گزارش پروژه نهایی درس شناسایی الگو

استاد : دکتر آبین

سید سروش مجد

شماره دانشجویی

۴۰۰۴۴۳۱۸۱

بهمن ماه ۱۴۰۰

مقدمه و تشریح ویژگی‌ها

در این پروژه می‌خواهیم ویژگی‌های مناسب برای پیدا کردن رابطه بین وقوع زلزله بیشتر از ۴.۵ ریشتر در ایران را با وضعیت و چینش سیاره‌های منظومه شمسی پیدا کنیم و ببینیم واقعا این چینش سیارات بر روی زلزله تاثیر می‌گذارد یا خیر. برای این کار فایل اکسلی که دارای تاریخ و زمان، موقعیات جغرافیایی، شدت و عمق زلزله و شهر استان محل وقوع زلزله است در اختیار ما قرار داده شده است. برای مرحله pre processing داده‌هایی که مقدار Nan داشتند را حذف کردیم (تعداد این داده‌ها خیلی کم بود) و تعداد کل داده‌ها ۵۴۰۵۸ عدد شد سپس به داده‌ها لیبل زدیم و داده‌هایی با شدت زلزله بیشتر از ۴.۵ ریشتر لیبل ۱ و داده‌هایی با شدت زلزله کمتر از ۴.۵ ریشتر لیبل ۰ خواهند داشت. بدای آنکه موقعیت سیاره‌ها را به دست بیاوریم از کتابخانه solarsystem استفاده کردیم که با ورودی تاریخ و ساعت ویژگی‌های مختلف که موقعیت سیارات منظومی شمسی (۱۲ عدد سیاره) بر اساس طول و عرض جغرافیایی و فاصله AU هستند در اختیار ما قرار می‌دهد و درنتیجه از آن‌ها می‌توانیم ویزگی‌های جدید "فاصله زمین تا آن سیاره" و "زاویه بین خورشید، زمین و آن سیاره" (زاویه از تقسیم ضرب داخلی دو بردار فاصله بر ضرب اندازه‌های دو بردار به دست می‌آید.) را استخراج کردیم و در کل ۵۳ ویژگی به دست آوردیم. ما اول نمی‌دانیم که کدام یک از این ویژگی‌ها تاثیر گذار هست یا نیست. در این پروژه این ویژگی‌ها را استخراج می‌کنیم و در نهایت بعد از plot کردن آن‌ها می‌فهمیم کدام یک از آن‌ها روی تشخیص وقوع یا عدم وقوع زلزله تاثیرگذاری بیشتری داشتند.

نرمال‌سازی داده‌ها

قبل از اعمال الگوریتم‌های یادگیری ماشین باید داده‌ها را استاندارد و نرمال سازی کنیم. چون می‌دانیم برای مثال اگر یک ویژگی عددی داده بین رنج ده هزار و صد هزار باشد و ویژگی دیگر بین ۰.۱ و یک باشد، ویژگی اول تاثیر بیشتری روی نتیجه و classify می‌گذارد و به سمت ویژگی‌هایی با مقدار بیشتر بایاس می‌شویم. همچنین باید ویژگی‌هایی را انتخاب کنیم که informative هستند و به درد بخور باشند تا classify وقت روی ویژگی‌هایی که اطلاعات کافی برای classification ندارند نگذارد. همچنین ممکن است بعضی ویژگی‌ها با یکدیگر correlation یا پیوستگی داشته باشند. به عنوان مثال وقتی فاصله با ماه زیاد شده است، زاویه هم زیاد شده است و این دو ویژگی یک اطلاعات یه ما می‌دهند. برای این کار pca اعمال می‌کنیم و این ستون‌هایی که پیوستگی دارند از این طریق این ستون‌هایی که دارای اطلاعات اضافه هستند حذف یا یکی شده و در نتیجه دیتا کمتر و با توجه به اینکه ویژگی اضافه نداریم سرعت شبکه بیشتر می‌شود.

میانگین، واریانس، ماکسیمم و مینیمم قبل از نرمال‌سازی:

	Lat	Long	...	tetaM	label
count	54058.000000	54058.000000	...	54058.000000	54058.000000
mean	32.920185	52.146044	...	0.053257	0.082042
std	3.833277	4.813332	...	0.144473	0.274430
min	22.095000	41.243000	...	0.000001	0.000000
25%	29.665000	47.930000	...	0.010502	0.000000
50%	32.658000	51.870000	...	0.018560	0.000000
75%	36.088000	56.490000	...	0.036455	0.000000
max	44.090000	66.230000	...	3.007796	1.000000

[8 rows x 55 columns]

میانگین، واریانس، ماکسیمم و مینیمم بعد از نرمال‌سازی:

	Lat	Long	...	tetaM	label
count	54058.000000	54058.000000	...	54058.000000	54058.000000
mean	0.492166	0.436349	...	0.017706	0.082042
std	0.174279	0.192633	...	0.048033	0.274430
min	0.000000	0.000000	...	0.000000	0.000000
25%	0.344169	0.267619	...	0.003491	0.000000
50%	0.480246	0.425301	...	0.006170	0.000000
75%	0.636190	0.610197	...	0.012120	0.000000
max	1.000000	1.000000	...	1.000000	1.000000

[8 rows x 54 columns]

می‌بینیم که میانگین ها،ماکسیمم، مینیمم ها و واریانس ها چقدر به یکدیگر نزدیک و داده‌ها نرمال شدند.

سنجش عملکرد(**F-Measure، Accuracy، Recall،Precision**) و پیشگویی (Prediction)

برای عمل Classification از دو الگویتم svm و رگرسیون لاجستیک استفاده شد که هردو تقریبا عملکرد یکسانی داشتند و svm کمی بهتر عمل کرد. برای برای سنجش عملکرد مدل لیبل پیشبینی شده با لیبل واقعی مقایسه کرده و accuracy، precision، recall، f1 score و confusion matrix را به دست می‌آوریم. می‌بینیم که precision برای داده‌هایی با برچسب صفر یا زلزله با ریشتر کمتر از ۴.۵ در روش رگرسیون لاجستیک ۰.۶۵ و در روش SVM ۰.۶۹ درصد و برای داده‌ها با برچسب ۱ یا زلزله با ریشتر بیشتر از ۴.۵ در روش رگرسیون لاجستیک ۰.۹۵ و در روش SVM ۰.۹۵ درصد می‌باشد. دقت تشخیص داده‌ها با لیبل ۱ در بهترین حالتش نزدیک به ۷۰ درصد است و دلیلش در ادامه تشریح داد خواهد شد. تعداد داده‌هایی که زلزله اتفاق نیفتاده است و لیبل برابر صفر دارند، ۳۴۶۹۰ و تعداد داده‌هایی که لیب ۱ دارند و زلزله اتفاق افتاده است برابر یا ۳۱۵۰ است. اگر یک کلاسی از نمونه‌هایمان خیلی زیاد باشد باعث می‌شود classifyer به سمتی برود که همیشه آن را که تعداد خیلی زیادی دارد درست تشخصی بدهد و برای دیتایی که کم هست (لیبل ۱) اهمیتی قائل نشود. precision ، recall و accuracy خیلی خوب جواب می‌دهد چون آن داده‌هایی که تعدادش بیشتر بوده (لیبل صفر) درست تشخیص داده شده‌اند و classifier به این دیتا ها بایاس می‌شود و این به این دلیل است که توزیع داده‌های ما یکسان نیست و در نتیجه مثلا accuracy معیار مناسبی نیست و در با اینکه دقت بالاست به نظر می‌رسد به این classifyerها اعتمادی نیست.

عملکرد رگرسیون لاجستیک:

	precision	recall	f1-score	support
0	0.95	0.98	0.96	9885
1	0.65	0.48	0.56	927
accuracy			0.93	10812
macro avg	0.80	0.73	0.76	10812
weighted avg	0.93	0.93	0.93	10812

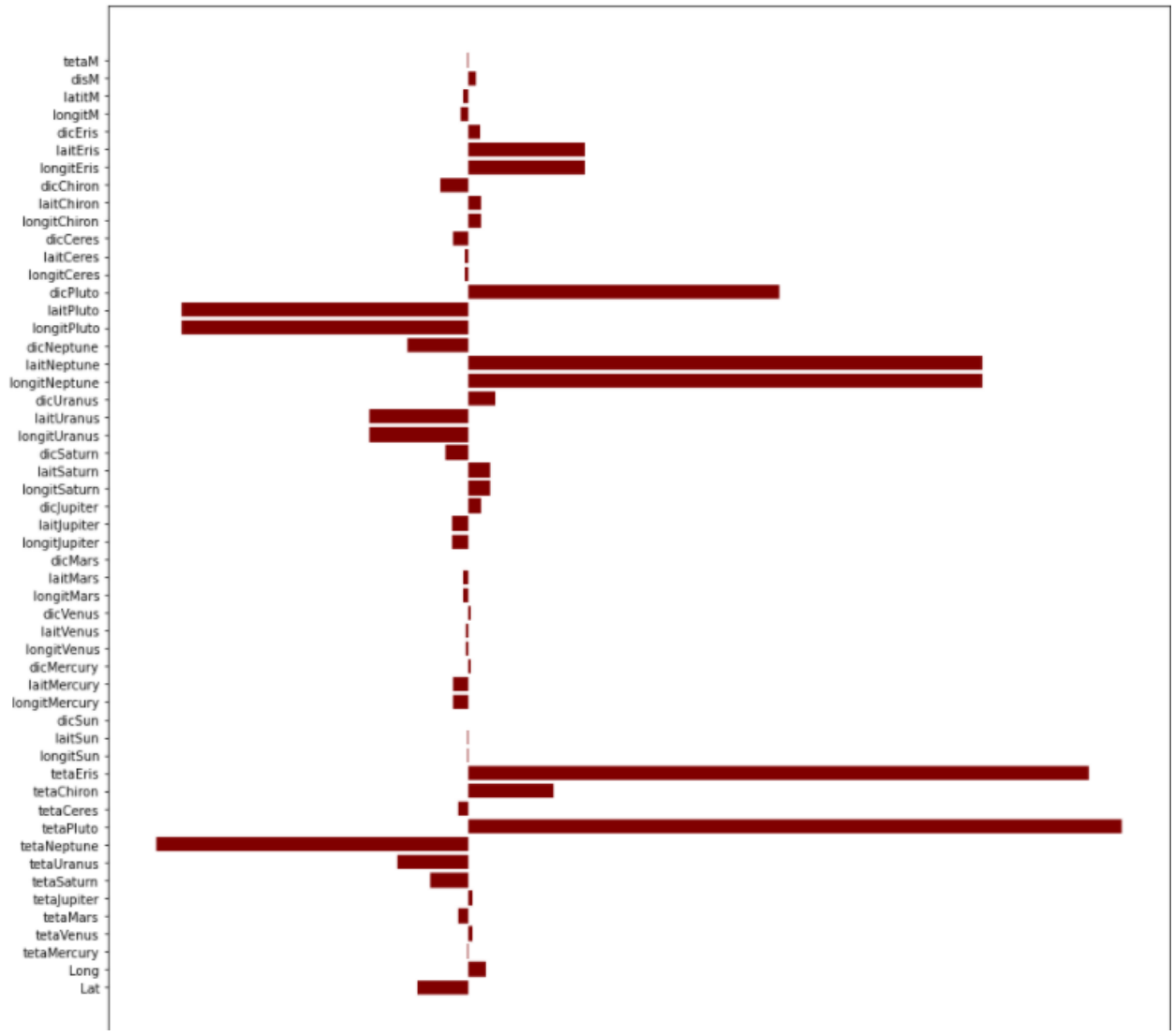
عملکرد SVM:

	precision	recall	f1-score	support
0	0.95	0.98	0.96	9885
1	0.69	0.43	0.53	927
accuracy			0.93	10812
macro avg	0.82	0.71	0.75	10812
weighted avg	0.93	0.93	0.93	10812

توی الگوریتم SVM، هر نمونه داده را به عنوان یک نقطه در فضای n بعدی روی نمودار پراکندگی داده‌ها رسم کرده (n تعداد ویژگی‌هایی است که یک نمونه داده دارد) و مقدار هر ویژگی مربوط به داده‌ها، یکی از مؤلفه‌های مختصات نقطه روی نمودار را مشخص می‌کند. سپس، با ترسیم یک خط راست، داده‌های مختلف و متمایز از یکدیگر دسته‌بندی می‌شوند.

همچنین برای پیشگویی، تاریخ، زمان و موقعیت مکانی به classifier می‌دهیم و یک لیبل به ما می‌دهد که به احتمال زیاد می‌گوید زلزله نیامده است(لیبل خروجی اکثرا صفر است چون احتمال زلزله آمدن پایین است).

Feature Importance به دست آمده از روش رگرسیون لاجستیک:

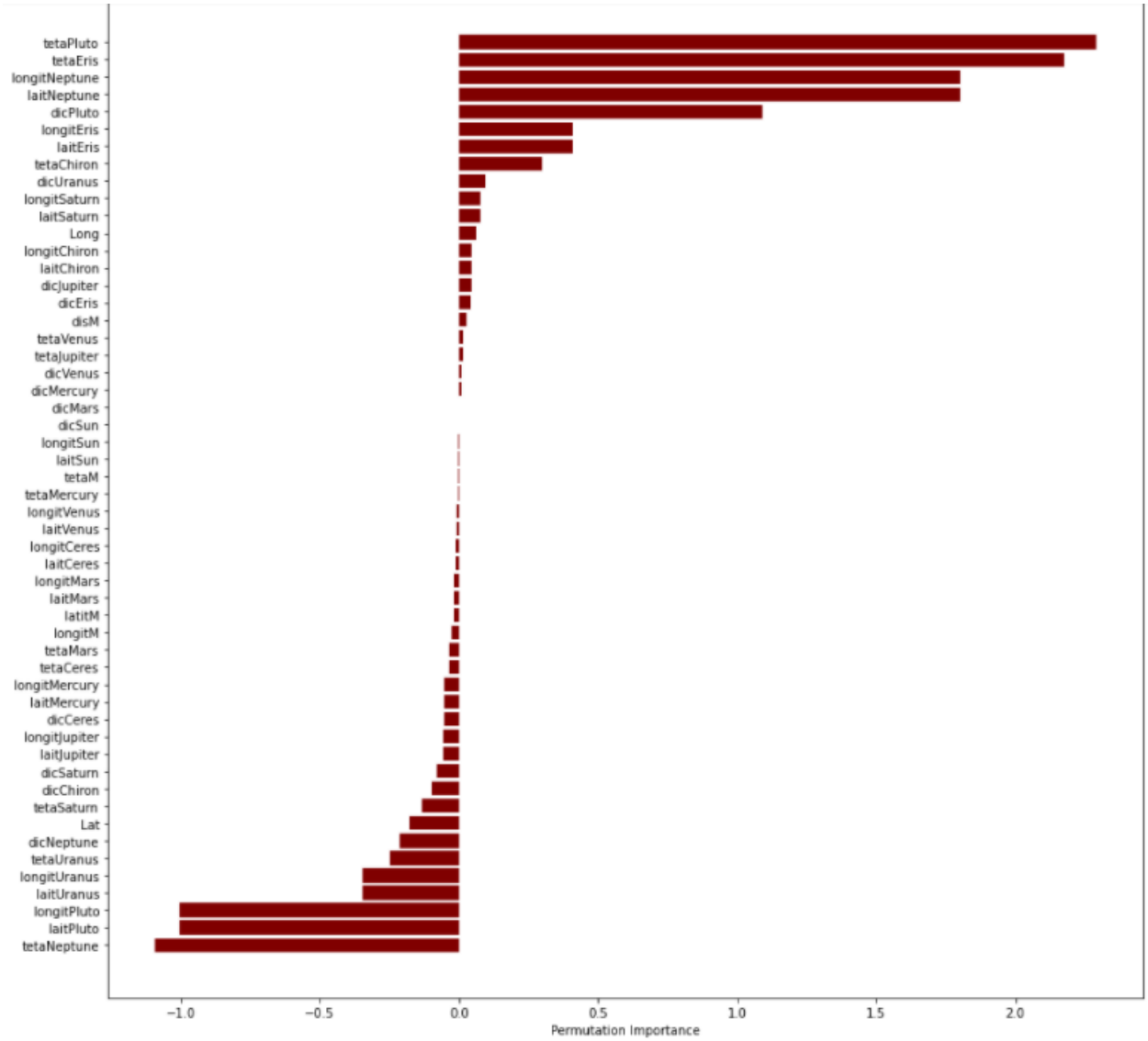


مقادیر عددی ویژگی‌ها از بالا به پایین در نمودار بالا:

```
Feature: 0, Score: -0.17766 Feature: 1, Score: 0.06545 Feature: 2, Score:
-0.00356 Feature: 3, Score: 0.01613 Feature: 4, Score: -0.03426 Feature:
5, Score: 0.01378 Feature: 6, Score: -0.13074 Feature: 7, Score: -0.24797
Feature: 8, Score: -1.09175 Feature: 9, Score: 2.29173 Feature: 10, Score:
-0.03574 Feature: 11, Score: 0.29833 Feature: 12, Score: 2.17592  Feature:
13, Score: -0.00191 Feature: 14, Score: -0.00191 Feature: 15, Score:
0.00094 Feature: 16, Score: -0.05109 Feature: 17, Score: -0.05109
Feature: 18, Score: 0.00794 Feature: 19, Score: -0.00934 Feature: 20,
Score: -0.00934 Feature: 21, Score: 0.01034 Feature: 22, Score: -0.01654
```

```
Feature: 23, Score: -0.01654 Feature: 24, Score: 0.00241 Feature: 25,
Score: -0.05798 Feature: 26, Score: -0.05798 Feature: 27, Score: 0.04626
Feature: 28, Score: 0.07477 Feature: 29, Score: 0.07477 Feature: 30,
Score: -0.07680 Feature: 31, Score: -0.34496 Feature: 32, Score: -0.34496
Feature: 33, Score: 0.09582 Feature: 34, Score: 1.80030 Feature: 35,
Score: 1.80030 Feature: 36, Score: -0.21362 Feature: 37, Score: -1.00532
Feature: 38, Score: -1.00532 Feature: 39, Score: 1.08874 Feature: 40,
Score: -0.01289 Feature: 41, Score: -0.01289 Feature: 42, Score: -0.05275
Feature: 43, Score: 0.04710 Feature: 44, Score: 0.04710 Feature: 45,
Score: -0.09899 Feature: 46, Score: 0.41138 Feature: 47, Score: 0.41138
Feature: 48, Score: 0.04012 Feature: 49, Score: -0.02695 Feature: 50,
Score: -0.01849 Feature: 51, Score: 0.02945 Feature: 52, Score: -0.00216
```

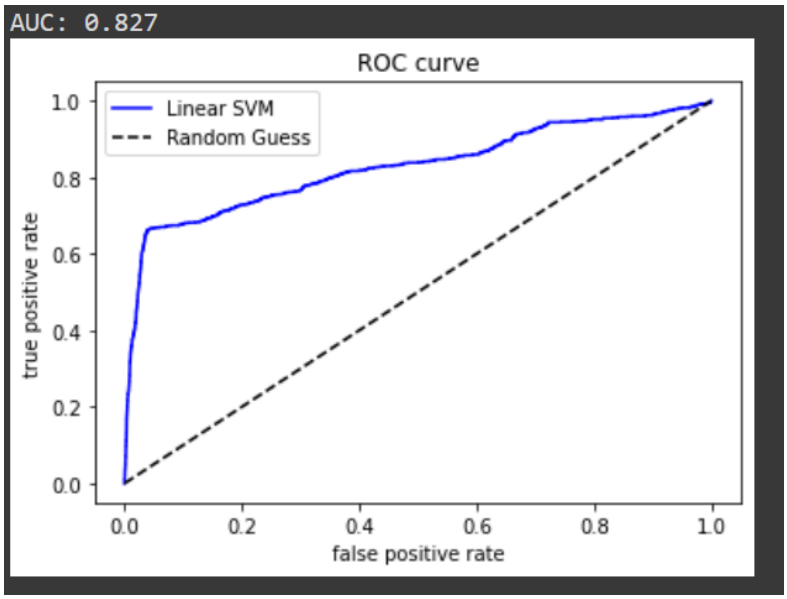
نمودار مرتب شده:



متوجه می‌شویم که چهار ویژگی اول از بالا در نمودار فوق بیشترین تاثیر را در classification دارند.

AUC و منحنی مشخصه عملکرد ROC

این منحنی توسط ترسیم true positive rate بر حسب false positive rate ایجاد می‌شود. بهترین عملکرد calssification در نقطه (۰,۱) رخ خواهد داد که بیشترین نرخ tpr را داریم. خطی که (۰,۰) را به (۱,۱) وصل می‌کند حدس تصادفی است. مساحت سطح زیر این منحنی AUC یا Area Under the ROC Curve است و نشان می‌دهد قدرت درستی نتایج یک آزمون چقدر می‌باشد و این به این بستگی دارد که آزمونمان چقدر توانایی درست تشخیص دادن true positive و True negative دارد. اگر AUC به یک نزدیک باشد نشان می‌دهد داده ها بالای خط نیمساز قرار دارند و روش classifier از قدرت تشخیص و درستی خوبی برخوردار است. برای SVM، مقدار AUC عدد قابل قبول ۰.۸۲۷ به دست آمد. ولی باز هم به دلایلی که قبل تر گفته شد خیلی قابل اطمنینان نیست.



لینک کد:

https://colab.research.google.com/drive/1Sas8SIdLdMZPwCb4xtXC-TJdyDRiF-OR