(هامون سجادی – حسین جمشیدیان)

(decision tree,perceptron,…):sider

* بارگذاری کتابخانه ها و خواندن فایل داده ها و impute کردن داده های گمشده در feaure ها با استفاده از استراتژی میانگین.

* الگوریتم درخت تصمیم را روی داده های اموزشی fit میکنیم .

* یکی از داده های اموزشی ما را بعنوان مثال انتخاب کرده و میبینیم به درستی پیش بینی میکند.

* دقت در درخت تصمیم 591. بدست می اید .

* یک loss function به نام log_loss تعریف میکنیم که همان cross_entrophy است که میزان عدم اطلاعات را میدهد و هر چه بالاتر باشد یعنی احتمال پیش بینی غلط بیشتر میشود.

* cross-validation مقدار دقت را 59.53 با انحراف معیار 2.71 میدهد که نتیجه میدهد که دقت بدست امده از test set تصادفی نبوده است.

* توضیحات learning curve زیر نمودار درخت تصمیم نوشته شده است .

*حال الگوریتم های مختلف را اجرا و دقت ان ها را با هم مقایسه کرده و با استفاده از cross validation از دقت بدست امده اطمینان حاصل کرده و با استفاده از grid search بهترین هایپر پارامترها را پیدا میکنیم .

* بعد از درخت تصمیم random forest را از نظر دقت مورد بررسی قرار میدهیم که تا اینجا الگوریتم برگزیده ی ما است.

* perceptron تایید نمیشود چه از نظر دقت و چه از نظر نمودار .

* logistic regression عملکرد بهتری از نظر دقت و خطا داد پس تا اینجا بهترین الگوریتم ما است .

* و درنهایت به ترتیب 1- kernel svm 2- xgboost 3- random forest 4- knn 5- logistic regression 6- linear svm 7- decision tree 8- perceptron بهترین به بدترین مدل های ما تا به اینجا بوده اند .

* از نظر average precision و roc curve نیز مورد بررسی قرار میدهیم که باز هم kernel svm بعنوان بهترین مدل انتخاب میشود.

*الگوریتم bagging را روی decision tree و knn و logistic regression و linear svm و kernel svm و xgboost اجرا میکنیم که بجز decision tree بر روی بقیه تاثیر بسزایی ندارد .

* بر روی bagging هایی که بدست اورده ایم و xgboost و kernel svm باز هم از نظر AP و ROC curve مقایسه میکنیم که باز kernel svm عملکرد بهتری دارد.

*در نهایت gradian boosting و adaboosting را هم اجرا میکنیم و با اینکه gradiant boosting دقت خوبی را با براورد کننده ی decision tree دارد ولی kernel svm با دقت 69.53 و همچنین از نظر AP=.78 از بین الگوریتم های بالا بهترین الگوریتم برای این نوع داده هاست .

جدول کلی از دقت ها و بهترین مدل ها از نظر دقت :

Model name	acuuracy	Cross-val score std	Acuuracy with grid search and best parametr	rank
Decision tree	59.10	59.53 2.71	60.53	12
perceptron	63.02	60.09 5.70	-	11
Logistic regression	66.38	63.55 4.24	63.74	5
knn	64.98	64.67 3.09	66.92	6
Linear svm	67.50	64.71 4.41	63.74	4
Kernel svm	69.53	67.38 4.35	69.53	1
xgboost	68.34	62.90 4.26	-	2
Bagging for decision tree	64.98	-	-	8
Gradinat boosting	67.50	-	-	3
Ada boosting	62.18	-	-	10
Bagging for kernel svm	63.86	-	67.38	9
Random forest	64.42	-	67.10	7

(decision tree,perceptron,…): Tox21

* لیبل ها دارای داده های Nan هستند و چون یک taskداریم میتوانیم سطرهایی که لیبل Nan دارند را حذف کنیم و همچنین روی feature ها mutation را انجام می دهیم .

* اولین مدل را درخت تصمیم میگیریم و دقت 9735. را برای test set بدست می اوریم و cross validation نیز مقدار دقت را 97.25با انحراف معیار 44. میدهد که دقت را که از test set بدست اورده ایم را تایید میکند.

* الگوریتم ها را به ترتیب چک میکنیم و در نهایت به ترتیب 1- xgboost 2- kernel svm 3- knn 4- linear svm 5- logisitic regression 6- decision tree 7- perceptron که هم از نظر دقت و هم از نظر cross validation و مقادیر دقت و انحراف معیاری که از ان بدست می اید و همچنین از نظر AP بررسی کرده ایم و xgboost تا اینجای کار بعنوان بهترین مدل انتخاب میشود .

* حال به سراغ ensemble ها میرویم و random forest و adaboosting و gradiant boosting و bagging را برای decision tree و knn و logistic regression و linear svm و xgboost اجرا میکنیم .

* با کمی تغییر در پارامتر های gradiant boosting , adaboosting به دقت های خوبی میرسیم که ada boosting بهترین دقت با مقدار 9752. میدهد .

* از نظر دیگر متریک ها مثل AP مدل bagging decision tree بهترین میشود .

*از نظر ROC curve مدل random forest بهترین میشود .

* ولی برای داده های طبقه بندی بهتراست که از نظر دقت مورد بررسی قرار دهیم بهترین مدل را برای مدل های بالا adaboosting با براوردگر decision tree انتخاب میکنیم ولی داده های های ما بالانس نیستند به همین دلیل از یک متریک دیگر استفاده میکنیم که در نهایت bagging for decision tree را انتخاب میکنیم .

جدول کلی از دقت ها و بهترین مدل ها از نظر دقت :

Model name	acuuracy	Cross-val score std	Acuuracy with grid search and best parametr	Average precision	rank
Decision tree	.9735	97.25 .44	97.25	.43	9
Perceptron	.9598	96.29 .86	96.28	-	11
Logistic regression	.9686	97.25 .27	97.38	.49	8
knn	.9724	97.01 .5	97.15	.51	6
Linear svm	.9708	97.25 .34	97.30	.51	7
Kernel svm	.9702	97.36 .36	97.47	.53	5
Xgboost	.9708	97.39 .43	-	.54	4
Random forest	.9686	-	97.43	.53	10
Adaboosting for decision tree	.9752	-	-	.49	1
Gradiant boosting	.9741	-	-	.51	3
Bagging for decision tree	.9746	-	-	.54	2