دانشگاه تربیت مدرس

دانشکده مهندسی برق و کامپیوتر

دانشکده علوم و فناوری های بین رشته ای

تمرین سوم درس یادگیری ماشین

یادگیری با نظارت (درخت تصمیم و رگرسیون لجستیک)

استاد محترم درس:

جناب آقای دکتر قادری

دستیاران آموزشی:

حانیه سرتیپی

نیلوفر مقدس

نیمسال اول ۱۴۰۴-۱۴۰۳

نکات مربوط به تحویل پروژه

- ✓ پروژه فقط با زبان برنامه نویسی پایتون قابل قبول می‌باشد.
- ✓ فایل تحویلی شما، یک فایل زیپ شده‌ی نهایی شامل گزارش کار (فایل pdf) و فایل کد پایتون با پسوند ipynb (Jupiter Notebook) می‌باشد. لطفا آن را به صورت زیر نام‌گذاری و ارسال نمایید.
- ✓

HW_۳_[LastName]_[FirstName]

- ✓ گزارش کار خود را در یک فایل pdf تحویل دهید و از گذاشتن صرفا اسکرین شات‌های پشت سرهم از کد در گزارش کار خودداری کنید.
- ✓ توجه داشته باشید که در فایل ارسالی پایتون، خروجی هر سلول (شامل نمودار، خروجی عددی و غیره) حتما ذخیره شده و قابل مشاهده باشد.
- ✓ لازم است حتما نتایج بدست آمده را گزارش و تحلیل کنید.
- ✓ علاوه بر مهارت حل سوالات، نوشتن پاسخ مینی پروژه‌ها در فرمت گزارش فنی (فصل بندی و صفحه بندی مناسب، رعایت اصول نگارش و ...) برای دانشجویان تحصیلات تکمیلی اهمیت دارد، این مورد نیز در ارزشیابی لحاظ می‌شود.
- ✓ در صورت فراموشی در ارسال کد پایتون، هیچ نمره‌ای به شما تعلق نخواهد گرفت.
- ✓ در صورت مشاهده تشابه در هر بخش از انجام پروژه، نمره هر دو نفر صفر لحاظ می‌گردد.

برای پاسخ به سوالات دانشجویان در مورد مینی‌پروژه ها، دو راه ارتباطی وجود دارد:

۱. ایمیل: برای پرسش سوال از طریق ایمیل، در قسمت To، ایمیل دستیار آموزشی این مینی پروژه، خانم سرتیپی (haniehsartipi[at]gmail.com) را قرار دهید و در قسمت Cc، ایمیل niloofar.moghaddas[at]gmail.com را قرار دهید.

۲. گروه تلگرام: می توانید برای پرسش سوال از مینی پروژه سوم در گروه، خانم سرتیپی را با شناسه تلگرامی @hani_srtp، در پیام خود نام ببرید.

- ✓ ایمیل دستیار آموزشی: haniehsartipi@gmail.com

مقدمه

هدف از انجام این پروژه اجرای الگوریتم های درخت تصمیم و رگرسیون لجستیک برای طبقه بندی داده‌ها و مقایسه عملکرد این دو مدل است.

- **سوگیری قیاسی (inductive bias)**

سوگیری قیاسی به معنای استفاده از دانش پیشینی یا فرضیات درباره‌ی داده‌ها برای انتخاب الگوریتمی است که با ویژگی‌های مجموعه داده سازگارتر باشد. به عبارت دیگر، به جای آزمایش بی‌هدف تمام مدل‌ها، سوگیری قیاسی به ما کمک می‌کند تا با توجه به ساختار و ماهیت داده‌ها، الگوریتم‌های مناسب‌تر را انتخاب کنیم.

برای مثال، اگر فرض کنیم که برخی از ویژگی‌ها اهمیت بیشتری دارند، رگرسیون لجستیک می‌تواند گزینه مناسبی باشد، زیرا این الگوریتم بر اساس وزن‌دهی ویژگی‌ها عمل می‌کند. اما اگر فرض برابری اهمیت ویژگی‌ها را در نظر بگیریم، KNN گزینه‌ی مناسبی است.

- **هایپرپارامترهای درخت تصمیم**

criterion: این هایپرپارامتر مشخص می‌کند که برای تقسیم داده‌ها در هر گره، از کدام معیار استفاده می‌شود.

max_depth: عمق حداکثری درخت تصمیم را محدود می‌کند.

min_samples_leaf: حداقل تعداد نمونه‌ها برای گره‌های برگ.

min_samples_split: حداقل تعداد نمونه‌ها برای تقسیم گره.

- **معیار ارزیابی**

Precision: بیانگر این است که چه تعداد از نمونه هایی که مدل به عنوان آن کلاس تشخیص داده است، واقعا از آن کلاس خاص بوده اند.

Recall: برای هر کلاس بیانگر این است که چقدر از کل نمونه های آن کلاس، توسط مدل به درستی تشخیص داده شده اند.

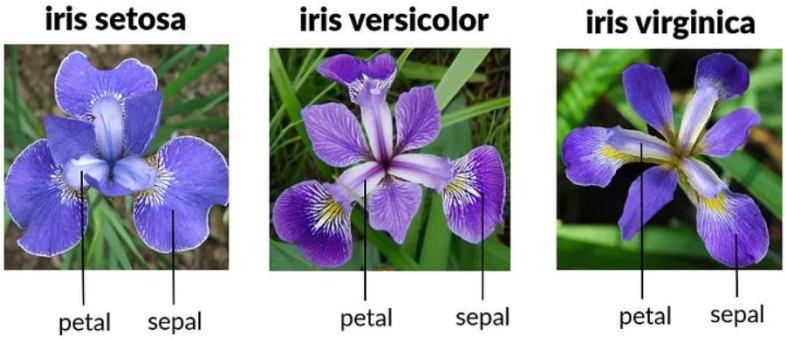
F_1-score: یک معیار جامع تر برای ارزیابی مدل است و تعادلی بین precision و recall برقرار می کند و زمانی که تعادل بین این دو مهم است، مورد استفاده قرار می گیرد.

Accuracy: **صحت**، نسبت تعداد نمونه هایی که به درستی طبقه بندی شده اند به تعداد کل نمونه ها را نشان می‌دهد. این معیار در مجموعه داده‌های متوازن (balance) استفاده می‌شود. مجموعه داده متوازن، مجموعه ای است که در آن تعداد نمونه‌های کلاس‌ها با هم برابر (یا تقریبا برابر) باشد.

پروژه سوم: درخت تصمیم و رگرسیون لجستیک

- **مجموعه داده**

مجموعه داده Iris را در نظر بگیرید. این مجموعه داده شامل ۱۵۰ نمونه و ۴ ویژگی است. متغیر هدف شامل ۳ نوع گل Iris است که هدف ما در این تمرین طبقه بندی درست این ۳ نوع گل از یکدیگر است، با استفاده از الگوریتم های درخت تصمیم و رگرسیون لجستیک.



برای دسترسی به این مجموعه داده می‌توانید کد زیر را اجرا کنید:

```
from sklearn.datasets import load_iris
import pandas as pd
iris=load_iris()
```

```
df=pd.DataFrame(data=iris.data,columns=iris.feature_names)
```

```
df['target']=iris.target
```

سوال ۱) آیا استاندارد سازی داده‌ها برای هر دو الگوریتم درخت تصمیم و رگرسیون لجستیک نیاز است؟ در صورت لزوم داده‌ها را استاندارد کنید.

- **درخت تصمیم**

سوال ۲)

الف) با توجه به توضیحات داده شده در مورد سوگیری قیاسی، کدام الگوریتم را برای آموزش مدل خود انتخاب می‌کنید؟ چرا؟ (درخت تصمیم؟ رگرسیون لجستیک؟ هر دو امکان پذیر است؟)

ب) داده‌های خود را به ۳ بخش train,validation,test تقسیم کنید. (۷۰٪ train، ۱۵٪ test، ۱۵٪validation) و از Girdsearchcv استفاده کنید تا بهترین هایپرپارامترها را پیدا کنید. برای آموزش درخت تصمیم خود از الگوریتم Decisiontreeclassifier استفاده کنید. (هنگام تقسیم‌بندی داده‌های خود shuffle=True قرار دهید)

پ) گزارش کنید کم و زیاد کردن عمق درخت، چه تاثیری بر overfitting یا underfitting شدن مدل دارد؟(صحت مدل را گزارش کنید)

ت) بهترین مدلی که در قسمت" ب " بدست آورده‌اید را با مجموعه داده‌های تست ارزیابی کنید. صحت (accuracy) و ماتریس درهم ریختگی (confusion matrix) مدل را گزارش کنید.

همان طور که مشاهده می‌کنید داده‌های ما balance هستند و تعداد نمونه‌های هر کلاس با سایر کلاس‌ها برابر است.

```
# check if our data is balance or not
label_counts = df['target'].value_counts()
label_counts
```

```
target
0    50
1    50
2    50
Name: count, dtype: int64
```

فرض کنید داده های ما balance نبودند(تعداد نمونه ها در کلاس های مختلف برابر نبود) در این شرایط، چه معیارهایی به جز صحت برای ارزیابی مدل لازم هستند؟ در مورد این معیارها جستجو کنید و سه معیار را توضیح دهید.

- **رگرسیون لجستیک**

سوال ۳)

الف) این بار از رگرسیون لجستیک برای طبقه بندی کلاس‌ها استفاده کنید. داده‌های خود را به سه بخش train,validation,test تقسیم کنید. (train ٪۷۰، test ۱۵٪، validation۱۵٪) و از Girdsearchcv استفاده کنید تا بهترین هایپرپارامترها (c, penalty) را پیدا کنید.

ب) مدل خود را با داده‌های مجموعه تست ارزیابی کنید و صحت، ماتریس درهم‌ریختگی را گزارش کنید.

- **مقایسه و تحلیل نتایج**

سوال ۴)

الف) نتایج بدست آمده از صحت هر دو مدل را گزارش کنید و با یکدیگر مقایسه کنید. پاسخ خود به سوال سوگیری قیاسی را در این بخش ارزیابی کنید.

ب) با توجه به تقسیم بندی ما در این مساله (train ٪۷۰، test ۱۵٪، validation۱۵٪) ،اگر از model selection برای انتخاب بهترین هایپرپارامترها استفاده کنیم نسبت به حالتی که هایپرپارامترها را دستی تعیین می‌کنیم، آیا همیشه به بیشترین میزان صحت مدل بر روی مجموعه داده‌ی تست می‌رسیم؟ توضیح دهید. (راهنمایی: به متوازن بودن تقسیم بندی دقت کنید، test ۱۵٪ و validation٪۱۵)