به نام خدا

گزارش دوم کارآموزی

بررسی هفت الگوریتم ماشین‌لرنینگ و دیپ‌لرنینگ

بر روی ده دیتاست انتخابی

طیبه محبی

فهرست

مقدمه:

در هر بخش، نتیجه تست ۷ الگوریتم روی یک دیتاست نشان داده شده است. دیتاست‌های بررسی شده عبارت اند از:

1. Wine
2. Breast cancer w
3. Balance Scale
4. Dermatology
5. Diabetes
6. Vowel
7. Glass
8. Titanic
9. Iris
10. Ionosphere

الگوریتم‌های استفاده شده عبارت اند از:

1. Logistic Regression
2. Random Forest
3. SVG
4. KNN
5. Decision Tree
6. XGBoost
7. Neural Network

تمامی کدهای استفاده‌شده، در فایل زیپ با فرمت ipynb موجودند و قابل ران‌کردن می‌باشند. دیتاست‌های استفاده‌شده نیز در فایل زیپ موجود می‌باشند.

تنها معیار استفاده‌شده برای سادگی در این گزارش، accuracy یا دقت است. مسلما بررسی معیارهایی همچون f1-score نیز با استفاده از پایتون کار مشکلی نیست و به راحتی می‌توان با ایجاد تغییرات کوچکی، این معیارها را نیز بررسی کرد.

دیتاست اول: Wine

ALGORITHM NAME	ACCURACY
Logistic Regression	**0.93333**
Random Forest	**0.97778**
SVG	**0.82222**
KNN	**80.0**
Decision Tree	**0.91111**
XGboost	**0.95556**
Neural Network Batch Size: 32 Epoch: 75	**100.0**

- با وجود کوچک بودن دیتاست، شبکه عصبی نتیجه بسیار خوبی از خود نشان داده و با دقت ۱۰۰ درصد پیش بینی را انجام داده است.
- این نتیجه با تغییر تعداد لایه ها، متناسب کردن فیچرها و تنظیم بچ سایز و ایپاک به دست آمده است.
- اا‌ز بین دیگر الگوریتم ها، رندوم فارست بهترین نتیجه را به دست آورده است.
- تعداد کلاس ها، سه میباشد.

دیتاست دوم: BreastW

ALGORITHM NAME	ACCURACY
Logistic Regression	**0.94737**
Random Forest	**0.95322**
SVG	**0.95322**
KNN	**94.73684210526315**
Decision Tree	**0.94737**
XGboost	**0.95322**
Neural Network Batch Size: 32 Epoch: 100	**98.54014598540147**

- شبکه عصبی اینجا هم از تمامی الگوریتم‌های دیگر بهتر عمل کرده است.
- لازم به ذکر است تقریبا تمامی دیتاست های استفاده شده، نسبتا کوچک هستند و زیر ۱۰۰۰ رکورد دارند، به همین دلیل نمیتوان از شبکه عصبی، انتظار دقت بالایی داشت. اما در اینجا میبینیم که با تنظیم تعداد لایه ها و دنسیتی آن ها، انتخاب ایپاک و بچ سایز مناسب و انتخاب درست فیچر ماتریس، میتوان حتی در داده های کوچک نیز به نتیجه مناسبی رسید.
- از بین دیگر الگوریتم ها، رندوم فارست و ایکس جی بوست بهترین نتیجه را به دست آورده اند.
- تعداد کلاس ها، دو میباشد.

دیتاست سوم: Balance Scale

ALGORITHM NAME	ACCURACY
Logistic Regression	**0.89172**
Random Forest	**0.83439**
SVG	**0.93631**
KNN	**89.80891719745223**
Decision Tree	**0.75796**
XGboost	**0.87898**
Neural Network Batch Size: 32 Epoch: 100	**97.6**

- اینجا هم شبکه عصبی بهترین نتیجه را به دست آورده و از دقت خوبی برخوردار است.
- از بین دیگر الگوریتم ها، SVG بهترین نتیجه را به دست آورده است.
- تعداد کلاس ها، سه می باشد.

دیتاست چهارم: Dermatology

ALGORITHM NAME	ACCURACY
Logistic Regression	**0.94444**
Random Forest	**0.93333**
SVG	**0.65556**
KNN	**87.77777777777777**
Decision Tree	**0.86667**
XGboost	**0.93333**
Neural Network Batch Size: 32 Epoch: 75	**98.61111111111111**

- شبکه عصبی بهترین نتیجه را نشان داده است.
- از بین دیگر الگوریتم ها، لاجیستیک ریگرشن بهترین نتیجه را آورده است.
- تعداد کلاس ها، شش می باشد.

دیتاست پنجم: Diabetes

ALGORITHM NAME	ACCURACY
Logistic Regression	**0.79167**
Random Forest	**0.78125**
SVG	**0.77083**
KNN	**76.04166666666666**
Decision Tree	**0.71354**
XGboost	**0.76042**
Neural Network Batch Size: 32 Epoch: 50	**78.57142857142857**

- لاجیستیک ریگرشن، بهترین نتیجه را داده است.
- از بین دیگر الگوریتم ها، شبکه عصبی از بهترین دقت برخوردار بوده است.
- میبینیم که اینجا شبکه عصبی دقت خیلی بالایی (هرچند قابل قبول) ندارد. پس باید آزمایش های بیشتری انجام داد تا دید آیا میتوان دقت آن را افزایش داد یا خیر.

دیتاست ششم: Vowel

ALGORITHM NAME	ACCURACY
Logistic Regression	**0.66129**
Random Forest	**0.96371**
SVG	**0.84677**
KNN	**96.37096774193549**
Decision Tree	**0.75806**
XGboost	**0.91935**
Neural Network Batch Size: 32 Epoch: 100	**54.04040404040404**

- الگوریتم رندوم فارست و KNN، بالاترین دقت را داشته اند.
- در این دیتاست، بین دقت الگوریتم های مختلف تفاوت بسیار زیادی وجود دارد!
- این مسئله خود جای بررسی دارد و احتمالا باید بعضی از ستون ها و فیچر ها بیشتر بررسی شوند تا ببینیم این مسئله چه دلیلی دارد
- شبکه عصبی دقت بسیار پایینی از خود نشان داده
- که این مورد احتمالا به دلیل زیادی تعداد کلاس ها(۱۱ تا)، کم بودن داده ها و متناسب نبودن فیچرهای انتخاب شده برای ساخت مدل بوده است.

دیتاست هفتم: Glass

ALGORITHM NAME	ACCURACY
Logistic Regression	**0.66667**
Random Forest	**0.87037**
SVG	**0.42593**
KNN	**72.22222222222221**
Decision Tree	**0.68519**
XGboost	**0.87037**
Neural Network Batch Size: 32 Epoch: 75	**76.74418604651163**

- به غیر از رندوم فارست و ایکس جی بوست، الگوریتم های دیگر دقت آنچنان بالایی نداشته اند.
- این مسئله احتمال به دلیل بیش از حد کم بودن رکوردها میباشد.
- شبکه عصبی دقت آنچنان خوبی از خود نشان نداده، اما با توجه به نتایج دیگر الگوریتم ها، قابل قبول است.
- تعداد کلاس ها، شش می باشد.
- SVG در این دیتاست از دقت بسیار کمی برخوردار است، که نشانگر آن است که الگوریتم مناسبی برای این مسئله نیست.

دیتاست هشتم: Titanic

ALGORITHM NAME	ACCURACY
Logistic Regression	**0.79730**
Random Forest	**0.81532**
SVG	**0.68468**
KNN	**72.52252252252252**
Decision Tree	**0.78829**
XGboost	**0.81081**
Neural Network Batch Size: 32 Epoch: 100	**87.07865168539325**

- شبکه عصبی بیشترین دقت را نشان داده است.
- از بین الگوریتم های دیگر، رندوم فارست و ایکس جی بی بیشترین دقت را داشته اند.
- تعداد کلاس ها، دو می باشد.

دیتاست نهم: Iris

ALGORITHM NAME	ACCURACY
Logistic Regression	**0.97368**
Random Forest	**0.97368**
SVG	**0.97368**
KNN	**97.36842105263158**
Decision Tree	**0.97368**
XGboost	**0.97368**
Neural Network	**80.0**
Batch Size: 20	**100.0**
Epoch: 100	**53.3333**

- این مسئله، مسئله جالبی است، زیرا با چند بار ران کردن الگوریتم و تغییر داده های تست و تمرین، دقت الگوریتم های دیگر ثابت مانده ولی دقت شبکه عصبی به شدت تغییر می کند.
- دیتاست آیریس، یک دیتاست کوچک و تمرینی است و ۱۵۰ رکورد بیشتر ندارد.
- می توان نتیجه گیری کرد که برای آیریس، شبکه عصبی الگوریتم مناسبی نیست زیرا نتایج آن قابل اطمینان نیستند.
- تعداد کلاس ها، سه می باشد.

دیتاست دهم: Ionosphere

ALGORITHM NAME	ACCURACY
Logistic Regression	**0.89773**
Random Forest	**0.98864**
SVG	**0.98864**
KNN	**93.18181818181817**
Decision Tree	**0.92045**
XGboost	**0.98864**
Neural Network Batch Size: 32 Epoch: 100	**98.57142857142858**

- رندوم فارست، اس وی جی، ایکس جی بوست و شبکه عصبی بهترین نتایج را به دست آورده اند.
- این دیتاست، نسبتا تعداد رکوردهای بیشتری داشت، فلذا می توان به نتیجه شبکه عصبی تا حد خوبی اعتماد کرد.
- تعداد کلاس ها، دو می باشد.

با تشکر