به نام خدا

محمدعلی رشادی

گزارش تمرین سری ۲

تمرین ۱.

۱.

از یک مجموعه تهی شروع کرده و هر ویژگی را به مجموعه اضافه کرده و معیار AUC را سنجیدیم. بدیهی است که وقتی AUC به مقدار ۱۰۰ برسد دیگر ویژگی جدیدی لازم نیست به مجموعه اضافه شود، الگوریتم انتخاب ویژگی پیشرو با هفت ویژگی زیر توانست به مقدار AUC ماکسیمم برسد.

['battery_power', 'blue', 'dual_sim', 'mobile_wt', 'ram', 'sc_h', 'talk_time', 'three_g']

۲.

نتایج لاجستیک رگرسیون روی الگوریتم نوشته شده بدین صورت است :

	precision	recall	f1-score	support
0	0.90	0.88	0.89	297
1	0.88	0.91	0.90	303
accuracy			0.89	600
macro avg	0.89	0.89	0.89	600
weighted avg	0.89	0.89	0.89	600

۳.

PCA را با n_components=7 ساختیم.

.۴

نتایج PCA روی داده بدین صورت است که نشان دهنده برابری نتایج با الگوریتم پیاده سازی شده و موفق بودن آن است.

	precision	recall	f1-score	support
0	0.89	0.88	0.89	297
1	0.88	0.90	0.89	303
accuracy			0.89	600
macro avg	0.89	0.89	0.89	600
weighted avg	0.89	0.89	0.89	600

.۵,۶

آ) ویژگی battery power را با مقادیر مختلف به ۳ bin تقسیم کرده و برای حل مشکل در مدلسازی برای bin ها از لیبل ۱و۲و۳ استفاده کردیم.

	battery_power	blue	clock_speed	dual_sim	fc	four_g	int_memory
0	1	0	2.2	0	1	0	7
1	1	1	0.5	1	0	1	53
2	1	1	0.5	1	2	1	41
3	1	1	2.5	0	0	0	10
4	3	1	1.2	0	13	1	44

ب) نتایج وان هات انکدینگ را شاهد هستیم.

fc_18	fc_19	four_g_0	four_g_1	three_g_0	three_g_1	touch_screen_0	touch_screen_1	wifi_0	wifi_1
0	0	1	0	1	0	1	0	0	1
0	0	0	1	0	1	0	1	1	0
0	0	0	1	0	1	0	1	1	0
0	0	1	0	0	1	1	0	1	0
0	0	0	1	0	1	0	1	1	0

ج) با توجه به بررسی و کاوش های انجام شده روی دیتاست در تمرین قبلی میدانیم ویژگی خاصی نداریم که چولگی داشته باشد و از طرفی دیتاهای زمانی نیز نداریم پس نیاز خاصی به تبدیل لگاریتمی و نمایی ندارین بلکه به دلیل بازه های متفاوت ویژگی ها ترجیح دادیم از نرمالسازی استاندارد بهره ببریم تا ویژگی ها نسبت به یکدیگر تاثیر یکسانی در مدلسازی داشته باشند.

د) با استفاده از طول و عرض گوشی ویژگی مساحت را به وجود آوردیم.

masahat
15120
1799140
2167308
2171776
1464096

۷)

آ)

	precision	recall	f1-score	support
0	0.93	0.94	0.93	297
1	0.94	0.93	0.94	303
accuracy			0.94	600
macro avg	0.93	0.94	0.93	600
weighted avg	0.94	0.94	0.94	600

ب)

	precision	recall	f1-score	support
0	0.97	1.00	0.99	297
1	1.00	0.97	0.98	303
accuracy			0.98	600
macro avg	0.99	0.99	0.98	600
weighted avg	0.99	0.98	0.98	600

ج)

	precision	recall	f1-score	support
0	0.94	0.97	0.95	297
1	0.97	0.94	0.95	303
accuracy			0.95	600
macro avg	0.95	0.95	0.95	600
weighted avg	0.95	0.95	0.95	600

د)

	precision	recall	f1-score	support
0	0.53	0.79	0.64	297
1	0.61	0.32	0.42	303
accuracy			0.55	600
macro avg	0.57	0.56	0.53	600
weighted avg	0.57	0.55	0.53	600

همه)

	precision	recall	f1-score	support
0	0.94	0.94	0.94	297
1	0.94	0.94	0.94	303
accuracy			0.94	600
macro avg	0.94	0.94	0.94	600
weighted avg	0.94	0.94	0.94	600

سوالات تئوری :

۸.

در Bootstrapping ، داده‌های آموزش به صورت تصادفی و با استفاده از جایگذاری انتخاب می‌شوند. نمونه‌هایی که انتخاب نشده‌اند نیز برای تست مورد استفاده قرار می‌گیرد. در این روش، بر خلاف کراس ولیدیشن تعداد نمونه‌های انتخاب شده در هر تکرار متفاوت است. نرخ خطای مدل در این روش نیز برابر با میانگین نرخ خطا در هر تکرار است. زمانی که توزیع داده های ما نرمال نبود می توانیم از این روش برای استنباط و بررسی معناداری ضرایب آماری خود بهره ببریم.

۹.

به ۵ تکرار از ۲-fold، 5x2 cross validation گفته می شود. به عبارتی دیگر دیتای آموزش و ولیدیشن به نسبت ۵۰-۵۰ جدا میشود و این عمل ۵ بار تکرار می پذیرد.

این روش به عنوان راهی برای به دست آوردن نه تنها تخمین خوب از خطای تعمیم، بلکه همچنین برآورد خوبی از واریانس آن خطا (به منظور انجام آزمون‌های آماری) رایج شد.

۱۰.

با توجه به مفهوم بایاس و واریانس نمودار elbow به دنبال این است که خوشه های مختلف را از همدیگر بتواند به خوبی جدا کرده و به یک پیچیدگی مناسب دست پیدا کند، درست است که از لحاظ تحلیلی این پیچیدگی تا حدی قابل قبول است ولی در عمل این تضمین نمیکند که ما در واقعیت هم همین تعداد خوشه داشته باشیم بلکه ممکن است خوشه های مختلف با یکدیگر ارتباط تنگاتنگی داشته باشند.

سوال امتیازی

۳.

معیار MCC بیان‌گر کیفیت کلاس‌بندی برای یک مجموعه باینری می‌باشد. (MCC (Matthews correlation coefficient، سنجه‌ای است که بیان‌گر بستگی مابین مقادیر مشاهده شده از کلاس باینری و مقادیر پیش‌بینی شده از آن می‌باشد. مقادیر مورد انتظار برای این کمیت در بازه ۱- و ۱ متغیر می‌باشد. مقدار ۱+، نشان دهنده پیش‌بینی دقیق و بدون خطای الگوریتم یادگیر از کلاس باینری می‌باشد. مقدار ۰، نشان دهنده پیش‌بینی تصادفی الگوریتم یادگیر از کلاس باینری می‌باشد. مقدار ۱-، نشان دهنده عدم تطابق کامل مابین موارد پیش‌بینی شده از کلاس باینری و موارد مشاهده شده از آن می‌باشد.

تمرین ۲.

۱.

با استفاده از ۳ ویژگی گفته شده به ساخت الگوریتم رگرسیون خطی با تابع خطا MSE پرداختیم که نتیجه MSE روی داده آزمایش به صورت زیر حاصل شد :

MSE: 3154787.5153122097

۲.

با استفاده از پکیج سایکیت لرن نیز رگرسیون خطی ساخته و روی داده آزمایش تست کردیم که خطای MSE به شرح ذیل حاصل شد:

MSE: 3240131.3034362276

همانطور که می بینیم خطای MSE هم در الگوریتم پیاده سازی شده توسط ما و هم پکیج آماده تقریبا یکسان بوده و همچنین بزرگ است پس نتیجه می شود سه ویژگی به تنهایی برای مدلسازی کافی نبوده است و باید از ویژگی های دیگر نیز بهره برد.

۳.

اینبار پس از پیش پردازش و حذف داده های نامربوط و استفاده از ویژگی های بیشتر به ساخت مدل پرداختیم که نتایج به شرح ذیل حاصل شد:

Ridge Regression MSE : 27.39234027083926

Lasso Regression MSE : 28.436480247714194

به وضوح می بینیم که این بار خطا به شدت کاهش یافت و تعداد ویژگی های بیشتر توانست عملکرد مدل را به شدت بهبود بخشد.

سوال امتیازی

۱. در این بخش رگرسیون رو با خطای Absolute Error پیاده کردیم که خطا در داده آزمایش به شرح ذیل حاصل شد :

Absolute Error = 29.162655360158997