01.关于数据的整理

机器学习以数据为主导，当做数据科学的工作的时候，我们第一步需要的是整理数据， 本次动手实验会结合开源，结合.NET 的数据整理工具进行。

一 .NumSharp

在Python我们经常用Numpy做数据处理

什么是Numpy?

NumPy(Numerical Python) 是 Python 语言的一个扩展程序库，支持大量的维度数组与矩阵运算，此外也针对数组运算提供大量的数学函数库。

NumPy（Numeric Python）提供了许多高级的数值编程工具，如：矩阵数据类型、矢量处理，以及精密的运算库。专为进行严格的数字处理而产生。多为很多大型金融公司使用

在.NET Core中 ，有很好的原生第三方库 NumSharp 。https://github.com/SciSharp/NumSharp

```
#r "nuget: NumSharp"
```

```
using NumSharp;
```

```
var study_score_list = new float[]{50.0f, 50.0f, 47.0f ,97.0f ,49.0f ,3.0f ,53.0f
,42.0f ,26.0f ,74.0f ,82.0f ,62.0f ,37.0f ,15.0f ,70.0f ,27.0f ,36.0f ,35.0f
,48.0f ,52.0f ,63.0f ,64.0f };
var numpy_array = np.array(study_score_list);
```

```
foreach(var item in numpy_array)
{
    Console.WriteLine(item.ToString());
}
```

```
numpy_array.shape
```

```
numpy_array[0]
```

```
numpy_array.mean()
```

```
float[,] study_datas_array = new float[2,22]{{10.0f ,11.5f ,9.0f ,16.0f , 9.25f ,
1.0f ,11.5f ,9.0f ,8.5f ,14.5f ,15.5f,
    13.75f ,9.0f ,8.0f, 15.5f ,8.0f , 9.0f , 6.0f ,10.0f ,12.0f ,12.5f,12.0f },
{50.0f, 50.0f, 47.0f ,97.0f ,49.0f ,3.0f ,53.0f ,42.0f ,26.0f ,74.0f ,82.0f ,62.0f
,37.0f ,15.0f ,70.0f ,27.0f ,36.0f ,35.0f ,48.0f ,52.0f ,63.0f ,64.0f}};
```

```
NDArray study_datas_numpy = np.array(study_datas_array);
```

```
study_datas_numpy.shape
```

```
study_datas_numpy[0][0]
```

```
var avg_study = study_datas_numpy[0].mean();
var avg_grade = study_datas_numpy[1].mean();
```

```
avg_study
```

```
avg_grade
```

2. 通过.NET Core中的DataFrame对数据进行整理

再来看看Python , 当我们整理数据过程中，除了数据一些格式转换外，需要检查数据，找出是否有空值，有什么数据特征，一般用pandas解决。

pandas 是基于NumPy 的一种工具，该工具是为解决数据分析任务而创建的。Pandas 纳入了大量库和一些标准的数据模型，提供了高效地操作大型数据集所需的工具。pandas提供了大量能使我们快速便捷地处理数据的函数和方法。

在.NET Core 中有原生对应的方案

.NET 的 DataFrame 类型，使数据探索变得容易。

```
#r "nuget:Microsoft.Data.Analysis"
```

```
using Microsoft.Data.Analysis;
```

```
StringDataFrameColumn nameList = new StringDataFrameColumn("Strings");
PrimitiveDataFrameColumn<float> timelist = new PrimitiveDataFrameColumn<float>
("StudyTime");
PrimitiveDataFrameColumn<float> scorelist = new PrimitiveDataFrameColumn<float>
("StudyScore");
```

```
var student_list = new string[]{"Dan", "Joann", "Pedro", "Rosie", "Ethan",
"Vicky", "Frederic", "Jimmie", "Rhonda", "Giovanni", "Francesca", "Rajab",
"Naiyana", "Kian",
"Jenny","Jakeem","Helena","Ismat","Anila","Skye","Daniel","Aisha" };
```

```
student_list.Length
```

```
for(var i = 0 ; i< student_list.Length ;i++)
{
    nameList.Append(student_list[i]);
}
```

```
for(var i = 0 ; i< study_datas_numpy.shape[1] ;i++)
{
   timelist.Append(study_datas_numpy[0][i]);
   scorelist.Append(study_datas_numpy[1][i]);
}
```

```
nameList
```

```
DataFrame df = new DataFrame(nameList ,timelist, scorelist);
```

```
df
```

```
df.Info()
```

```
var order = df.OrderBy("StudyScore");
```

```
order
```

```
var find_score = df["StudyScore"].ElementwiseEquals(74);
```

```
find_score
```

```
var score = df.Filter(find_score);
```

```
score
```

三.数据图表工具XPlot

在Python中，Matplotlib是一个2D绘图库，它以各种硬拷贝格式和跨平台的交互式环境生成出版质量级别的图形。通过Matplotlib，开发者可以仅需要几行代码，便可以生成绘图。一般可绘制折线图、散点图、柱状图、饼图、直方图、子图等等。Matplotlib使用Numpy进行数组运算，并调用一系列其他的Python库来实现硬件交互。

而对应.NET Core有XPlot https://fslab.org/XPlot/

```
#i "nuget:https://pkgs.dev.azure.com/dnceng/public/_packaging/dotnet-
tools/nuget/v3/index.json"
#r "nuget:xplot.plotly.interactive"
```

```
using XPlot.Plotly;
using System.Linq;
```

```
IList<float> data_list=new List<float>(){50.0f, 50.0f, 47.0f ,97.0f ,49.0f ,3.0f
,53.0f ,42.0f ,26.0f ,74.0f ,82.0f ,62.0f ,37.0f ,15.0f ,70.0f ,27.0f ,36.0f
,35.0f ,48.0f ,52.0f ,63.0f ,64.0f};
```

```
var chart = new PlotlyChart();


Chart.Bar(data_list).Show();
```

```
chart
```

```
display(Chart.Bar(data_list));
```