تمرین دوم درس یادگیری ماشین: دسته‌بندی به کمک رگرسیون لجستیک

در این تمرین قصد داریم یک مدل رگرسیون لجستیک را آموزش دهیم تا عملیات دسته‌بندی دادگان را برای ما انجام دهد. برای این کار قصد داریم از مجموعه داده Fashion MNIST استفاده کنیم. در این تمرین بایستی از LogisticRegression در کتابخانه scikit-learn استفاده کنید. به این صورت که در یک مدل لجستیک پارامترهای زیر را تنظیم نمایید و مدل را ارزیابی کنید. همینطور یک SGDClassifier را بر روی این دادگان ارزیابی کنید و با مدل لجستیک مقایسه نمایید. مراحل آموزش و آزمون را انجام دهید و دقت نهایی هر مرحله را گزارش کنید و نتایج را تحلیل کنید.

```
solver='liblinear', C=0.001, multi_class='auto', random_state=0, penalty='l2'
```

برای این تمرین بایستی گزارشی در قالب PDF بنویسید و در آن به سوالات زیر پاسخ دهید. کدی که می‌نویسید بایستی بر اساس این سوالات نوشته شده باشد. نام فایل‌های ارسالی باید نام‌خانوادگی و شماره دانشجویی شما باشد.

در کد خود:

۱. دو مجموعه آموزش و تست Fashion MNIST را در کد خود لود کنید.
۲. دو مجموعه آموزش و تست ورودی را نرمالسازی کنید و مقادیر را از مقیاس بین 0 تا 255 به مقیاس بین 0 و 1 ببرید.
۳. مدل لجستیک را با پارامترهای بالا ایجاد کنید. مدل را آموزش داده و دقت آموزش و تست را بدست آورید.
۴. برای پارامتر C مقادیر دیگری از قبیل 0.001،0.01، 0.05، 0.1، 0.2، 0.3، را مورد بررسی قرار دهید و بهترین نتیجه را گزارش کنید. برای انجام این عمل از GridSearchCV با مقدار 5 برای k استفاده نمایید.
۵. ماتریس درهم‌ریختگی (Confusion Matrix) را برای حالت تست محاسبه و رسم نمایید.
۶. یک مدل SGDClassifier ایجاد کنید و آن را با همین دادگان آموزش و تست ارزیابی کنید.

در گزارش خود:

۱. مجموعه داده Fashion MNIST را معرفی نمایید.
۲. چرا این دادگان نرمال شده‌اند؟
۳. دقت‌های دو بخش آموزش و تست را گزارش کنید.
۴. تحلیل خود را از دقت‌های بدست آمده بنویسید.
۵. پارامترهای مدل رگرسیون لجستیک که در بالا مشخص شده‌اند و مقادیر ممکن برای هر پارامتر را معرفی کنید.
۶. برای پارامتر C مقادیر دیگری را در بخش ۴ پیاده‌سازی مورد بررسی قرار دادیم. کدام مقدار دقت بیشتری در حالت ارزیابی دارد؟ تحلیل کنید که چرا این مقدار از بقیه مقادیر بهتر عمل کرده است.
۷. ماتریس درهم‌ریختگی (Confusion Matrix) پیش‌بینی مدل در بخش ۵ پیاده‌سازی را تحلیل نمایید.
۸. دقت آموزش و تست مدل SGDClassifier در بخش ۶ پیاده‌سازی را گزارش نمایید.
۹. کدام مدل دسته‌بندی بهتری انجام داده است؟ علت آن چیست؟

راهنمایی:

برای آشنایی با نحوه استفاده از LogisticRegression که توسط scikit-learn پیاده‌سازی شده است توصیه می‌شود از داکیومنتیشن اصلی آن استفاده کنید. (لینک)

به عنوان مثال در قطعه کد زیر یک مدل لجستیک ایجاد شده است.

```
from sklearn.linear_model import LogisticRegression
... = LogisticRegression(C=..., solver=..., random_state=...)
```

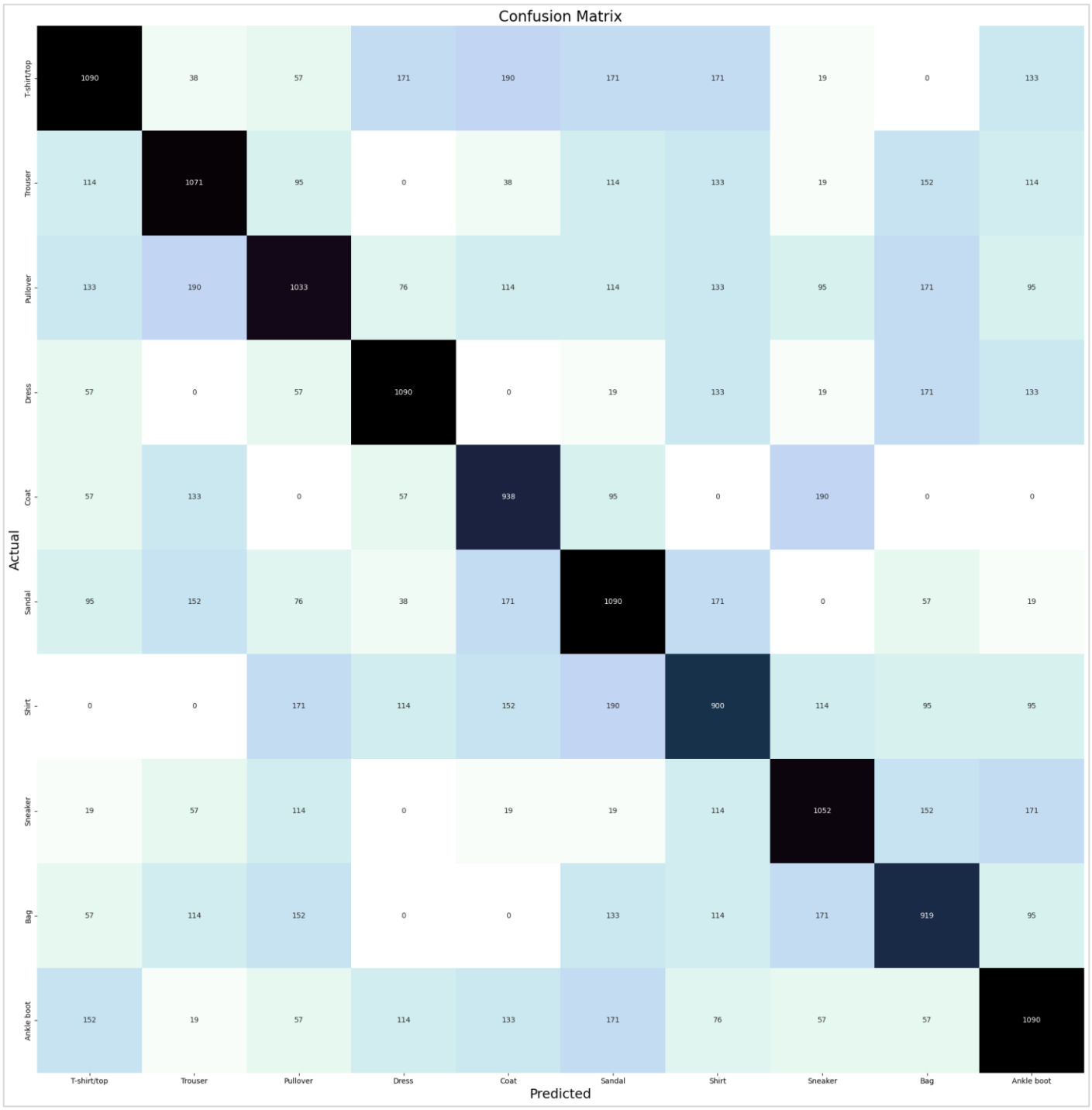
برای یادگیری با نحوه کار k-fold cross validation و GridSearchCV به تمرین قبل مراجعه کنید.

Confusion Matrix

برای بدست آوردن ماتریس درهم‌ریختگی از تابع کتابخانه scikit learn استفاده کنید.

```
from sklearn.metrics import confusion_matrix
```

برای رسم این ماتریس مانند تمرین قبل از کتابخانه seaborn استفاده کنید. دقت داشته باشید که سطر و ستون‌های این ماتریس به درستی تعیین گردند. (نمونه خروجی در صفحه بعد برای شما قرار گرفته شده است)


Confusion Matrix
Actual
Predicted
T-shirt/top
Trouser
Pullover
Dress
Coat
Sandal
Shirt
Sneaker
Bag
Ankle boot
1090 38 57 171 190 171 171 19 0 133
114 1071 95 0 38 114 133 19 152 114
133 190 1033 76 114 114 133 95 171 95
57 0 57 1090 0 19 133 19 171 133
57 133 0 57 938 95 0 190 0 0
95 152 76 38 171 1090 171 0 57 19
0 0 171 114 152 190 900 114 95 95
19 57 114 0 19 19 114 1052 152 171
57 114 152 0 0 133 114 171 919 95
152 19 57 114 133 171 76 57 57 1090

موفق باشید