بسمه تعالی

مقدمه ای بر یادگیری ماشین

نیمسال اول ۹۸-۹۹

تمرین عملی سری چهارم

مدرس: صابر صالح

● مهلت تحویل تمرین ها: ۱۳۹۸/۱۰/۳ ●

۱ مقدمه

پس از آموختن شیوه های مختلف تحلیل اکتشافی داده ها، اکنون زمان بدست آوردن پیشبینی صحیح بر اساس داده های مرتب شده می باشد. با استفاده از کتابخانه های موجود برای پایتون، انواع مختلفی از الگوریتم های یادگیری ماشین به راحتی قابل دسترس هستند. در برخورد با هر مسئله، لازم است تا ابتدا شرایط مسئله و هدف آن به صورت دقیق بررسی شود. سپس با توجه به شرایط مسئله الگوریتم مناسب برای حل آن به کار گرفته می شود. مسئله داده های تایتانیک که از قبل با آن آشنایی دارید، یک مسئله طبقه بندی در یادگیری با نظارت است. همین شناسایی اولیه مسئله، تعداد زیادی از الگوریتم های یادگیری ماشین را از دایره ابزارهای ما برای حل مسئله حذف می کند.

۲ آشنایی با مسئله

در این تمرین، تعدادی از الگوریتم هایی که در درس دیده اید را برای پیشبینی در داده های تایتانیک استفاده خواهید کرد. لازم است که ابتدا هر یک از الگوریتم ها و نحوه استفاده آن ها را بررسی کرده و سپس در حل مسئله از آن ها استفاده کنید.

- Logistic Regression
- KNN or k-Nearest Neighbors
- Support Vector Machines
- Naive Bayes classifier
- Random Forest
- Neural Network

حال تصور کنید که تمام الگوریتم ها را پیاده سازی کرده اید. کدام یک بهترین عملکرد را داشته است؟ معیار مقایسه الگوریتم ها کدام است؟ ارزیابی الگوریتم های یادگیری ماشین از مهم ترین قسمت های هر پروژه یادگیری ماشین می باشد. حتی امکان آن وجود دارد که الگوریتم شما برای یک متریک پاسخ بسیار مناسب و برای معیاری دیگر پاسخی غیر قابل قبول بدهد. برای پاسخ به این سوال، معیارهای گوناگونی ارائه شده اند. هر یک از این معیارها برای ارزیابی در شرایطی خاص به کار رفته و نمایانگر اثربخشی الگوریتم از دیدگاه مخصوص به معیار می باشند. از بین این معیارها برخی از قبل برای شما آشنا هستند. سایر معیارها و نحوه اندازه گیری آنها به طور خلاصه در ادامه آمده است.

- Classification Accuracy
- Logarithmic Loss
- Confusion Matrix (CM)
- Area under Curve
- F1 Score
- MAE

برای آشنایی با این معبارها میتوانید به این لینک مراجعه نمایید.

Classification Metrics

۳ پیاده سازی

در ابتدا داده ها را به دو دسته ی آموزش و تست تقسیم نمایید. ۸۰ درصد داده ها را به داده های آموزش و ۲۰ درصد باقیمانده را به داده های تست اختصاص دهید. در استفاده از ویژگی ها و ساخت ویژگی های جدید محدودیتی وجود ندارد اما استفاده از ویژگی هایی که به طور قابل توجهی (درصد های پایین دقت) پاسخ های مناسب حاصل نکنند امتیاز کمتری در مقایسه با ویژگی های موثرتر خواهند داشت.

۱.۳ Logistic Regression

با اعمال الگوریتم بر روی داده های تست، ضریب های هر ویژگی در مدل بدست آمده را مشخص کرده و آنها را به ترتیب زیاد به کم مرتب نمایید.

- کدام ویژگی بیشترین اثر بر روی افزایش احتمال زنده ماندن را دارد؟
- کدام ویژگی بیشترین اثر بر روی کاهش احتمال زنده ماندن را دارد؟
- دقت الگوریتم را بر روی داده های آموزش و تست مشخص کنید.
- متریک Confusion Matrix را به دست آورید.

۲.۳ KNN

- الگوریتم را بر روی داده های تست اعمال نمایید. از نرم های ۱، ۲ و ۳ به منظور تعیین فاصله بین داده ها استفاده کنید.
- همچنین به ازای نرم ۲، این الگوریتم را به ازای مقادیر k بین ۱ تا ۱۰ اجرا کرده و میزان خطا را به ازای k های مختلف بر روی داده های تست و آموزش رسم نمایید.
- به ازای چه مقداری از k کمترین خطا بر روی داده های تست مشاهده میشود؟
- با استفاده از k بدست آمده در قسمت قبل، Confusion Matrix را برای داده های تست بدست آورید. سپس آن را با Confusion Matrix قسمت قبل مقایسه کرده و نتایج را تحلیل نمایید.

۳.۳ Support Vector Machines

- الگوریتم SVM را با توابع کرنل linear, poly, rbf پیاده سازی نمایید. دقت الگوریتم را بر روی داده های تست و آموزش به ازای هریک از کرنل های بالا تعیین کنید. کدام کرنل بهترین دقت را بر روی داده های تست به ما میدهد؟
- مقدار F1 Score را برای SVM های بالا بدست آورده و اعداد بدست آمده را تحلیل کنید.
- با استفاده از PCA، بعد داده ها را به ۲ کاهش دهید. سپس بر روی این داده ها الگوریتم SVM با بهترین کرنل را پیاده سازی نمایید. با استفاده از کتابخانه ی mlxtend.plotting، داده های تمرینی و decision boundary را نمایش دهید. آیا overfit اتفاق افتاده است؟

۴.۳ Naive Bayes Classifier

در یادگیری ماشین Naive Bayes Classifier مدل های بسیار ساده ای بر اساس اعمال قانون Bayes هستند. این مدل ها مرتبه خطی از پارامتر بر اساس اندازه بردار ویژگی (feature) دارند و بسیار scalable هستند.

- با استفاده از کتابخانه scikit-learn یک Gaussian Naive Bayes طراحی کنید و مقدار smoothing را در حالت دیفالت نگه دارید.
- کاربرد ضریب smoothing را توضیح دهید.
- AUC را بدست آورید و گزارش کنید. مزیت ان نسبت به accuracy چیست؟

۵.۳ Random Forest

در بین مدل های یادگیری ماشین کلاسیک مدل های ensemble محبوبیت زیادی دارند. از جمله این مدل ها می توان به random forest اشاره کرد. در این روش مدل های پایه ما در مدل ensemble درخت های تصمیم هستند. در نهایت در حالت classification مد و در حالت regression میانگین تصمیم های درخت های مختلف به عنوان تصمیم نهایی اعلام می شود.

- با کمک گیری از کتابخانه scikit-learn مدل random forest را طراحی کنید در این مدل ها تعداد درخت ها و همین طور بیشینه عمق ان ها از ابر پارامتر های مهم است. سعی کنید این ۲ پارامتر را با استفاده از five fold cross validation بهینه کنید. (برای بقیه پارامترها می توانید مقدار پیش فرض را در نظر بگیرید.)
- پارامترهای بهینه را گزارش کنید.
- همین طور F1 score را گزارش کنید.

۶.۳ Artificial Neural Network

شبکه های عصبی مصنوعی انواعی از مدل های یادگیری ماشین هستند که ساختار آنها الهام گرفته از ساختار شبکه عصبی در انسان است. ساده ترین نوع شبکه های عصبی شبکه های عصبی تمام متصل هستند. ساختار این شبکه ها بسیار ساده است. در هر لایه یک ماتریس از وزن ها در ورودی آن لایه ضرب می شود و سپس یک تابع غیر خطی روی خروجی لایه اعمال می شود. در نهایت شبکه سعی می کند وزن ها را طوری تغییر دهید تا تابع هزینه مورد نظر بر روی داده ی تمرین کمینه شود.

- با کمک گیری از کتابخانه scikit-learn یک MLP سه لایه طراحی کنید. تابع activation را relu بگذارید. از adam به عنوان solver استفاده کنید. batch size را ۳۲ و همین طور تعداد epochs را ۲۰ در نظر بگیرید. یک بار شبکه شما به ترتیب ۲۰ و ۱۰ node در لایه های مخفی داشته باشد و یک بار به ترتیب ۱۰۰ و ۵۰ node داشته باشد. در آموزش شبکه های عصبی learning rate و ضریب regularization از ضرایب بسیار مهم هستند سعی کنید تا حد امکان ضرایب بهینه را با روش دلخواه انتخاب کنید. (برای بقیه پارامترها می توانید مقدار پیش فرض را در نظر بگیرید.)
- مدل با ۲۰ و ۱۰ node بهتر است یا مدل با ۱۰۰ و ۵۰ node؟ به نظر شما دلیل چیست؟
- همین طور Confusion Matrix را گزارش کنید.

۴ قسمت امتیازی

الگوریتمی برای داده های تایتانیک پیشنهاد دهید که دقت ۹۵ درصد و یا بالاتر را حاصل کند. دقت کنید که در این راه مجاز هستید از تغییر ویژگی ها و ساخت ویژگی های جدید استفاده کنید. همچنین درصد داده های تست و تمرین را مشابه قسمت های قبلی ۲۰ به ۸۰ تنظیم نمایید.

- برای الگوریتم خود تمامی معیارهای یاد شده در مقدمه را اندازه گیری کنید. الگوریتم خود را تنها برای خطاهای Confusion Matrix, F1 Score با روش های قبلی مقایسه کنید. بهترین و بدترین عملکرد الگوریتم در کدام خطا قابل مشاهده می باشد؟
- داده های آموزش و تست را به نسبت ۹۰ به ۱۰ تقسیم کرده و مجددا الگوریتم خود را پیاده سازی کنید. در این حالت مجددا خطاهای ذکر شده در قسمت قبل را اندازه گیری کرده و نتایج را تحلیل کنید.