In the name of God

استاد مربوطه: سرکار خانم استاد ثریا عنایتی

موضوع data science

اعضای گروه: نگین رحیمی، فاطمه مهران پور، یاسمن جعفری پور

علم داده‌ها یا "دیتا ساینس (Data Science) "مفهومی است که به ترکیبی از روش‌های علمی، فناوری و الگوریتم‌های ریاضی برای استخراج اطلاعات از داده‌ها و درک الگوها و روابط موجود در آنها اشاره دارد. داده‌ها به صورت گسترده در حوزه‌های مختلفی مانند علوم پایه، بهداشت، تجارت، حمل و نقل، تحقیقات بازار، اجتماعی و غیره تولید می‌شوند.

در علم داده‌ها، از تکنیک‌ها و روش‌های مختلفی مانند استخراج داده، پردازش داده، تحلیل داده، مدلسازی، پیش‌بینی و تفسیر داده‌ها استفاده می‌شود. با استفاده از این روش‌ها، اطلاعات مفید و قابل استفاده از داده‌ها استخراج می‌شود تا بتوان به تصمیم‌گیری‌های بهتری در مورد مسائل مختلف پرداخت.

علم داده‌ها شامل چندین حوزه کاربردی است که شامل تحلیل داده‌ها، هوش مصنوعی، یادگیری ماشین، استخراج اطلاعات، بینایی ماشین، پردازش زبان طبیعی و غیره می‌شود .این حوزه‌ها با استفاده از روش‌های آماری و محاسباتی، به دست‌آوردن اطلاعات جدید و قابل استفاده از داده‌ها کمک می‌کنند.

علم داده‌ها برای تجزیه و تحلیل داده‌ها و ارائه الگوها و روابط موجود در آنها، استفاده از ابزارها و فنون مختلفی از جمله برنامه‌نویسی، آمار، الگوریتم‌های ریاضی، پایگاه داده و غیره را نیازمند است. همچنین، توانایی تفسیر و تبدیل داده‌ها به اطلاعات مفید و قابل فهم برای تصمیم‌گیری‌های استراتژیک نیز در این حوزه بسیار مهم است.
سینتکس در علم داده‌ها به قواعد و ساختارهایی اشاره دارد که برای برنامه‌نویسی و تحلیل داده‌ها استفاده می‌شود. این قواعد شامل نحوه نوشتن دستورات و توابع، تعریف متغیرها و اشاره به داده‌ها و غیره است. در علم داده‌ها، سینتکس معمولاً با استفاده از زبان‌های برنامه‌نویسی مانند پایتون، R، جاوا و غیره تعریف می‌شود.

کتابخانه‌ها در علم داده‌ها مجموعه‌ای از توابع، الگوریتم‌ها و ابزارهایی هستند که برای انجام عملیات مربوط به داده‌ها استفاده می‌شوند. این کتابخانه‌ها توسط جامعه علمی و برنامه‌نویسان توسعه داده می‌شوند و شامل ابزارهای گوناگونی مانند پردازش داده، تجزیه و تحلیل داده، مدلسازی، تصویرسازی و غیره هستند .برخی از معروف‌ترین کتابخانه‌های مورد استفاده در علم داده‌ها شامل NumPy ، Pandas، Matplotlibو Scikit-learn در پایتون و ggplot2 و dplyr در R می‌باشند. این کتابخانه‌ها به برنامه‌نویسان و تحلیلگران داده امکانات قدرتمندی را برای کار با داده‌ها فراهم می‌کنند و به سادگی و کارآیی در تحلیل و استخراج اطلاعات از داده‌ها کمک می‌کنند.
یک مثال از استفاده از داده‌ها و کتابخانه‌های علم داده در دیتا ساینس می‌تواند مربوط به پیش‌بینی قیمت خودرو با استفاده از مجموعه‌ای از داده‌های مربوط به خودروها و ویژگی‌های آنها باشد.

برای این منظور، می‌توان از کتابخانه Pandas در پایتون استفاده کرده و داده‌ها را در قالب یک جدول (DataFrame) ذخیره کرده و عملیات مورد نیاز را روی آن انجام داد. سپس با استفاده از کتابخانهScikit-learn ، مدلی را برای پیش‌بینی قیمت خودرو آموزش داده و از آن برای پیش‌بینی قیمت خودروهای جدید استفاده کرد.

به عنوان مثال، با استفاده از داده‌های مربوط به خودروها شامل ویژگی‌های مانند سال تولید، قدرت موتور، حجم موتور و ...، می‌توان با استفاده از الگوریتم‌های یادگیری ماشین مانند رگرسیون خطی، یک مدل را آموزش داده و با ورودی دادن ویژگی‌های جدید، قیمت خودرو را پیش‌بینی کرد.

این مثال نشان می‌دهد که با استفاده از داده‌ها و کتابخانه‌های علم داده، می‌توان به راحتی و با دقت بالا، پیش‌بینی‌ها و تحلیل‌های مربوط به داده‌ها را انجام داد.

کتابخانه های data science

1. NumPy

NumPyکه مخفف Numerical Python است، یک کتابخانه قدرتمند برای عملیات عددی در پایتون است. این کتابخانه ابزارهایی را برای کار با آرایه‌ها و ماتریس‌ها در پایتون فراهم می‌کند و عملیات‌های ریاضی و عددی پیشرفته را امکان‌پذیر می‌سازد.

با استفاده ازNumPy ، می‌توانید آرایه‌های چند بعدی را ایجاد و مدیریت کنید و عملیات ماتریسی و برداری را انجام دهید. همچنین، NumPyابزارهایی برای تولید داده‌های تصادفی، تبدیل داده‌ها، محاسبات آماری، فیلتر کردن داده‌ها و بسیاری از عملیات دیگر را فراهم می‌کند.

برای استفاده از کتابخانهNumPy ، ابتدا باید آن را نصب کنید. معمولاً با استفاده از مدیر بسته pip می‌توانید NumPy را نصب کنید. برای نصبNumPy ، دستور زیر را در خط فرمان وارد کنید:

```
pip install numpy
```

بعد از نصب، می‌توانید NumPy را در برنامه خود وارد کنید با استفاده از دستور زیر:

```
python
import numpy as np
```

حالا می‌توانید از توابع و ابزارهای موجود در کتابخانه NumPy استفاده کنید. به عنوان مثال، می‌توانید آرایه‌ها را ایجاد و عملیات ریاضی روی آن‌ها انجام دهید:

```
python
import numpy as np

#ایجاد یک آرایه یک بعدی
arr1 = np.array([1, 2, 3, 4, 5])

#ایجاد یک آرایه دو بعدی
arr2 = np.array([[1, 2, 3], [4, 5, 6]])

#جمع دو آرایه
result = arr1 + arr2

print(result)
```

این کد آرایه arr1 را با آرایه arr2 جمع می‌کند و نتیجه را در متغیر result ذخیره می‌کند. سپس نتیجه را چاپ می‌کند.

کتابخانه NumPyدارای مستندات جامعی است که شما می‌توانید از آن استفاده کنید تا با توابع و ابزارهای مختلف آشنا شوید. همچنین، بسیاری از منابع آموزشی و ویدئوها در دسترس هستند که می‌توانند به شما در یادگیری NumPy کمک کنند.

2. Pandas

کتابخانه Pandasیک کتابخانه قدرتمند برای تحلیل داده‌ها در پایتون است. این کتابخانه ابزارهایی را برای خواندن و نوشتن داده‌ها از منابع مختلف مانند فایل‌های CSV وExcel ، ایجاد و مدیریت داده‌های ساختار یافته (مانند جداول و دیتافریم‌ها)، تحلیل و تغییر داده‌ها، و انجام عملیات آماری و محاسبات مربوط به داده‌ها را فراهم می‌کند.

با استفاده ازPandas ، می‌توانید داده‌های خام را به ساختارهای داده‌ای مانند جداول تبدیل کنید و به راحتی با آن‌ها کار کنید. همچنین، Pandasامکاناتی برای ترکیب، تفکیک و تغییر داده‌ها، فیلتر کردن و مرتب‌سازی داده‌ها، انجام عملیات آماری و تحلیلی پیشرفته، و بسیاری از عملیات دیگر را فراهم می‌کند.

برای استفاده از کتابخانهPandas ، ابتدا باید آن را نصب کنید. معمولاً با استفاده از مدیر بسته pip می‌توانید Pandas را نصب کنید. برای نصبPandas ، دستور زیر را در خط فرمان وارد کنید:

```
pip install pandas
```

بعد از نصب، می‌توانید Pandas را در برنامه خود وارد کنید با استفاده از دستور زیر:

```
python
import pandas as pd
```

حالا می‌توانید از توابع و ابزارهای موجود در کتابخانه Pandas استفاده کنید. به عنوان مثال، می‌توانید یک فایل CSV را بخوانید و داده‌های آن را در یک دیتافریم ذخیره کنید:

```
python
import pandas as pd

# خواندن فایل CSV
data = pd.read_csv('data.csv')

# نمایش اولین چند ردیف از داده‌ها
print(data.head())
```

این کد فایل CSV به نام data.csv را می‌خواند و داده‌های آن را در یک دیتافریم ذخیره می‌کند. سپس اولین چند ردیف از داده‌ها را چاپ می‌کند.

کتابخانه Pandasدارای مستندات جامعی است که شما می‌توانید از آن استفاده کنید تا با توابع و ابزارهای مختلف آشنا شوید. همچنین، بسیاری از منابع آموزشی و ویدئوها در دسترس هستند که می‌توانند به شما در یادگیری Pandas کمک کنند.

3. Matplotlib

کتابخانه Matplotlibیک کتابخانه قدرتمند برای تولید نمودارها و تصاویر در پایتون است. این کتابخانه امکاناتی را برای تولید انواع نمودارها مانند خطی، میله‌ای، دایره‌ای، نقطه‌ای، سطحی و... فراهم می‌کند.

با استفاده از Matplotlib، می‌توانید داده‌های خود را به صورت گرافیکی نشان دهید و به راحتی با آن‌ها تعامل کنید. همچنین، این

کتابخانه امکاناتی برای سفارشی‌سازی نمودارها، افزودن عناصر مختلف مانند عنوان، برچسب‌ها و لگوها، و ایجاد نمودارهای پیچیده‌تر مانند نمودارهای چندبعدی را فراهم می‌کند.

برای استفاده از کتابخانهMatplotlib ، ابتدا باید آن را نصب کنید. معمولاً با استفاده از مدیر بسته pip می‌توانید Matplotlib را نصب کنید. برای نصب Matplotlib، دستور زیر را در خط فرمان وارد کنید:

```
pip install matplotlib
```

بعد از نصب، می‌توانید Matplotlib را در برنامه خود وارد کنید با استفاده از دستور زیر:

```
python
import matplotlib.pyplot as plt
```

حالا می‌توانید از توابع و ابزارهای موجود در کتابخانه Matplotlib استفاده کنید. به عنوان مثال، می‌توانید یک نمودار خطی از داده‌های خود رسم کنید:

```
python
import matplotlib.pyplot as plt

# داده‌ها
x = [1, 2, 3, 4, 5]
y = [2, 4, 6, 8, 10]

# رسم نمودار خطی
plt.plot(x, y)

# نمایش نمودار
plt.show()
```

این کد داده‌های x و y را تعریف می‌کند و سپس یک نمودار خطی از این داده‌ها رسم می‌کند. سپس با استفاده از تابع show()، نمودار را نمایش می‌دهد.

کتابخانه Matplotlibدارای مستندات جامعی است که شما می‌توانید از آن استفاده کنید تا با توابع و ابزارهای مختلف آشنا شوید. همچنین، بسیاری از منابع آموزشی و ویدئوها در دسترس هستند که می‌توانند به شما در یادگیری Matplotlib کمک کنند.

4. Scikit_learn

کتابخانه Scikit-learnیکی از کتابخانه‌های قدرتمند برای یادگیری ماشین در پایتون است. این کتابخانه ابزارها و توابعی را برای انجام وظایف یادگیری ماشین مانند تقسیم داده، استخراج ویژگی، انتخاب مدل، آموزش و ارزیابی مدل‌ها و ...فراهم می‌کند.

Scikit-learnدارای رابط برنامه‌نویسی ساده‌ای است که استفاده از آن بسیار آسان است. برای استفاده از این کتابخانه، ابتدا باید آن را نصب کنید. معمولاً با استفاده از مدیر بسته pip می‌توانید Scikit-learnرا نصب کنید. برای نصبScikit-learn ، دستور زیر را در خط فرمان وارد کنید:

```
pip install scikit-learn
```

بعد از نصب، می‌توانید Scikit-learn را در برنامه خود وارد کنید با استفاده از دستور زیر:

```
python
import sklearn
```

حالا می‌توانید از توابع و ابزارهای موجود در کتابخانه Scikit-learn استفاده کنید. به عنوان مثال، می‌توانید یک مدل رگرسیون خطی را با استفاده از داده‌های خود آموزش دهید:

```
python
from sklearn.linear_model import LinearRegression

# داده‌ها
X = [[1], [2], [3], [4], [5]]
y = [2, 4, 6, 8, 10]

# آموزش مدل رگرسیون خطی
model = LinearRegression()
model.fit(X, y)

# پیش‌بینی مقدار جدید
new_X = [[6]]
predicted_y = model.predict(new_X)
print(predicted_y)
```

این کد داده‌های X و yرا تعریف می‌کند و سپس یک مدل رگرسیون خطی با استفاده از این داده‌ها آموزش می‌دهد. سپس با استفاده از تابع ()predict، مقدار جدیدی را پیش‌بینی می‌کند و نتیجه را چاپ می‌کند.

کتابخانه Scikit-learnدارای مستندات جامعی است که شما می‌توانید از آن استفاده کنید تا با توابع و ابزارهای مختلف آشنا شوید. همچنین، بسیاری از منابع آموزشی و ویدئوها در دسترس هستند که می‌توانند به شما در یادگیری Scikit-learnکمک کنند.

5. کتابخانه ggplot2 در: R

Rچی هست:

Rیک زبان برنامه‌نویسی و محیط نرم‌افزاری است که برای آمار و تحلیل داده استفاده می‌شود. این زبان توسط گروهی از محققان آماری توسعه داده شده است و اکنون یکی از ابزارهای محبوب در علوم داده و آمار است.

ggplot2یکی از کتابخانه‌های قدرتمند R است که برای تولید نمودارهای آماری و تجزیه و تحلیل داده استفاده می‌شود. این کتابخانه بر اساس فلسفه "Grammar of Graphics" ساخته شده است که با استفاده از قواعد ساده‌ای، امکان تولید نمودارهای پیچیده را فراهم می‌کند.

برای استفاده از کتابخانه ggplot2، ابتدا باید آن را نصب کنید. معمولاً می‌توانید از مدیر بسته CRAN در RStudio استفاده کنید. برای نصبggplot2 ، دستور زیر را در کنسول R وارد کنید:

```
R
install.packages("ggplot2")
```

بعد از نصب، می‌توانید ggplot2 را در برنامه خود وارد کنید با استفاده از دستور زیر:

```
R
library(ggplot2)
```

حالا می‌توانید از توابع و ابزارهای موجود در کتابخانه ggplot2 استفاده کنید .به عنوان مثال، می‌توانید یک نمودار پراکندگی را برای دو متغیر ایجاد کنید:

```
R
# داده‌ها
x <- c(1, 2, 3, 4, 5)
```

```
y <- c(2, 4, 6, 8, 10)

ساخت نمودار پراکندگی #
ggplot(data = NULL, mapping = aes(x = x, y = y)) +
 geom_point()
```

این کد داده‌های x و y را تعریف می‌کند و سپس با استفاده از تابع ()ggplot و ()geom_point یک نمودار پراکندگی ایجاد می‌کند.

کتابخانه ggplot2دارای مستندات جامعی است که شما می‌توانید از آن استفاده کنید تا با توابع و ابزارهای مختلف آشنا شوید. همچنین، بسیاری از منابع آموزشی و ویدئوها در دسترس هستند که می‌توانند به شما در یادگیری ggplot2 کمک کنند.

6. کتابخانه dplyr در R :

کتابخانه dplyrیکی از کتابخانه‌های محبوب و قدرتمند در زبان برنامه‌نویسی R است که برای تحلیل و تغییر داده‌ها به صورت سریع و کارآمد استفاده می‌شود. این کتابخانه برای انجام عملیات‌های متداول در تحلیل داده مانند فیلتر کردن، مرتب سازی، تلفیق، خلاصه‌سازی و تغییر شکل داده‌ها استفاده می‌شود.

با استفاده از dplyr می‌توانید داده‌ها را به صورت پرونده‌ای (tidy data) سازماندهی کنید و عملیات مورد نیاز را بر روی آن‌ها انجام دهید. این کتابخانه قابلیت‌های بسیاری برای تحلیل داده‌ها ارائه می‌دهد، از جمله:

1. فیلتر کردن: با استفاده از توابع ()filter و ()subset می‌توانید داده‌های مورد نظر خود را بر اساس شرایط مشخصی انتخاب کنید.

2. مرتب سازی: توابع ()arrange و ()sort به شما امکان می‌دهند داده‌ها را بر اساس یک یا چند ستون مرتب کنید.

3. تلفیق داده‌ها: توابع ()bind_rows و ()bind_cols به شما امکان می‌دهند داده‌ها را به صورت عمودی یا افقی تلفیق کنید.

4. خلاصه‌سازی: با استفاده از توابع ()summarise و ()group_by می‌توانید داده‌ها را بر اساس یک یا چند ستون خلاصه کنید.

5. تغییر شکل داده‌ها: توابع ()mutate و ()transmute به شما امکان می‌دهند داده‌ها را با استفاده از توابع ریاضی، رشته‌های متنی و غیره تغییر دهید.

وجود این قابلیت‌ها و سایر قابلیت‌های dplyr باعث می‌شود که تحلیل داده‌ها در R سریع‌تر و قابل فهم‌تر شود.