بسم الله الرحمن الرحیم

موضوع :مینی پروژه – پاسخ سوال اول

درس مربوط : یادگیری ماشین لرنینگ

اساتید راهنما: آقایان دکتر علیاری و احمدی

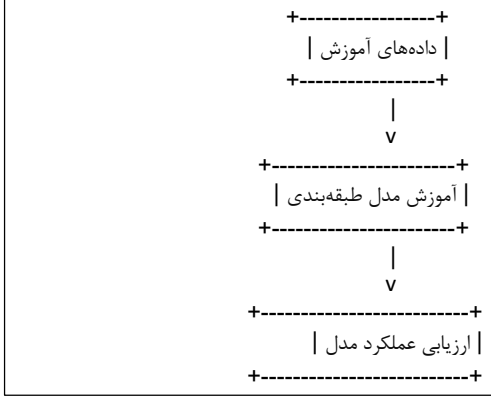
دانشجو: ابوذر بختیاری برزیده

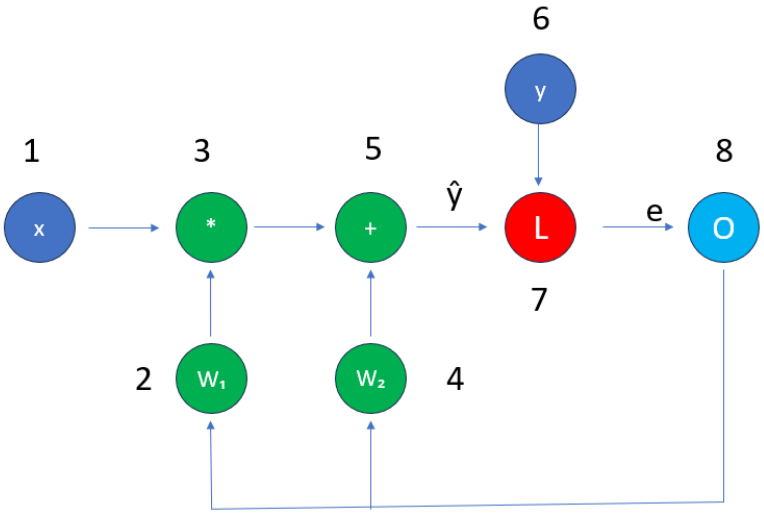
شماره دانشجویی: 4021634202

فروردین 1403

پاسخ سوال اول

-1 یک دیاگرام بلوکی معمولاً برای نمایش فرآیند آموزش و ارزیابی یک مدل طبقه‌بندی خطی مفید است. این دیاگرام بلوکی می‌تواند به شکل زیر باشد:





قسمت 1: در این بلوک دیاگرام ما چندین قسمت مختلف داریم که هر کدام را به اختصار و در انتها توضیحات نهایی را می دهیم.

1- در این قسمت مقدار ورودی ما می باشد که به چه صورت وارد می شود که ممکن است یک کلاس و یا چند کلاس باشد که ما می توانیم از مقدار ورودی آن ها را شناسایی کنیم. این مرحله شامل داده‌هایی است که برای آموزش مدل استفاده می‌شود. این داده‌ها به دو بخش تقسیم می‌شوند: داده‌های ورودی (ویژگی‌ها) و برچسب‌ها (کلاس‌ها).

2- این قسمت معادله ای وارد بلوک دیاگرام ما می شود ، این معادله می تواند معادله خط ، معادله خطوط و یا معادله صفحه ای باشد.

3- ضرب مقدار ورودی و معادله موجود ما می باشد.

4- در این قسمت ما معادله از قبل را داریم که ممکن است تصادفی به دست امده باشد و یا خیر در اطلاعات مسئله باشد.

5- در این قسمت ما جمع معادله تصادفی اولیه با حاصل ضرب مقداری ورودی و معادله موجود ما است در این زمان در حاصل جمع این دو جمع شونده ، ما انتظار داریم که متغییر y بدست بیاید ولی در این قسمت $\hat{y}$ بدست آمده است.

6- مقدار انتظار ما برای به وقوع پیوستن.

7- در این زمان باید اخلاف مقدار بدست امده و مقدار مورد انتظار ما بدست اید که این کار بر عهده تابع اتلاف می باشد که توابع مختلفی برای این کار موجود می باشند.

8- در ادامه با استفاده از اختلاف بدست امده ما می توانیم از روش های بهینه سازی برای مثال تابع گرادیان نزولی استفاده کنیم و معادلات در قسمت های دو و چهار را اپدیت کنیم تا در مرحله بعدی جواب بهتر ، نزدیک تر و حتی خود مقدار انتظار ما را برآورده کند و این روند را تا زمانی ادامه دهد که بهترین پاسخ ممکن را برای ما بدست اورد.

****

قسمت 2:

در کد زیر ما با استفاده از کتابخانه های پایتون به یک کتبخانه به نام (sklearn) استفاده می کنیم که بخشی از این کتابخانه دارای پایگاه داده های مختلفی می باشد از جمله breast cancer و make classification این پایگاه داده برای ساختن داده های طبقه بندی شده استفاده می شود.

- توضیح کوتاهی در مورد make classification (ساختن داده های طبقه بندی شده) : این تابع به ما اجازه می دهد تا داده های مصنوعی برای مسائل ظبقه بندی شده تولید کنیم و این تابع در اکثر موارد در مسائل یادگیری ماشین کاربرد دارد که دارای بخش ها و آرگومان های مهمی می باشد که به اختصار توضیح میدهم :
  - n_samples : تعداد نمونه هایی که باید تولید شوند
  - n_features : تعداد ویژگی ها یا ویژگی های یک نمونه
  - n_classes : تعداد کلاس های مسئله
  - n_slusters_per_class : تعداد خوش های هر کلاس
  - weights : وزن های هر کلاس
  - class_sep : جدایی بین مراکز داده های مختلف
  - and ect...
- ما با استفاده از ماژول inspect می باشم امکاناتی برای انجام عملیات های با روش ها ، کلاس ها ، توابع و سایر اشیاء در زمان اجرا فراهم کند دستور from inspect import classify_class_attrs از کتابخانه دستور

classify_class_attrs را فراخوانی می کند و این دستور ویژگی های کلاس از نوع طبقه بندی شده را ساماندهی می کند.

- ما با استفاده از ماژول matplotlib که مخفف ان plt است ، استفاده می کنیم این کتابخانه به ما این امکان را می دهد تا تصاویر و نمودار تحت پایتون رسم کنیم.

ما داده های مصنوعی طبقه بندی ایجاد می کنیم که دارای 1000 نمومه است ، دارای 2 ویژگی می باشد برای مثال قد و وزن ، دما و رطوبت ، میزان مقدار مصرف سیگار و افزایش سرطان ریه ، در مرحله بعد میزان ویژگی های اضافی است که می تواند وابسته به ویژگی های اصلی باشد که در این مثال مقدار ان صفر است ، در این ارگومان برای تولید داده های تصادفی از عدد آنتالمپوس استفاده می شود و با تغییر مقدار این ارگومان داده های تصادی تولید شده تغییر میکند ، در قسمت بعد تعداد خوشه های که برای هر کلاس ایجاد می شود مشخص می کند که در اینجا فقط یک خوشه داریم، در ارگومان بعدی جدایی بین مراکز داده های مختلف مقادیر کوچکتر یعنی که هر چه مقادیر کوچک تر نزدیک تر و هرچه مقادیر بزرگ تر دور تر می باشد در مرحله اخر هم تعداد کلاس ها تعیین شده است.

دستور scatter برای رسم نمودار پراکندگی استفاده می شود ، متغییر اول نشان دهنده این است که تمام مقادیر روی ردیف اول ماتریس و متغییر دوم تمام مقادیر روی ردیف دوم را نمایش بدهد ، متغییر اخر هم نمایش دهنده رنگ داده های ما می باشد.

از نظر چالشی بودن چالش خیلی زیاد چالش خاصی برای جدا کردن نداریم چون که ما میزان پراکندگی را 1.9 در مثال قبل قرار دادیم برای سخت تر کردن و چالش برانگیز کردن این مثال می توانیم این پراکندگی را کمتر کنیم که در مثال می بینیم.

توضیح مختصری دیگر: می‌توانید از کتابخانه Scikit-learn برای تولید دیتاست مورد نظر استفاده کنید. این کتابخانه دارای یک ماژول بنام datasetsاست که انواع مختلفی از دیتاست‌های آماده را ارائه می‌دهد. اما چون شما می‌خواهید یک دیتاست ساده با 1000 نمونه، 2 کلاس و 2 ویژگی ایجاد کنید، از تابع make_classificationدر این ماژول می‌توانید استفاده کنید.

در ادامه کدی آمده است که این دیتاست را ایجاد کرده و نمایش می‌دهد:

```
from sklearn.datasets import make_classification
import pandas as pd
```

```
# تولید دیتاست با 1000 نمونه، 2 ویژگی و 2 کلاس
X, y = make_classification(n_samples=1000, n_features=2, n_classes=2, random_state=42)

# تبدیل دیتاست به یک DataFrame
df = pd.DataFrame(X, columns=['Feature 1', 'Feature 2'])
df['Class'] = y

# نمایش اولیه از دیتاست
print("نمونه‌ها از دیتاست:")
print(df.head())
```

این کد یک دیتاست با 1000 نمونه، 2 ویژگی و 2 کلاس ایجاد می‌کند. سپس این دیتاست به صورت یک DataFrame پانداس نمایش داده می‌شود.

****

قسمت 3:

بررسی سرطان شایع پستان(breast_cancer)«علت انتخاب این است که اساتید محترم در ایم مورد مفصل توضیح داده اند»

load_breast_cancer این تابع از ماژول sklearn.datasets برای بارگیری مجموعه داده های سرطان پستان استفاده می کند train_test_split این تابع از ماژول sklearn.model_selection برای تقسیم داده‌ها به دو بخش آموزشی و آزمایشی استفاده می کند LogisticRegression و SGDClassifier: این دو تابع از ماژول sklearn.linear_model برای ایجاد مدل‌های طبقه‌بندی لجستیک و SGD مورد استفاده

قرار می‌گیرند accuracy_score: این تابع از ماژول sklearn.metrics برای ارزیابی دقت مدل استفاده می‌کند.

این قسمت داده هایی که در دیتابیس سرطان پستان دارد را بارگزاری میکند و در دو متغییر قرار می دهد در این تابع یک ارگومان دارد که به معنای این است که داده ها را به صورت دو ارایه و جداگانه بر دو متغییر برگرداند در خط بعدی کد هم اندازه متغییر ها را نشان می دهیم.

این قسمت داده ها و متغییر ها را دو صورت تقسیم بندی می کنیم ، هم از متغییر ابتدایی دو نمونه اموزش و ازمایش استفاده می شود در تابع استفاده شده هم متغییر ابتدایی را ویژگی و متغییر دوم را بردار برچسب ها نام گذاری می شود و اندازه داده های تست را 20 درصد می گذارد بنابراین داده های اموزش باقی می مانند و در خط بعدی هم اندازه ان ها را نشان می دهد.

: Logistic Regression with prediction X_train

در خط اول در یک متغییر تابع رگرسیون لاجستیک را قرار می دهیم تا بتوانیم از ان استفاده کنیم این تابع یک نمونه جدید از مدل رگرسیون لجستیک را ایجاد می‌کند و در متغییر قرار می دهد و در خط دوم تابعی است که فرایند اموزش مدل را کنترل می کند و مدل های ما را اموزش میدهد و در خط سوم تابع پیش بینی کننده ای روی متغییر اولیه ما تعریف می شود که با استفاده داده ی اموزش داده شده ی ایکس پیش بینی را انجام می دهد و در خط چهارم هم با استفاده از تابع میزان دقت می توانیم دقت بین داده پیش بینی شده و داده اموزش داده شده بدست اوریم

: Logistic Regression with prediction X_test

در خط اول در یک متغییر تابع رگرسیون لاجستیک را قرار می دهیم تا بتوانیم از ان استفاده کنیم این تابع یک نمونه جدید از مدل رگرسیون لجستیک را ایجاد می‌کند و در متغییر قرار می دهد در خط دوم تابعی است که فرایند اموزش مدل را کنترل می کند و مدل های ما را اموزش میدهد در خط سوم تابع پیش بینی کننده ای روی متغییر تست و ازمون ما تعریف می شود و در

تابع یک مولفه خواهیم داشت در خط چهارم میزان دقت بین داده پیش بینی شده و تست ما صورت می پذیرد

: Stochastic gradient descent X_train

در خط اول ما در یک متغییر تابع گرادیان نزولی را قرار می دهیم که ویژگی هایی دارد در مرحله اول الفا که نرخ یادگیری می باشد و ان را کنترل می کند که مدل چقدر در هر بهینه سازی پارامتر ها به روز می شود و در اینجا خیلی کم می باشد در مولفه ی بعدی تکرار هایی که الگوریتم برای بهینه سازی انجام می دهد نشان می دهد و در اینجا حداکثر 1000 تکرار می باشد در خط دوم تابعی است که فرایند اموزش مدل را کنترل می کند و مدل های ما را اموزش میدهد در خط سوم تابع پیش بینی کننده ای روی متغییر اموزش ما تعریف می شود و در تابع یک مولفه خواهیم داشت در خط چهارم میزان دقت بین داده پیش بینی شده واموزش ما صورت می پذیرد

: Stochastic gradient descent X_test

در خط اول ما در یک متغییر تابع گرادیان نزولی را قرار می دهیم که ویژگی هایی دارد در مرحله اول الفا که نرخ یادگیری می باشد و ان را کنترل می کند که مدل چقدر در هر بهینه سازی پارامتر ها به روز می شود و در اینجا خیلی کم می باشد در مولفه ی بعدی تکرار هایی که الگوریتم برای بهینه سازی انجام می دهد نشان می دهد و در اینجا حداکثر 1000 تکرار می باشد در خط دوم تابعی است که فرایند اموزش مدل را کنترل می کند و مدل های ما را اموزش میدهد در خط سوم تابع پیش بینی کننده ای روی متغییر تست ما تعریف می شود و در تابع یک مولفه خواهیم داشت در خط چهارم میزان دقت بین داده پیش بینی شده و تست ما صورت می پذیرد

: perceptron prediction with train data

در خط اول ابتدا تابع پرسپترون را در متغییری قرار می دهیم و دارای مولفه می باشد که الفا برای نرخ یادگیری و مولفه بعدی برای کنترل تکرار ها می باشد که در مثال قبل کامل توضیح دادم در خط دوم مدل یادگیری می باشد و امزش مربوط به ان در خط سوم با استفاده از تابع پیش بینی

کننده با استفاده از داده های ایکس اموزش داده شده ی خودمون در خط چهارم با استفاده ازتابعی میران دقت بین تابع پیش بینی شده و قسمت اموزش داده بررسی می کنیم

: perceptron prediction with test data

در خط اول ابتدا تابع پرسپترون را در متغییری قرار می دهیم و دارای مولفه می باشد که الفا برای نرخ یادگیری و مولفه بعدی برای کنترل تکرار ها می باشد که در مثال قبل کامل توضیح دادم در خط دوم مدل یادگیری می باشد و امزش مربوط به ان در خط سوم با استفاده از تابع پیش بینی کننده با استفاده از داده های ایکس تست خودمون در خط چهارم با استفاده ازتابعی میران دقت بین تابع پیش بینی شده و قسمت تست بررسی می کنیم