بارگذاری (load) داده برای پروژه های Machine Learning

فرض کنید می خواهید یک پروژه یادگیری ماشین را شروع کنید، اولین و مهمترین چیزی که نیاز دارید چیست؟ مهمترین چیز داده است که برای شروع هر پروژه یادگیری ماشین باید بارگیری کنیم. با توجه به داده ها، متداول ترین قالب داده برای پروژه های یادگیری ماشین، CSV (comma-separated values) است.

اساسا، CSV یک قالب ساده فایل است که برای ذخیره داده های جدولی(اعداد و متن) مانند یک صفحه گسترده با متن ساده(spreadsheet in plain text) استفاده می شود. در پایتون، می توان به روش های مختلف داده CSV را بارگیری کرد، اما قبل از بارگیری داده های CSV، باید برخی ملاحظات را در نظر بگیریم.

ملاحظات حین بارگذاری داده CSV :

قالب داده CSV متداول ترین قالب برای داده های یادگیری ماشین است، اما در حین بارگیری داده درون پروژه های ML خود باید ملاحظات اساسی زیر را در نظر بگیریم.

سر برگ فایل(file header) :

در فایل های داده CSV، سربرگ شامل اطلاعات برای هر فیلد است. ما باید از جداکننده مشابه برای فایل سربرگ و فایل داده استفاده کنیم، زیرا این فایل سربرگ است که تعیین کننده چگونگی تفسیر فیلدهای داده است.

در ادامه دو حالت مرتبط با سربرگ فایل CSV مطرح شده است که باید مد نظر داشته باشیم.

حالت ۱ : زمانی که فایل داده دارای یک فایل سربرگ می باشد – اگر فایل داده دارای یک فایل سربرگ باشد، نام ها را به صورت خودکار به هر ستون داده تخصیص می دهد.

حالت ۲ : زمانی که فایل داده دارای فایل سربرگ نمی باشد. - – اگر فایل داده دارای یک فایل سربرگ نباشد، باید نام ها را به صورت دستی به هر ستون داده تخصیص دهیم.

در هر دو حالت، باید صریحا مشخص کنیم که آیا فایل CSV ما سربرگ دارد یا خیر.

نظرات (Comments) :

نظرات در هر فایل داده ای، اهمیت ویژه خود را دارند. در فایل داده CSV ، نظرات با قرار دادن نماد هش (hash #) در ابتدای یک خط مشخص می شود. هنگام بارگیری داده CSV درون پروژه های یادگیری ماشین، باید نظرات را در نظر بگیریم، زیرا در صورت داشتن نظرات درون فایل، ممکن است لازم باشد بیان کنیم، که بر اساس متدی که برای بارگیری انتخاب می کنیم، آن نظرات را می پذیریم یا خیر.

جدا کننده (delimiter):

در فایل های داده CSV، کاراکتر ویرگول (،) جداکننده استاندارد است. نقش جدا کننده، تفکیک مقادیر در فیلد ها است. در نظر گرفتن نقش جدا کننده، حین بارگذاری فایل CSV درون پروژه های یادگیری ماشین مهم است، زیرا میتوانیم از جدا کننده های متفاوتی مانند tab و فاصله (white space) نیز استفاده کنیم. اما در صورت استفاده از جدا کننده ای غیر از جدا کننده استاندارد، باید صریحا آن را مشخص کنیم.

نقل قول ها(quotes) :

در فایل های داده CSV، علامت نقل قول دوتایی (" ") ، کاراکتر پیش فرض نقل قول است. در نظر گرفتن نقش نقل قول ها، حین بارگذاری فایل CSV درون پروژه های یادگیری ماشین مهم است، زیرا علاوه بر علامت نقل قول دوتایی، می توانیم از سایر کاراکتر های نقل قول نیز استفاده کنیم. اما در صورت استفاده از یک کاراکتر نقل قول متفاوت، غیر از نمونه استاندارد، باید صریحا آن را مشخص کنیم.

متدهای بارگذاری فایل داده CSV:

در هنگام کار با پروژه های یادگیری ماشین، حیاتی ترین کار، بارگیری صحیح داده به درون آن است. متداول ترین قالب داده برای پروژه های یادگیری ماشین CSV است، و برای تجزیه و تحلیل، دارای حالات متنوع و مشکلات متفاوتی است. در این بخش، به بررسی سه مورد از راهکارهای رایج برای بارگذاری فایل داده CSV در پایتون می پردازیم.

بارگیری CSV با کتابخانه استاندارد پایتون:

اولین و پر کاربرد ترین روش برای بارگیری فایل داده CSV ، استفاده از کتابخانه استاندارد پایتون است که ماژول های داخلی متنوعی را برای ما فراهم می کند، به عنوان مثال ماژول csv و تابع *reader()*. در ادامه مثالی از بارگیری فایل داده CSV به کمک آن، ارائه شده است.

مثال: در این مثال، از مجموعه داده iris *flower* استفاده می کنیم که می تواند در دایرکتوری محلی ما دانلود شود. پس از بارگیری فایل داده، میتوانیم آن را به آرایه NumPy تبدیل و از آن برای پروژه های یادگیری ماشین استفاده کنیم. در ادامه اسکریپت پایتون برای بارگیری فایل داده CSV قرار دارد. ابتدا، باید به صورت زیر، ماژول csv فراهم شده توسط کتابخانه استاندارد پایتون را وارد(import) کنیم.

```
import csv
```

سپس، برای تبدیل داده بارگیری شده به آرایه NumPy، باید ماژول Numpy را وارد کنیم.

```
import numpy as np
```

حال، مسیر کامل فایل، که روی دایرکتوری محلی ذخیره شده و در بر دارنده فایل داده CSV است را ارائه کنید.

```
path = r"c:\iris.csv"
```

سپس، از تابع csv.reader() برای خواندن داده از فایل CSV ، استفاده کنید.

```
with open(path,'r') as f:
reader = csv.reader(f,delimiter = ',')
headers = next(reader)
data = list(reader)
data = np.array(data).astype(float)
```

به واسطه این خط از اسکریپت، می توان نام سربرگ ها را چاپ کرد.

```
print(headers)
```

اسکریپت زیر شکل داده را چاپ می کند. مانند تعداد سطر ها و ستون ها در فایل.

```
print(data.shape)
```

این خط از اسکریپت، سه خط اول فایل داده را ارائه می کند.

```
print(data[:3])
```

خروجی:

```
['sepal_length',     'sepal_width',     'petal_length',
'petal_width']
(150, 4)
[[5.1 3.5 1.4 0.2]
[4.9 3. 1.4 0.2]
[4.7 3.2 1.3 0.2]]
```

بارگیری CSV به کمک *NumPy* :

روشی دیگر برای بارگیری فایل داده CSV، استفاده از *NumPy* و تابع *numpy.loadtxt()* است. مثال زیر نمونه ای از بارگیری فایل داده CSV به کمک آن است.

مثال: در این مثال، از مجموعه داده Pima Indians که شامل داده های بیماران دیابتی است استفاده می کنیم. این مجموعه داده، یک مجموعه داده عددی است که سر برگ ندارد. همچنین می توان آن را درون دایرکتوری محلی خود دانلود کرد. پس از بارگیری فایل داده، می توان آن را به آرایه *NumPy* تبدیل کرد و از آن در پروژه های یادگیری ماشین استفاده کرد. کد زیر اسکریپت پایتون برای بارگیری فایل داده CSV است.

```
from numpy import loadtxt
path = r"C:\pima-indians-diabetes.csv"
datapath= open(path, 'r')
data = loadtxt(datapath, delimiter=",")
print(data.shape)
print(data[:3])
```

خروجی:

```
(768, 9)
[[ 6. 148. 72. 35. 0. 33.6 0.627 50. 1.]
[ 1. 85. 66. 29. 0. 26.6 0.351 31. 0.]
[ 8. 183. 64. 0. 0. 23.3 0.672 32. 1.]]
```

بارگیری CSV با *Pandas* :

روشی دیگر برای بارگیری فایل داده CSV استفاده از *Pandas* و تابع *pandas.read_csv()* است. این یک تابع بسیار منعطف است که یک pandas.DataFrame را باز می گرداند و فورا می توان از آن برای رسم (plotting) استفاده کرد. مثال زیر نمونه ای از بارگیری فایل داده CSV به کمک آن است.

مثال: در اینجا، دو اسکریپت پایتون را پیاده سازی می کنیم. اسکریپت اول با مجموعه داده Iris و سربرگ، و دیگری با استفاده از مجموعه داده *Pima Indians* که یک مجموعه داده عددی بدون سربرگ است. هر دو مجموعه داده را می توان درون دایرکتوری محلی دانلود کرد.

اسکریپت ۱ : کد زیر اسکریپت پایتون برای بارگیری فایل داده CSV روی مجموعه داده Iris با استفاده از *Pandas* است.

```
from pandas import read_csv
path = r"C:\iris.csv"
data = read_csv(path)
print(data.shape)
print(data[:3])
```

خروجی:

```
(150, 4)
sepal_length sepal_width petal_length petal_width
0 5.1          3.5          1.4           0.2
```

```
1  4.9          3.0          1.4          0.2
2  4.7          3.2          1.3          0.2
```

اسکریپت ۲ : کد زیر اسکریپت پایتون برای بارگیری فایل داده CSV به همراه ارائه نام سربرگ ها است که روی مجموعه داده دیابتی های Pima Indians با استفاده از Pandas اعمال می شود.

```
from pandas import read_csv
path = r"C:\pima-indians-diabetes.csv"
headernames = ['preg', 'plas', 'pres', 'skin', 'test',
'mass', 'pedi', 'age', 'class']
data = read_csv(path, names=headernames)
print(data.shape)
print(data[:3])
```

خروجی :

```
(768, 9)
   preg   plas     pres   skin   test   mass    pedi   age
class
0      6    148       72     35      0   33.6   0.627    50
1
1      1     85       66     29      0   26.6   0.351    31
0
2      8    183       64      0      0   23.3   0.672    32
1
```

تفاوت بین سه راهکار استفاده شده در بالا برای بارگیری فایل داده CSV ، توسط مثال های ارائه شده به سادگی قابل فهم است.