گزارش مینی پروژه درس یادگیری ماشین

اعضای گروه

محمد سفید

شماره دانشجویی: 40206864

عرفان مجیدی

شماره دانشجویی: 40211034

اسفند 1403

کدهای مربوط به تمرین اول در لینک های زیر قابل مشاهدا است:
Google colab: Untitled0.ipynb - Colab
Github: mohammad-sefid/Machine_Learning

پردازش داده

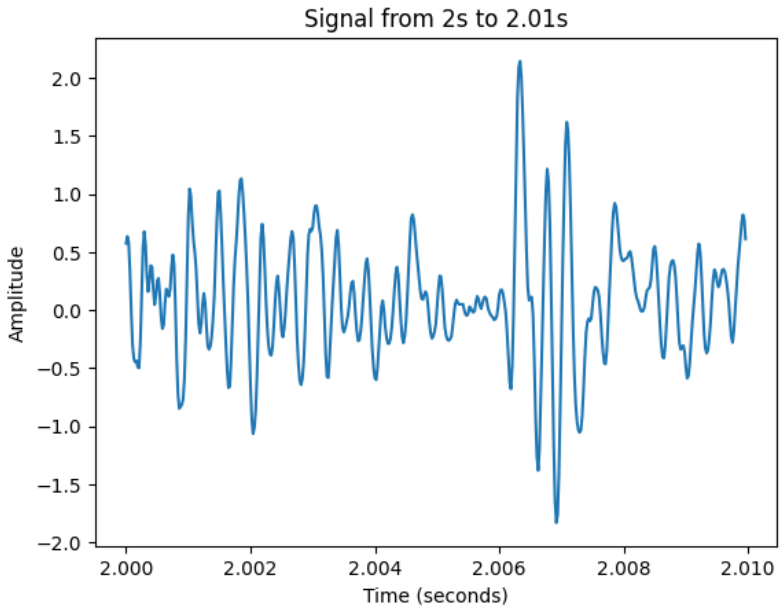
CWRU Dataset I

آ. 1. فرمت فایل دانلود شده (.mat) است که با استفاده از دستور loadmat قابل فراخوانی می‌باشد.

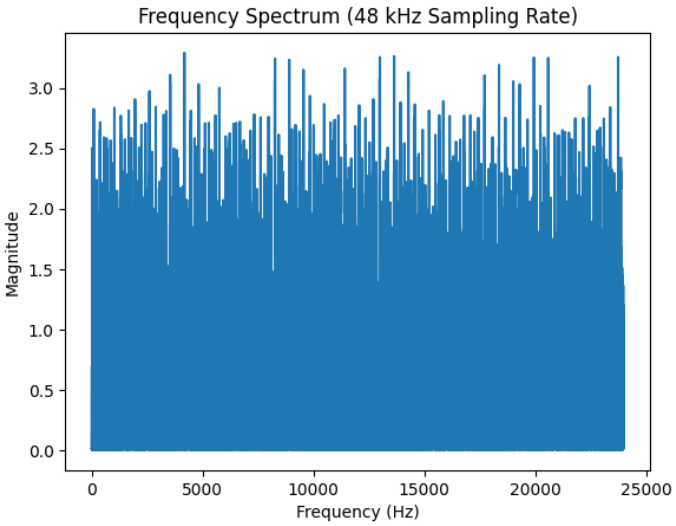
2. با فراخوانی داده و ذخیره‌سازی آن در یک متغیر به داده با کلاس دیکشنری(Dict) می‌رسیم. اجزای دیکشنری شامل '__header__', '__version__', '__globals__', 'X109_DE_time', 'X109_FE_time', 'X109RPM' است.

3. سیگنال 'X109_DE_time' را انتخاب کرده و در متغیری به نام selected_signal ذخیره کردیم.

ب. در این بخش با استفاده از کتابخانه matplotlib سیگنال را برای کل بازه زمانی [2, 2.01] ثانیه با فرکانس نمونه‌برداری 48KHz نمایش می‌دهیم.



ج. 1– سپس با استفاده از تعریف تبدیل فوریه طیف فرکانسی سیگنال را نمایش می‌دهیم.

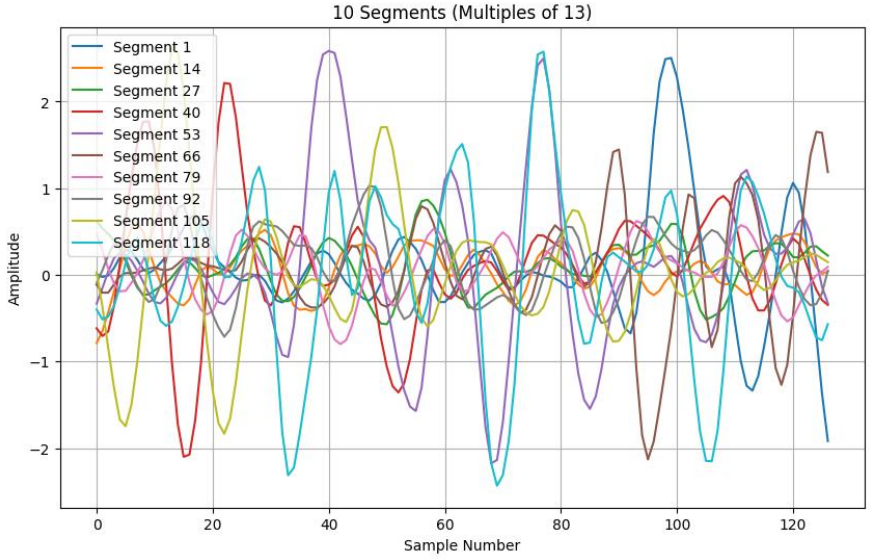


2. فرکانس غالب سیگنال برابر 4174.700128721232 Hz است.

د. سیگنال را به قطعاتی با اندازه 128 با فرض هم‌پوشانی 64 تقسیم کرده و در یک آرایه numpy ذخیره کردیم.

ه. 1. نمونه‌های به‌دست‌آمده در بخش قبل را در یک دیتافریم از کتابخانه pandas ذخیره‌سازی کردیم. (به این دلیل که نمونه ها شامل 3 بعد هستند بعد سوم را که یک تنها مقدار ثابت دارد حذف کردیم و نمونه‌ها را دوبعدی در نظر می‌گیریم.)

2. با استفاده از دیتافریم ایجاد شده تعداد 10 قطعه مضرب با مضرب 13 را به صورت زیر نمایش می‌دهیم:



و. 1. با استفاده از تعاریف توابع میانگین، انحراف معیار و ریشه میانگین مربعات را محاسبه کردیم.

2. تابع ایجاد شده را برای محاسبه ویژگی‌های تمام نمونه‌های ایجاد شده به‌کارمی‌بریم و در یک دیتافریم جدید(features) ذخیره‌سازی کردیم.

3. سپس این دیتافریم را در یک فایل csv ذخیره‌سازی کردیم و به شکل زیر نمایش می‌دهیم:

Mean	Standard Deviation	RMS
0.09328704166666664	0.6795687853352902	0.6859418387480093
0.06799272916666667	0.26504870273126	0.2736308207038372
0.10262487500000002	0.3033532594011311	0.32024219734141157
0.11966055208333322	0.7269789852687464	0.7367612182703862
0.09058415625000002	0.9707486043866357	0.97496581595568
0.07526997916666665	0.6349854419164747	0.6394310605605668
0.07698169791666666	0.34006375241790726	0.3486682341749541
0.1185063645833333	0.3917326904705702	0.40926551189914545
0.0447003125	0.7775365511611815	0.7788203941405378
0.006354552083333333	0.9890708903467889	0.9890913034010406

Iris Dataset II

آ. 1. دیتاست iris شامل 150 نمونه از سه گونه مختلف setosa, virginica, versicolor است که هر نمونه شامل چهار ویژگی طول کاسبرگ(sepal length)، عرض کاسبرگ(sepal length)، طول گلبرگ(petal length) و عرض گلبرگ(petal width) برحسب سانتی‌متر می‌باشد.

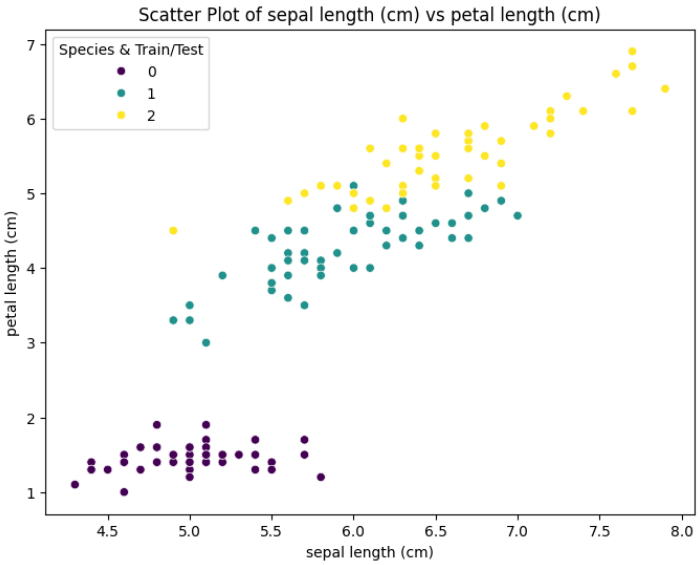
2. در این قسمت از کتابخانه scikit.learn دیتاست iris را فراخوانی کردیم. داده‌ها را به شکل اسم ویژگی‌ها در ستون‌ها و داده‌ها در سطرها نمایش می‌دهیم.

3. در این بخش داده‌ها را در یک دیتا فریم ذخیره کردیم. داده‌ها را به دو بخش مجزای آموزش و تست تقسیم کردیم. 70٪ داده‌ها را برای آموزش و 30٪ آن‌ها را برای تست در نظر گرفتیم.

4. داده‌ها را به تفکیک آموزش و تست در هم ادغام کردیم و در یک جدول نشان می‌دهیم:

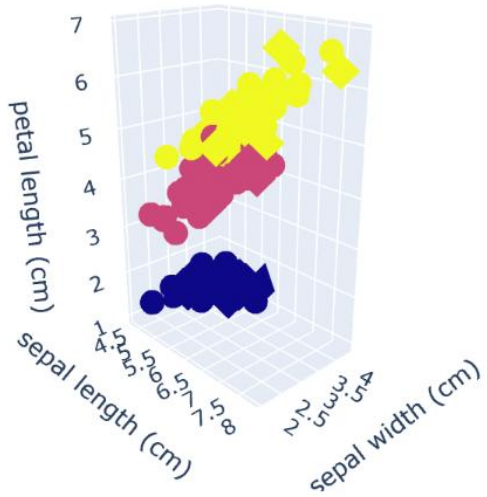
	species	train_test
81	1	train
133	2	train
137	2	train
75	1	train
109	2	train
..	...	...
142	2	test
85	1	test
86	1	test
16	0	test
10	0	test

ب. 1. طبق خواسته سوال دو ویژگی از نمونه‌ها(sepal length, petal length) را انتخاب کردیم و در یک نمودار دوبعدی به شکل زیر نمایش می‌دهیم:



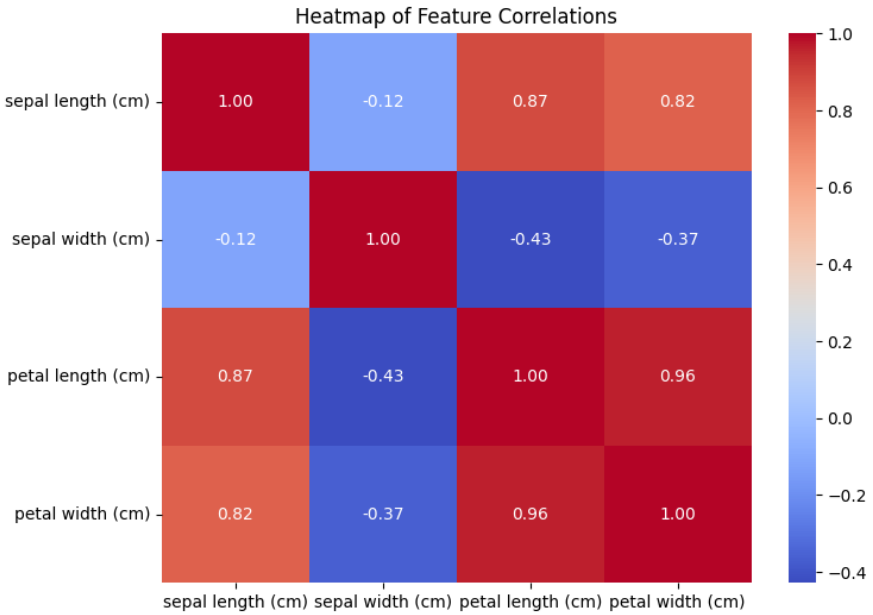
نمودار پراکندگی نمونه‌ها طبق دو ویژگی ذکر شده را نمایش می‌دهد و هر سه نوع گل به رنگ‌های متمایز نمایش داده شده‌اند.

2. طبق خواسته سوال سه ویژگی از نمونه‌ها(sepal length, sepal width, petal length) را انتخاب کردیم و در یک نمودار سه‌بعدی نمایش می‌دهیم:



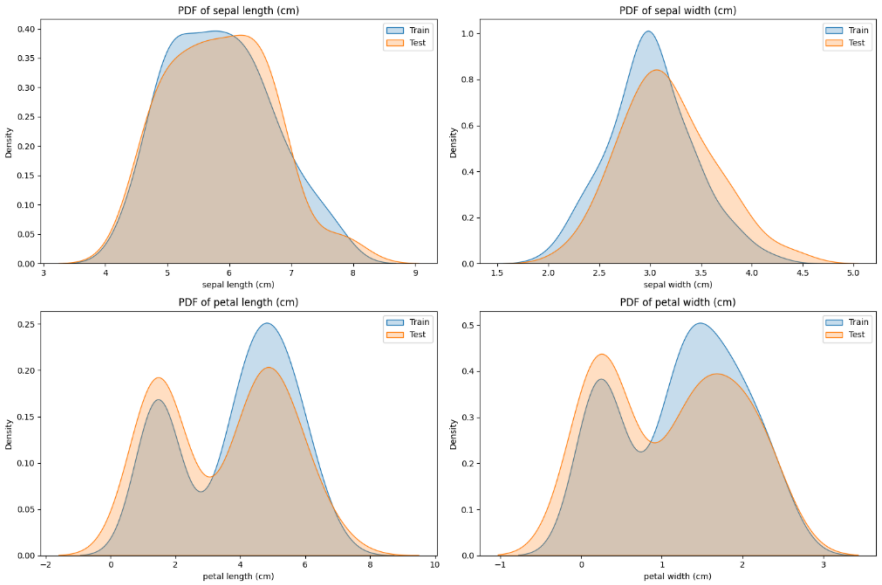
نمودار پراکندگی نمونه‌ها طبق سه ویژگی ذکر شده را نمایش می‌دهد و هر سه نوع گل به رنگ‌های متمایز نمایش داده شده‌اند.

3. در این قسمت نقشه حرارتی ویژگی‌ها که نشان‌دهنده ارتباط ویژگی‌ها به یکدیگر می‌باشد را نمایش می‌دهیم:



به عنوان مثال: همانطور که مشاهده می‌کنید petal length, petal width کرلیشن زیادی نشان می‌دهند که نشان‌دهنده ارتباط آن‌ها به یکدیگر است.

4. در این بخش طبق خواسته صورت سوال تابع ویژگی‌های دادگان را به تفکیک آموزش و تست را برای هر چهار ویژگی دیتاست iris نمایش می‌دهیم:



همانطور که مشاهده می‌کنید نتایج قابل قبولی از آموزش داده‌ها برای هر یک از ویژگی‌ها به دست‌آمده است.

ج. داده‌ها را برای ویژگی sepal length براساس سه بازه [4, 5], [5, 6], [6, 8] به‌ترتیب برچسب 'بلند' ,'متوسط' ,'کوتاه' می‌زنیم و بدین‌ترتیب داده‌های پیوسته را به گسسته تبدیل می‌کنیم و به عنوان یک ویژگی در دیتافریم ذخیره می‌کنیم. خروجی دیتافریم را به صورت زیر نمایش می‌دهیم:

	sepal length (cm)	sepal_length_category
0	5.1	متوسط
1	4.9	کوتاه
2	4.7	کوتاه
3	4.6	کوتاه
4	5.0	کوتاه
..	...	...
145	6.7	بلند
146	6.3	بلند
147	6.5	بلند
148	6.2	بلند
149	5.9	متوسط

د. ویژگی‌های آماری مختلف را برای گونه setosa به‌دست آوردیم. ویژگی های آماری به وسیله متد ()describe شامل مجموع، میانگین، انحراف از معیار، چارک اول، میانه، چارک سوم و حداکثر مقدار برای این گونه برای هر چهار ویژگی دیتاست iris است.

Setosa features:

	sepal length (cm)	sepal width (cm)	petal length (cm)	\
count	50.00000	50.000000	50.000000	
mean	5.00600	3.428000	1.462000	
std	0.35249	0.379064	0.173664	
min	4.30000	2.300000	1.000000	
25%	4.80000	3.200000	1.400000	
50%	5.00000	3.400000	1.500000	
75%	5.20000	3.675000	1.575000	
max	5.80000	4.400000	1.900000	

	petal width (cm)
count	50.000000
mean	0.246000
std	0.105386
min	0.100000
25%	0.200000
50%	0.200000
75%	0.300000
max	0.600000