به نام خداوند رنگین کمان

نیمسال اول 1402-1401

Machine Learning
Homework-1

استاد درس: دکتر منصور رزقی آهق

دانشجو: بهار مهدوی
40152521337

Tarbiat Modares University

تمرین 1:

از داده های ضمیمه شده به دلخواه سه مورد را انتخاب کنید. داده ها را به ۸۰ درصد داده آموزشی و ۲۰ درصد داده آزمایشی تقسیم کنید و روش های زیر را اعمال کنید.

اطلاعات دیتاست های انتخابی در سایت GEO:

GLI-85:

Accession	Title	Source name	Tissue	Histology	Characteristics	Age	Gender	Grade	Living	Survival time	Survival cluster	Hc_74_samples	Hc_85_samples	Array
GSM99588	brain, Thalamus: Extract51_le1	brain, Thalamus	brain, Thalamus	GBM	tumor number (as appears in figure 3): 1681	40	MALE	4	ALIVE	927	SC1	HC2B	HC2B	133A
GSM99482	brain, Right Temporal Parietal: Extract83_le1	brain, Right Temporal Parietal	brain, Right Temporal Parietal	GBM	tumor number (as appears in figure 3): 1032	34	FEMALE	4	DECEASED	90	SC2	HC1B	HC1B	133A

CLL-SUB-111:

Group	Accession	Title	Source name
-	GSM46667	V0455	Peripheral blood lymphocytes
-	GSM46668	V0457	Peripheral blood lymphocytes

SMK-CAN-187:

Accession	Title	Source name	Sample from smoker NOT diagnosed with cancer	Sample_id	Age	Gender	Race	Cancer_status	Smoking_status	Packyears	Bronch_status	Presence_of_hemoptysis	Presence_of_lymphadenopathy	Mass_size_greater_than_3cm	Biomarker_score	Subjective_assessment	Sample from smoker diagnosed with cancer	Sample from smoker with suspect lung cancer
GSM93997	Smoker NOT diagnosed with cancer Sample 283	Bronchial Epithelium		283	34	M	OTHER	No Cancer	quit less than 10 years	17	non-diagnostic	0	0	0	-2.253539513	Low		
GSM94019	Smoker diagnosed with cancer Sample 57	Bronchial Epithelium		57	63	M	CAU	Cancer	quit less than 10 years	75	diagnostic	0	1	1	8.900589368	High		

ابتدا داده ها از فرمت mat. به فرمت csv. تبدیل و پس از فراخوانی به روش train_test_split به Train 80% و Test 20% تقسیم گردید. (در تقسیم داده ها به Train 80% و Test 20% به روش KFold (k=5 fold) چون دقیقا نمی‌دونستم چطور دقت هر fold را بدست آورده و بعد میانگین یا بهترین آن را برای ادامه مراحل بعدی انتخاب کنم مرتب خطا گرفته و بنابراین از بکارگیری آن صرفه نظر کرده و در ادامه از روشی دیگر برای تقسیم بندی داده ها استفاده شد)

الف) از روش های کلاس بندی مبتنی بر رگرسیون خطی با تابع های هزینه ی MAE و MSE برای کلاس بندی این داده ها استفاده کنید. دقت آزمایش (Test) و آموزش (Train) را گزارش کنید.

Dataset	MAE (Train)	MSE (Train)	Train Accuracy	MAE (Test)	MSE (Test)	Test Accuracy (1-MSE)
GLI-85	0.0	0.0	100 %	0.28	0.13	87 %
CLL-SUB-111	0.0	0.0	100 %	0.44	0.33	69 %
SMK-CAN-187	0.0	0.0	100 %	0.37	0.20	80 %
Mean	0.0	0.0	100 %	0.36	0.22	~ 78.5 %

در Linear regression میانگین دقت train 100% و دقت test حدود 78.5% بوده و MSE خطای کمتری را نسبت به MAE نشان میدهد.

ب) همچنین از روش های Lasso و Ridge نیز برای اینکار اســتفاده نمایید. ← از چه روشــی برای تخمین بهترین پارامتر می توان استفاده کرد؟ دقت این روش‌ها را برحسب معیار (Accuracy) گزارش کنید.

Dataset	Lasso Accuracy (%)	Ridge Accuracy (%)	Linear regression Accuracy (%)
GLI-85	64.70588235294117	52.94117647058824	87
CLL-SUB-111	60.86956521739131	52.17391304347826	69
SMK-CAN-187	55.263157894736835	55.263157894736835	80
Mean	60.15336	53.44342	78.66666

من آلفا را بین رنج 0.02 تا 2 به طور Random تغییر داده و ســعی کردم بهترین آن را انتخاب کنم ولی متاســفانه روش غیر Manual آن برای تخمین دقیق بهترین پارامتر را پیدا نکرده و به همین خاطر، متاســفانه بر خلاف انتظار از روش Linear regression نســبت به روش های Lasso و Ridge، درصــد Accuracy بالاتر و بهتری را گرفتم که نشان میدهد مدل های بکار برده شده یا دارای خطاست و یا خوب کار نکرده است!

ج) از روش Logistic Regression Classifier برای کلاس‌بندی این داده‌ها استفاده کنید. نتایج را با روش‌های قبلی مقایسه کنید

Dataset	Classification Acc. (%)	Logistic Acc. (%)	Lasso Acc. (%)	Ridge Acc. (%)	Linear regression Acc. (%)
GLI-85	82	17.647058823529417	64.70588235294117	52.94117647058824	87
CLL-SUB-111	74	26.086956521739136	60.86956521739131	52.17391304347826	69
SMK-CAN-187	74	26.315789473684216	55.263157894736835	55.263157894736835	80
Mean	76.6666	22.96699	60.15336	53.44342	78.66666

از روش Logistic Regression Classifier برای Classification اســتفاده شــده اســت و مطمئن نیســتم از چه نظر میتوان درصــد دقت بدســت آمده آن را با روشهای Regression قبلی مقایسه کرد، به همین خاطر به گزارش آن در جدول بسنده شد...

کدهای مرتبط با تمرین 1:

- **GLI-85_Mahdavi.ipynb**
- **CLL-SUB-111_Mahdavi.ipynb**
- **SMK-CAN-187_Mahdavi.ipynb**

تمرین 2:

داده های فوق را در نظر بگیرید. (این مسئله را برای داده های ORL نیز انجام دهید.)

$$X = [X_1, X_2, \ldots, X_N] \in \mathbb{R}^{D \times N}$$

Dataset	D (feature)	N (sample)	Train (80%)	Test (20%)	K=30 % N Train	Test Accuracy (%)
GLI-85	22283 (16383 .csv)	85	68	17	20	79.40994234925168
CLL-SUB-111	11340	111	89	22	27	51.68301734818086
SMK-CAN-187	19993 (16383 .csv)	187	150	37	45	61.70868623938255
ORL10P	10304	100	80	20	24	-754.2800761350527 ?!

که D تعداد ویژگی ها و N تعداد نمونه‌های مربوط به داده آموزشی باشد، آنگاه روش فوق را برای کلاس بندی داده‌ها به کار ببرید. اگر Y داده آزمایشی (تست) باشد در این صورت:

$$\min ||Y - XW||_2^2 \text{ , } ||W||_0 < k \oplus$$

که k برابر ۳۰٪ داده های آموزشی است.

مسئله $\oplus$ توسط OMP می‌تواند حل شود. بعد از حل این مسئله عنصر iام W یعنی W_i میزان تاثیر داده X_i در ساختن Y را نشان می‌دهد. چگونه می توان از این W_iها برای کلاس بندی استفاده کرد؟ پیاده سازی (کدنویسی) کنید.

OMP یک روش حل sparse regression بوده و معمولا در feature selection داده های با تعداد feature بالا مثل داده های زیستی به کار میرود. در این مسئله قصد بر این است که تعداد زیادی از W ها صفر شده و حداکثر 30% از آنها non-zero باقی بمانند تا از این طریق تعداد زیادی از feature ها عملا صفر و حذف گردند.

کدهای مرتبط با تمرین 2:

- GLI-85_OMP_Mahdavi.ipynb
- CLL-SUB-111_OMP_Mahdavi.ipynb
- SMK-CAN-187_OMP_Mahdavi.ipynb
- ORL10P_OMP_Mahdavi.ipynb