis
- Accessible to everybody, and reusable in various contexts
- Built on NumPy, SciPy, and matplotlib
- Open source, commercially usable - BSD license

## Classification

Identifying to which category an object belongs to.

**Applications**: Spam detection, Image recognition.
**Algorithms**: SVM, nearest neighbors, random forest, ... — Examples

## Regression

Predicting a continuous-valued attribute associated with an object.

**Applications**: Drug response, Stock prices.
**Algorithms**: SVR, ridge regression, Lasso, ... — Examples

## Clustering

Automatic grouping of similar objects into sets.

**Applications**: Customer segmentation, Grouping experiment outcomes
**Algorithms**: k-Means, spectral clustering, mean-shift, ... — Examples

## Dimensionality reduction

Reducing the number of random variables to consider.

**Applications**: Visualization, Increased efficiency
**Algorithms**: PCA, feature selection, non-negative matrix factorization. — Examples

## Model selection

Comparing, validating and choosing parameters and models.

**Goal**: Improved accuracy via parameter tuning
**Modules**: grid search, cross validation, metrics. — Examples

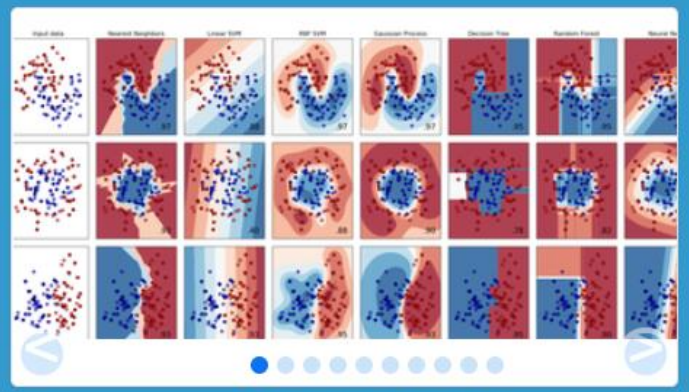
## Preprocessing

Feature extraction and normalization.

**Application**: Transforming input data such as text for use with machine learning algorithms.
**Modules**: preprocessing, feature extraction. — Examples

scikit learn

家 安装 文档 例子



# scikit学习

*Python中的机器学习*

- 简单有效的数据挖掘和数据分析工具
- 可供所有人访问，并可在各种环境中重复使用
- 基于NumPy，SciPy和matplotlib构建
- 开源，商业上可用 - BSD许可证

## 分类

确定对象属于哪个类别。

**应用**：垃圾邮件检测，图像识别。
**算法**： SVM，最近邻居，随机森林，.....

— 例子

## 回归

预测与对象关联的连续值属性。

**应用**：药物反应，股票价格。
**算法**： SVR，岭回归，套索，.....

— 例子

## 聚类

将类似对象自动分组为集合。

**应用程序**：客户细分，分组实验结果
**算法**： k-Means，谱聚类，均值漂移，.....

— 例子

## 维度降低

减少要考虑的随机变量的数量。

**应用**：可视化，提高效率
**算法**： PCA，特征选择，非负矩阵分解。

— 例子

## 型号选择

比较，验证和选择参数和模型。

**目标**：通过参数调整提高准确性
**模块**： 网格搜索，交叉验证，指标。

— 例子

## 预处理

特征提取和规范化。

**应用程序**：转换输入数据（如文本）以用于机器学习算法。
**模块**： 预处理，特征提取。

— 例子

# 加载datasets模块中数据集





**datasets模块常用数据集加载函数及其解释**

- sklearn库的datasets模块集成了部分数据分析的经典数据集，可以使用这些数据集进行数据预处理，建模等操作，熟悉sklearn的数据处理流程和建模流程。
- datasets模块常用数据集的加载函数与解释如下表所示。
- 加载后的数据集可以视为一个字典，几乎所有的sklearn数据集均可以使用data，target，feature_names，DESCR分别获取数据集的数据，标签，特征名称和描述信息。

| 数据集加载函数 | 数据集任务类型 | 数据集加载函数 | 数据集任务类型 |
|---|---|---|---|
| load_ boston | 回归 | load_breast_cancer | 分类，聚类 |
| fetch_california_housing | 回归 | load_iris | 分类，聚类 |
| load_digits | 分类 | load_wine | 分类 |

# 加载datasets模块中数据集





数据集总览

| 数据集大小 | 数据集名称 | 调用方式 | 适用算法 | 数据规模 |
| --- | --- | --- | --- | --- |
| 小数据集 | 波士顿房价数据集 | load_boston() | 回归 | 506*13 |
| – | 鸢尾花数据集 | load_iris() | 分类 | 150*4 |
| – | 糖尿病数据集 | load_diabetes() | 回归 | 442*10 |
| – | 手写数字数据集 | load_digits() | 分类 | 5620*64 |
| 大数据集 | Olivetti脸部图像数据集 | fetch_olivetti_faces() | 降维 | 400 * 64 * 64 |
| – | 新闻分类数据集 | fetch——20newsgroups（） | 分类 | - |
| – | 带标签的人脸数据集 | fetch_lfw_people() | 分类；降维 | - |
| – | 路透社新闻语料数据集 | fetch_revl() | 分类 | 804414*47236 |

# 将数据集划分为训练集和测试集



**常用划分方式**

- 在数据分析过程中，为了保证模型在实际系统中能够起到预期作用，一般需要将样本分成独立的两部分：
  - 训练集（train set）：用于训练模型。
  - 测试集（test set）：用于检验最优的模型的性能。

# 将数据集划分为训练集和测试集



## train_test_split函数

- sklearn的model_selection模块提供了train_test_split函数，能够对数据集进行拆分，其使用格式如下

*sklearn.model_selection.**train_test_split**(*arrays, **options)*

| 参数名称 | 说明 |
| --- | --- |
| *arrays | 接收一个或多个数据集。代表需要划分的数据集，若为分类回归则分别传入数据和标签，若为聚类则传入数据。无默认。 |
| test_size | 接收float，int，None类型的数据。代表测试集的大小。如果传入的为float类型的数据则需要限定在0-1之间，代表测试集在总数中的占比；如果传入为int类型的数据，则表示测试集记录的绝对数目。该参数与train_size可以只传入一个。在0.21版本前，若test_size和train_size均为默认则testsize为25%。 |
| train_size | 接收float，int，None类型的数据。代表训练集的大小。该参数与test_size可以只传入一个。 |
| random_state | 接收int。代表随机种子编号，相同随机种子编号产生相同的随机结果，不同的随机种子编号产生不同的随机结果。默认为None。random_state就是为了保证程序每次运行都分割一样的训练集合测试集 |
| shuffle | 接收boolean。代表是否进行有放回抽样。若该参数取值为True则stratify参数必须不能为空。 |

# 将数据集划分为训练集和测试集



**train_test_split函数**

- train_test_split函数根据传入的数据，分别将传入的数据划分为训练集和测试集。
- 如果传入的是1组数据，那么生成的就是这一组数据随机划分后训练集和测试集，总共2组。如果传入的是2组数据，则生成的训练集和测试集分别2组，总共4组。

# 使用sklearn转换器进行数据预处理



## sklearn转换器三个方法

- sklearn把相关的功能封装为转换器（transformer）。

| 方法名称 | 说明 |
| --- | --- |
| fit | fit方法主要通过分析特征和目标值，提取有价值的信息，这些信息可以是统计量，也可以是权值系数等。 |
| transform | transform方法主要用来对特征进行转换。 |
| fit_transform | fit_transform方法就是先调用fit方法，然后调用transform方法。 |

# 使用sklearn转换器进行数据预处理与降维



## sklearn转换器

- 在数据分析过程中，各类特征处理相关的操作都需要对训练集和测试集分开操作，需要将训练集的操作规则，权重系数等应用到测试集中。
- 如果使用pandas，则应用至测试集的过程相对烦琐，使用sklearn转换器可以解决这一困扰。

以理服人

# 使用sklearn转换器进行数据预处理与降维



以理服人

**sklearn部分预处理函数与其作用**

| 函数名称 | 说明 |
| --- | --- |
| MinMaxScaler | 对特征进行离差标准化。 |
| StandardScaler | 对特征进行标准差标准化。 |

# 使用sklearn估计器构建模型

Tedu.cn 达内教育

以理服人

## sklearn估计器

- 聚类算法实现需要sklearn估计器（estimator）。sklearn估计器和转换器类似，拥有fit和predict两个方法。两个方法的作用如下。

| 方法名称 | 说明 |
| --- | --- |
| fit | fit方法主要用于训练算法。该方法可接收用于有监督学习的训练集及其标签两个参数，也可以接收用于无监督学习的数据。 |
| predict | predict用于预测有监督学习的测试集标签，亦可以用于划分传入数据的类别。 |

# 使用sklearn中的聚类算法





sklearn常用的聚类算法模块cluster提供KMeans聚类算法：

| 函数名称 | 参数 | 适用范围 | 距离度量 |
| --- | --- | --- | --- |
| KMeans | 簇数 | 可用于样本数目很大，聚类数目中等的场景。 | 点之间的距离 |

Kmeans构建的模型返回参数：

kmeans.labels_：返回每个元素聚类后的类别

kmeans.cluster_centers_：返回类别的聚类中心

# 谢谢