首都师范大学

人工智能数学基础Python实践
第八章 概率估计

张苗苗

信息工程学院

主要内容

- 在Python中实现以下操作：
  - 最大似然估计
  - 最大后验估计(提高)

- 所需的python库：
  - scipy
  - pandas
  - scikit-learn

Pandas库

- Pandas是一个强大的分析结构化数据的工具集；Pandas 提供了方便的类表格和类SQL的操作，同时提供了强大的缺失值处理方法，可以方便的进行数据导入、选取、清洗、处理、合并、统计分析等操作。
- 与Numpy相比，Pandas更适合做数据的预处理，而numpy更适合做数据的运算。

Pandas库

Pandas中常用方法

加载数据的方法（支持大多数文件格式）

- **read_excel()**：从excel文件中读取数据；
- **read_csv()**：从csv文件中读取数据；
- **read_clipboard()**：从剪切板中数据；
- **read_html()**：从网页中读取数据；
- **read_json()**：从 json 格式文本中读取数据；
- **read_pickle()**：从pickle文件中读取数据；
- **……**

具体使用时，查看相应文档中各参数的说明。

读取Excel文件的核心参数

- **io**：**文件路径**，可以是本地文件也可以是网络文件，支持**xls**、**xlsx**、**xlsm**等格式；
- **sheet_name**：**表单序号或名称**，可以是一个列表，**同时读取多个表单**，默认为第一个表单；
- **hearder**：**表头**，可以是整数或整数列表；
- **names**：**指定列名**；
- **index_col**：**索引列**，可以是整数或整数列表；
- **usecols**：**使用到的列**；
- **dtype**：指定**每一列的数据类型**；
- **skiprows**：**跳过多少行**；
- **nrows**：**解析多少行**；
- **na_values**：指定**哪些值被看做是缺失值**；
- **……**

练习1-- 最大似然法求解模型参数

根据数据集中搜集到的样本数，利用最大似然法估计总体分布的模型参数：

观察数据

```
import pandas as pd
import matplotlib.pyplot as plt
messages = pd.read_csv('data/QQ_data.csv')  #读取数据
fig = plt.figure(figsize=(12,5))
plt.title('Frequency of QQmessages')
plt.xlabel('Number of QQmessages')
plt.ylabel('Frequency')
plt.hist(messages['numbers'].values,range=[0, 60], bins=60,
histtype='stepfilled')  #画直方图
plt.show()
```

练习1-- 最大似然法求解模型参数

确定模型，求解

```
import numpy as np
import pandas as pd
import scipy.stats as stats
import scipy.optimize as opt
messages = pd.read_csv('data/QQ_data.csv') #读取数据
y_obs = messages['numbers'].values
np.seterr(invalid='ignore') #设置浮点错误的处理方式
def poisson_logprob(mu, sign=-1):
    # 根据泊松模型和参数值返回观测数据的总似然值
    return np.sum(sign*stats.poisson.logpmf(y_obs, mu=mu))
# 求解，最小化一个变量的标量函数
freq_results = opt.minimize_scalar(poisson_logprob)
print("参数 mu 的估计值: %s" % freq_results['x'])
```

Pandas库中DataFrame数据类型

DataFrame

DataFrame 可以看作是一种**既有行索引，又有列索引的二维数组**，类似于Excel表或关系型数据库中的二维表，是**Pandas中最常用的基本结构**。

DataFrame的创建

- 可通过**值为一维ndarray，list, dict 或者Series的字典或列表**；**二维的ndarray**；**单个Series、列表、一维数组**；**其他的DataFrame**等创建；
- 创建DataFrame时，可通过 **index** 和 **columns** 参数指定 **行索引** 和 **列索引** ，若没有指定索引，则**默认为从0开始的连续数字**；
- 通过多个Series创建DataFrame时，多个Series对象会自动对齐。**若指定了index** ，则会**丢弃所有未和 index匹配的数据** 。如果指定的索引不存在，则**对应的值默认为NaN**。

Pandas库中DataFrame数据类型

DataFrame中常见属性和方法

DataFrame 常见属性

- shape：获取**形状信息**，结果为一个**元组**；
- dtypes：获取**各字段的数据类型**，结果为**Series**；
- values：获取**数据内容**，结果通常为**二维数组**；
- columns：获取**列索引**，即字段名称，结果为**Index**；
- index: 行索引，即行的标签，结果为**Index**。
- axes：同时获取行和列索引，结果为**Index的列表**；

DataFrame 常见方法

方法	说明
info()	**显示基本信息**，包括行列索引信息、每列非空元素数量，每列数据的类型，整体所占内存大小等
head(n)	**获取前n行数据，n默认为5**，结果为**DataFrame**
tail(n)	**获取后n行数据，n默认为5**，结果为**DataFrame**
describe()	数据的整体描述信息，包括：非空值数量、平均值、标准差、最小值、最大值等，结果为**DataFrame**
count()	统计各列中非空值的数量，结果为**Series**
sample(n, axis)	随机从数据中**按行或列抽取n行或n列**
apply(fun, axis)	对**每一行或每一列元素执行函数**
applymap(fun)	对**每一个数据执行函数**
to_dict()	**转化为dict类型对象**，可指定字典中值的类型，如list
to_excel(文件名)	将数据**保存到Excel文件**中去

Pandas库中DataFrame数据类型

DataFrame中常见方法

方法	说明
sort_values(by)	**根据值进行排序**，可以指定一列或多列，返回新的对象
sort_index()	根据索引进行排序，**原始索引不一定有序**，返回新的对象
rank()	对每一列的值进行排名，从小到大，从1开始
isna()、isnull()	对每一个元素判断**是否为缺失值**
dropna()	**删除缺失值**，可指定删除行或列、缺失值满足的条件等
fillna(value)	**用value值填充空值**，返回新的对象
rename()	**重命名，通过columns对列索引重命名，index对行索引重命名**
set_index()	**设置索引列**，可以用一个已有列名作为索引，返回新的对象
groupby()	**对数据进行分组**，例如根据某列或多列进行分组
d_1.append(d_2)	将d_2中的行添加到d_1的后面，会**自动对齐**，没有内容的部分**默认为NaN**
sum()、mean()、max()、min()、median()、std()、var()	对每一列数据**求和**、**求平均数**、**最大值**、**最小值**、**中位数**、标准差、方差
nunique()	统计每一列中**不重复的元素个数**

Scikit-learn库

- Scikit-learn库是一个通用型开源机器学习库，它几乎涵盖了所有的机器学习算法，并且搭建了高效的数据挖掘的框架。

Scikit-learn库中朴素贝叶斯库概述

- 朴素贝叶斯是一类比较简单的算法，scikit-learn中朴素贝叶斯类库的使用也比较简单。相对于决策树，KNN之类的算法，朴素贝叶斯需要关注的参数是比较少的，这样也比较容易掌握。

- 在scikit-learn中，一共有3个朴素贝叶斯的分类算法类。分别是GaussianNB，MultinomialNB和BernoulliNB。其中GaussianNB就是先验为高斯分布的朴素贝叶斯，MultinomialNB就是先验为多项式分布的朴素贝叶斯，而BernoulliNB就是先验为伯努利分布的朴素贝叶斯。

Scikit-learn库中朴素贝叶斯库概述

- 这三个类适用的分类场景各不相同
  - 一般来说，如果样本特征的分布大部分是连续值，使用GaussianNB会比较好。
  - 如果样本特征的分布大部分是多元离散值，使用MultinomialNB比较合适。
  - 而如果样本特征是二元离散值或者很稀疏的多元离散值，应该使用BernoulliNB。

练习2-- 朴素贝叶斯进行用户满意度预测

P209-综合实例1

汽车测评数据

根据汽车的属性（买入价、维护费、车门数、可容纳人数、后备箱大小、安全性），测评用户满意度（不可接受，可接受，好，非常好）

练习2-- 朴素贝叶斯进行用户满意度预测

步骤：

1. 准备数据
   - 从数据集中获取数据
   - 将数据集分成训练集和测试集
2. 创建朴素贝叶斯模型
3. 利用训练集数据对模型进行训练
4. 利用训练好的模型进行预测

练习2-- 朴素贝叶斯进行用户满意度预测

```
clf = BernoulliNB()          #伯努利朴素贝叶斯分布

clf.fit(train_X, train_Y)    #训练模型

predicted = clf.predict(test_X)    #预测

np.mean(predicted == test_Y)    #查看正确率
```

首都师范大学

谢谢！