

"به نام خدا"

مروارید لعل نور 9623097

گزارش تمرین شماره 2-3 درس یادگیری ماشین

در این تمرین ابتدا توابع مورد نیاز برای محاسبه میانگین و انحراف معیار و متغیرهای آماری را در ابتدای کد تعریف کرده و سپس برای هماهنگ سازی فرمت داده ها با فرمت قابل محاسبه برای این توابع ، توابعی مانند `str_column_to_float` , `str_column_to_int` , `separate_by_class` و غیره را تعریف کردیم .

بخش الف (

همانطور که خواسته شده داده ها را به دو قسمت با نسبت 30 به 70 برای تست و ترین تقسیم کرده و با استفاده از توابع تعریف شده که کار تبدیل ستونی رشته ها به عدد و محاسبه را انجام میدهند عمل یادگیری را انجام دادیم که در نتیجه تست کردن همانطور که در کد آورده شده است صحت این یادگیری حدودا 89 درصد است ( ماتریس `confusion` نیز در نتایج کد قابل مشاهده است )

وقتی برای بار دوم این کار را تکرار میکنیم ، صحت عملکرد حدود 94 درصد بدست می آید که به صورت میانگین میتوان صحت را در این قسمت برابر 91.5 درصد در نظر گرفت .

بخش ب (

در این قسمت از ما خواسته شده بود که با `4-fold cross validation` الگوریتم را اجرا کنیم . به این ترتیب دادگان را به قسمت های 37 تایی تقسیم کردیم ( از آنجایی که 150 بر 4 بخش پذیر نیست ) و چهار بار و هر بار با یکی از آن قسمت ها به عنوان داده تست و دیگری به عنوان داده ترین الگوریتم را اجرا کردیم ( ابتدا نیز دادگان را `shuffle` کردیم تا به صورت رندوم این تقسیم بندی انجام شود ) . نتایج در این بخش به این صورت شد که برای بار اول صحت عملکرد برابر 80 درصد ، برای بار دوم 78 درصد ، برای بار سوم 78 درصد و باری بار چهارم 83 درصد بدست آمد که به صورت میانگین برابر با 79.5 درصد می باشد .

بخش پ و ت (

برای به کار بردن الگوریتم `Gaussian naïve Bayes` از تعریف کلاسی به این نام استفاده کردیم که در آن الگوریتم `Gaussian naïve Bayes` مطابق با روند محاسبه در مسایل پیاده شده است و مقدار صحت و ماتریس `confusion` در کد ( با پسوند `.ipynp` ) مشخص شده است . که به طور میانگین این مقدار برابر 37.125 درصد (!!!؟) بدست آمده که در مقایسه با بخش های پیشین به صورت قابل توجهی کمتر است .

نتیجه :

با توجه به آنکه پیش فرض الگوریتم `Gaussian naïve Bayes` مستقل بودن خصوصیات نسبت به یکدیگر است و با این فرض صحت الگوریتم ما کاهش یافته است ، میتوان نتیجه گرفت که در این دیتاست خصوصیات به یکدیگر وابسته اند و شرط استقلال میتواند روند یادگیری را دچار اشتباه کند و بنابرین عملکرد الگوریتم کاهش یابد .